Multiple-Input
Heavy-Traffic Real-Time Queues

June 12, 2001

Lukasz Kruk[1]
Department of Mathematical Sciences
Carnegie Mellon University
kruk+@andrew.cmu.edu

John Lehoczky
Department of Statistics
Carnegie Mellon University
jpl@stat.cmu.edu

Steven Shreve[2]
Department of Mathematical Sciences
Carnegie Mellon University
shreve@cmu.edu

Shu-Ngai Yeung
Department of Statistics
Carnegie Mellon University
syeung@stat.cmu.edu.

ABSTRACT

A single queueing station which serves K input streams is considered. Each stream is an independent renewal process, with customers having random lead-times. Customers are served by processor sharing across streams. Within each stream, two disciplines are considered – earliest-deadline-first and first-in-first-out. The set of current lead times of the K streams is modeled as a K-dimensional vector of random counting measures on $\mathbb{R}$, and the limit of this vector of measure-valued processes is obtained under heavy traffic conditions.

Short title: Multiple-Input Queues
Keywords: Due dates, heavy traffic, queueing, diffusion limits, random measures
AMS subject classification: Primary 60K25; Secondary 60G57, 60J65

[1]Supported by the Center for Nonlinear Analysis (NSF Grant No. DMS-98-03791).
[2]Supported by NSF Grant No. DMS-98-02464.

1 Introduction

Over the last 10 years, communication technology has become dramatically more sophisticated. There are now many different types of communication services available and an ever increasing demand for those services. Real-time communication services, while currently a small portion of the total demand, are of growing importance. In view of the increase in demand for real-time services (such as video-conferencing in which voice and video must be delivered in a timely fashion to ensure continuity of the image and sound) it is becoming important to develop models and analysis techniques first to understand and then to better control these networks. For real-time traffic, in addition to standard measures of stability and delay, one must also be concerned whether individual application packets are meeting their timing requirements. Thus measuring average packet delay is not sufficient to determining whether a particular queue scheduling policy can satisfy real-time traffic requirements.

There has been important recent work that does develop analytic methods to determine stability and calculate delay. An excellent introduction to the scheduling of computer networks is given in in the textbook by Kershav[12], Chapter 9. In addition, the paper by Zhang [20] surveys scheduling algorithms and associated performance bounds. More specifically, Cruz developed a calculus for network delay in the single node [6] and the multiple node [7] cases. The theory he developed is different from standard queueing theoretic models, because he used non-probabilistic data stream arrival models which satisfy burstiness constraints. Using this assumption he was able to derive bounds on delay and buffering requirements for network elements. Parekh and Gallager [14], [15] studied a new network resource allocation policy called the generalized (or weighted) processor sharing (GPS) algorithm or weighted fair queueing (WFQ) and combined this with a leaky bucket admission control policy. Demers et. al. [8] presented an analysis of these approaches. For these scheduling and control mechanisms the authors developed worst-case packet delay and worst-case session backlogs for both the single node and multiple node cases. Related research was published by Chang [4]. He also addressed the stability problem by determining conditions on queueing networks that result in bounded queue lengths and bounded delays for customers and gave conditions for the queue length distribution to have an exponential tail.

Most of these papers present a worst case analysis leading to bounds for the system which may be quite pessimistic. Thus it is important to find methods that offer more realistic answers. An additional drawback is that the methods do not apply directly to real-time traffic where performance must be measured at the packet level, rather than at an aggregated level. This paper introduces methods for some common queue scheduling policies to assess whether those packet timing requirements can be met.

In this paper, we consider a single node system with K independent arrival streams (sessions). Each stream creates an arrival process of packets, and each packet has an individual service time and deadline, the time by which its transmission must be completed. Packets from individual streams are queued in separate buffers. The K streams are jointly serviced by a single server who allocates its total service capacity across the streams. We consider the *head-of-the-line processor sharing* service policy (HOL-PS). This policy assumes that the total service capacity is evenly allocated to each stream having packets ready for transmission. Thus, if k of the K streams have one or more packets available for service, then each receives a fraction $1/k$ of the servers capacity. We also introduce a weighting scheme to allow different streams to receive different relative amounts of service. We define a set of weights, $\{w_i,\ 1 \leq i \leq K\},\ w_i \geq 0,\ \sum_{i=1}^{K} w_i = 1$. If two streams i and j have packets ready for transmission, then the relative service rate applied to streams i and j is the ratio w_i/w_j.

We assume that the packets from each stream are queued in different buffers either in deadline order or in FIFO order. Ordinarily packets from a common stream will have a common, deterministic deadline; however, our approach permits us to consider the more general case in which packets from a single stream have random deadlines. For real-time systems the service policy should take into account the packet timing requirements, and the Earliest-Deadline-First (EDF) algorithm for a single queue is generally optimal. The FIFO queueing discipline arises as a special case of EDF when the packets have constant deadlines.

Our approach follows the methods of Doytchinov, Lehoczky and Shreve [9]. (We note that the single-station results of [9] have recently been extended to feed-forward networks; see [18].) We study the heavy traffic case in which the total traffic intensity on the server approaches 1. The HOL-PS queue discipline is work conserving, so under reasonable conditions on the arrival and service processes, the total workload on the system will converge to a drifted, reflected Brownian motion process. It is, however, more complicated to determine how that workload will be distributed across the K different queues and whether the packets in those queues will meet their timing requirements. To determine this, we define a lead-time for each packet (the lead-time is the time until the packet deadline) and a *lead-time profile* for each of the K queues. Under heavy traffic conditions, we prove that these profiles, when conditioned on the workload, converge to a deterministic form (including a null profile when the scaled queue length process converges to 0) which depends upon the task deadlines, the queue discipline and the traffic intensities for each of the queues. The ability of this service discipline to meet packet timing requirements can be inferred from these profiles.

This paper is organized as follows. Section 2 presents the model, assumptions and the notation. Section 3 introduces the heavy traffic assumptions. Section 4 develops the measure-valued processes that will be needed to study the lead-time profiles. Sections 5 and 6 present limit theorems for those measure-valued processes assuming individual packet streams are queued using either EDF or FIFO respectively. Section 7 presents simulation results illustrating the accuracy of the theory. Appendix A presents an analysis when the traffic streams are not balanced and one of the streams is dominant, and Appendix B provides a technical result needed for Appendix A.

2 The model

We have a sequence of single-station queueing systems, each with K arrival processes. The queueing systems are indexed by superscript (n), and the arrival processes within each queueing system are indexed by k.

The inter-arrival times for the k-th arrival process are $\{u_{k,j}^{(n)}\}_{j=1}^{\infty}$, a sequence of positive, independent, identically distributed random variables with common mean $\frac{1}{\lambda_k^{(n)}}$ and standard deviation $\alpha_k^{(n)}$. The service times are $\{v_{k,j}^{(n)}\}_{j=1}^{\infty}$, another sequence of positive, independent, identically distributed random variables with common mean $\frac{1}{\mu_k^{(n)}}$ and standard deviation $\beta_k^{(n)}$. We assume that each queue is empty at time zero.

For arrival stream k, we define the *customer arrival times*

$$S_{k,0}^{(n)} \triangleq 0, \qquad S_{k,j}^{(n)} \triangleq \sum_{i=1}^{j} u_{k,i}^{(n)}, \quad j \geq 1, \tag{2.1}$$

the *customer arrival process*

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

and the *work arrival process*

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

The work which has arrived to queue k by time t is then $V_k^{(n)}(A_k^{(n)}(t))$.

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

$$\mathbb{P}\{L_{k,j}^{(n)} \leq \sqrt{n}y\} = G_k(y), \tag{2.4}$$

where G_k is a right-continuous cumulative distribution function for each $k = 1, \ldots, K$. We define

$$y_k^* \triangleq \min\{y \in \mathbb{R}; G_k(y) = 1\}, \tag{2.5}$$

and assume that y_k^* is finite for every $k = 1, \ldots, K$. We assume that for every n, the sequences $\{u_{k,j}^{(n)}\}$, $\{v_{k,j}^{(n)}\}$ and $\{L_{k,j}^{(n)}\}$ are mutually independent over $k = 1, \ldots, K$ and $j = 1, 2, \ldots$.

We now describe the allocation of service capacity to each of the K arrival streams in the n-th queueing system. To this end, we introduce nonnegative random *weight* processes $w_k^{(n)}(t)$, $k = 1, \ldots, K$, such that $\sum_{k=1}^{K} w_k^{(n)}(t) = 1$

for every t almost surely. The process $w_k^{(n)}(t)$ represents the proportion of the server's capacity which would be assigned to queue k at time t if all the queues are nonempty at that time. If one or more queues is empty but queue k is not, then $w_k^{(n)}(t)$ divided by the sum of the weights associated with the nonempty queues is the proportion of the server's capacity assigned to queue k. We allow the weight processes to take the value zero, but we make the following assumption, which guarantees that the proportion just described is defined.

Assumption 2.1 *At every time t when there is at least one customer present in the n-th queueing system, there is at least one nonempty queue ℓ for which $w_\ell^{(n)}(t) > 0$.*

An important consequence of Assumption 2.1 is that the server is always fully utilized whenever there are customers anywhere in the system.

Throughout Sections 2–5, we assume that customers within each of the K queues are served using the Earliest-Deadline-First (EDF) queue discipline, i.e., the server always serves the customer with the shortest lead-time. Preemption is permitted (we assume preempt-resume). There is no set up, switch-over, or other type of overhead. Late customers (customers with negative lead-times) stay in queue until served to completion. In Section 6, customers are served using the First-In-First-Out (FIFO) discipline within each queue.

Denoting by $W_k^{(n)}(t)$ the work remaining in queue k at time t, we define the *idleness rate* for queue k to be

$$\dot{I}_k^{(n)}(t) \triangleq \begin{cases} 1, & \text{if } W_k^{(n)}(t) = 0, \\ 0, & \text{if } W_k^{(n)}(t) > 0. \end{cases} \tag{2.6}$$

The *entitlement rate* for queue k is then defined as

$$\dot{T}_k^{(n)}(t) \triangleq \frac{w_k^{(n)}(t)}{w_k^{(n)}(t) + \sum_{\ell \neq k} w_\ell^{(n)}(t)(1 - \dot{I}_\ell^{(n)}(t))}. \tag{2.7}$$

Because of Assumption 2.1, the denominator in (2.7) is not zero as long as there is at least one customer present in the system. If this is not the case and if $w_k^{(n)}(t) = 0$, we interpret $\frac{0}{0}$ in (2.7) as 0. The entitlement rate for queue k is the fraction of the server queue k is entitled to receive at time t and would receive if there is work in queue k at that time.

We may alternatively represent the server entitlement rate for a queue in terms of joint idleness rates. By a *multi-index* $\alpha = (\ell_1, \ldots, \ell_{|\alpha|})$, we shall mean a finite set of distinct indices $\{\ell_1, \ldots, \ell_{|\alpha|}\}$ between 1 and K. We denote by $|\alpha|$ the cardinality of the set α. Given a multi-index α, we define the *joint idleness rate*

$$\dot{J}_\alpha^{(n)}(t) \triangleq \begin{cases} 1, & \text{if } W_\ell^{(n)}(t) = 0 \ \forall \ell \in \alpha, \ W_\ell^{(n)}(t) > 0 \ \forall \ell \notin \alpha, \\ 0, & \text{otherwise.} \end{cases} \tag{2.8}$$

By convention,

$$\dot{J}_{\emptyset}^{(n)}(t) = \begin{cases} 1, & \text{if } W_\ell^{(n)}(t) > 0 \ \forall \ell, \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

Finally, we define $A_k(m)$ to be the set of all multi-indexes of cardinality m in which the index k appears and $A_k^c(m)$ to be the set of all multi-indexes of cardinality m in which the index k does not appear. By convention, $A_k(0) = \emptyset$ and $A_k^c(0) = \{\emptyset\}$. Then the idleness rate for queue k is

$$\dot{I}_k^{(n)}(t) = \sum_{m=1}^{K} \sum_{\alpha \in A_k(m)} \dot{J}_\alpha^{(n)}(t), \tag{2.10}$$

and the entitlement rate $\dot{T}_k^{(n)}(t)$ can be written as

$$\begin{aligned} \dot{T}_k^{(n)}(t) &= \sum_{m=0}^{K-1} \sum_{\alpha \in A_k^c(m)} \frac{w_k^{(n)}(t)}{1 - \sum_{\ell \in \alpha} w_\ell^{(n)}(t)} \dot{J}_\alpha^{(n)}(t) \\ &\quad + \sum_{m=1}^{K} \sum_{\alpha \in A_k(m)} \frac{w_k^{(n)}(t)}{w_k^{(n)}(t) + 1 - \sum_{\ell \in \alpha} w_\ell^{(n)}(t)} \dot{J}_\alpha^{(n)}(t). \end{aligned} \tag{2.11}$$

It can happen in the above expressions that the denominator $1 - \sum_{\ell \in \alpha} w_\ell^{(n)}(t)$ or $w_k^{(n)}(t) + 1 - \sum_{\ell \in \alpha} w_\ell^{(n)}(t)$ is zero, but this can only occur if the numerator $w_k^{(n)}(t)$ is also zero. In such a case, we interpret $\frac{0}{0} \cdot 0$ as 0. At each time at most one positive term appears on the right-hand side of (2.10) and (2.11). We define *idleness*, *joint idleness* and *entitlement* by integrating rates:

$$I_k^{(n)}(t) = \int_0^t \dot{I}_k^{(n)}(s)ds, \quad J_\alpha^{(n)}(t) = \int_0^t \dot{J}_\alpha^{(n)}(s)ds, \quad T_k^{(n)}(t) = \int_0^t \dot{T}_k^{(n)}(s)ds. \tag{2.12}$$

In particular,

$$I_k^{(n)}(t) = \sum_{m=1}^{K} \sum_{\alpha \in A_k(m)} J_\alpha^{(n)}(t), \tag{2.13}$$

$$\begin{aligned} T_k^{(n)}(t) &= \sum_{m=0}^{K-1} \sum_{\alpha \in A_k^c(m)} \int_0^t \frac{w_k^{(n)}(s)}{1 - \sum_{\ell \in \alpha} w_\ell^{(n)}(s)} \dot{J}_\alpha^{(n)}(s)\, ds \\ &\quad + \sum_{m=1}^{K} \sum_{\alpha \in A_k(m)} \int_0^t \frac{w_k^{(n)}(s)}{w_k^{(n)}(s) + 1 - \sum_{\ell \in \alpha} w_\ell^{(n)}(s)} \dot{J}_\alpha^{(n)}(s)\, ds. \end{aligned} \tag{2.14}$$

The *netput* for queue k is

$$N_k^{(n)}(t) \triangleq V_k^{(n)}\big(A_k^{(n)}(t)\big) - T_k^{(n)}(t), \tag{2.15}$$

which is the amount of work in queue at time t if it has never been empty prior to time t. Because queue k might be empty, it can have *unused entitlement*, defined to be

$$U_k^{(n)}(t) \triangleq -\min_{0\leq s\leq t} N_k^{(n)}(s). \tag{2.16}$$

The actual service received by queue k up to time t is

$$R_k^{(n)} \triangleq T_k^{(n)}(t) - U_k^{(n)}(t), \tag{2.17}$$

and the workload at time t is

$$W_k^{(n)}(t) = N_k^{(n)}(t) + U_k^{(n)}(t) = V_k^{(n)}\big(A_k^{(n)}(t)\big) - R_k^{(n)}(t). \tag{2.18}$$

The unused entitlement increases at time t if and only if queue k is empty at time t and $w_k^{(n)}(t) > 0$, and in this case, $\dot{U}_k^{(n)}(t) = \dot{T}_k^{(n)}(t)$. This implies that

$$U_k^{(n)}(t) = \sum_{m=1}^{K} \sum_{\alpha\in A_k(m)} \int_0^t \frac{w_k^{(n)}(s)}{w_k^{(n)}(s) + 1 - \sum_{\ell\in\alpha} w_\ell^{(n)}(s)} \dot{J}_\alpha^{(n)}(s)\, ds. \tag{2.19}$$

Indeed, this is precisely the part of (2.14) corresponding to queue k being empty.

All the above processes are independent of the service discipline within each queue, provided that the processes $w_k^{(n)}(t)$ remain unchanged. However, the *queue length processes* $Q_k^{(n)}(t)$, which are the number of customers in each queue $k = 1, \dots, K$ at time t, depend on the queue discipline.

3 Heavy traffic assumptions

We assume that for each k, the following limits exist and are all positive:

$$\lambda_k = \lim_{n\to\infty} \lambda_k^{(n)},\ \mu_k = \lim_{n\to\infty} \mu_k^{(n)},\ \alpha_k = \lim_{n\to\infty} \alpha_k^{(n)},\ \beta_k = \lim_{n\to\infty} \beta_k^{(n)}. \tag{3.1}$$

The *traffic intensity* for queue k in queueing system n is $\rho_k^{(n)} = \frac{\lambda_k^{(n)}}{\mu_k^{(n)}}$, and the *limiting traffic intensity* for queue k is $\rho_k = \frac{\lambda_k}{\mu_k}$. We assume

$$\sum_{k=1}^{K} \rho_k = 1 \tag{3.2}$$

and make the *heavy traffic assumption*

$$\lim_{n\to\infty} \sqrt{n}(\rho_k - \rho_k^{(n)}) = \gamma_k > 0 \tag{3.3}$$

for each $k = 1, \dots, K$. In particular, there is a finite constant c such that

$$\sqrt{n}\left|\rho_k^{(n)} - \rho_k\right| \leq c \tag{3.4}$$

for all n and k. We also impose the usual Lindeberg condition on the inter-arrival and service times: for $k = 1, \ldots, K$,

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{E}\left[\left(u_{k,j}^{(n)} - (\lambda_k^{(n)})^{-1}\right)^2 \mathbb{I}_{\left\{\left|u_{k,j}^{(n)} - (\lambda_k^{(n)})^{-1}\right| > c\sqrt{n}\right\}}\right] \qquad (3.5) \\ = &\lim_{n\to\infty} \mathbb{E}\left[\left(v_{k,j}^{(n)} - (\mu_k^{(n)})^{-1}\right)^2 \mathbb{I}_{\left\{\left|v_{k,j}^{(n)} - (\mu_k^{(n)})^{-1}\right| > c\sqrt{n}\right\}}\right] = 0 \quad \forall c > 0. \end{aligned}$$

We introduce the *heavy traffic scaling* for the workload and queue length processes

$$\widehat{W}_k^{(n)}(t) = \frac{1}{\sqrt{n}} W_k^{(n)}(nt), \quad \widehat{Q}_k^{(n)}(t) = \frac{1}{\sqrt{n}} Q_k^{(n)}(nt), \tag{3.6}$$

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

$$\widehat{A}_k^{(n)}(t) = \frac{1}{\sqrt{n}} \left[A_k^{(n)}(nt) - \lambda_k^{(n)} nt\right]. \tag{3.7}$$

For $k = 1, \ldots, K$ and $y \leq y_k^*$, we define the respective *partial work arrival process* and *centered scaled partial work arrival process*

$$V_{k,y}^{(n)}(t) = \sum_{j=1}^{\lfloor t \rfloor} v_{k,j}^{(n)} \mathbb{I}_{\{L_{k,j}^{(n)} \leq \sqrt{n}\, y\}}, \tag{3.8}$$

$$\widehat{V}_{k,y}^{(n)}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} \left(v_{k,j}^{(n)} \mathbb{I}_{\{L_{k,j}^{(n)} \leq \sqrt{n}\, y\}} - \left(\mu_k^{(n)}\right)^{-1} G_k(y)\right). \tag{3.9}$$

and we define also

$$\widehat{Z}_{k,y}^{(n)}(t) = \frac{1}{\sqrt{n}} \left[V_{k,y}^{(n)}\left(A_k^{(n)}(nt)\right) - G_k(y) \rho_k^{(n)} nt\right]. \tag{3.10}$$

Note that the term $\left(\mu_k^{(n)}\right)^{-1} G_k(y)$ appearing in (3.9) is the expectation of $v_{k,j}^{(n)} \mathbb{I}_{\{L_{k,j}^{(n)} \leq \sqrt{n}\, y\}}$. The variance of this random variable is

$$\left(\beta_{k,y}^{(n)}\right)^2 \triangleq G_k(y) \left(\beta_k^{(n)}\right)^2 + \left(\mu_k^{(n)}\right)^{-2} G_k(y) \left(1 - G_k(y)\right), \tag{3.11}$$

which has limit

$$\left(\beta_{k,y}\right)^2 = G_k(y) (\beta_k)^2 + (\mu_k)^{-2} G_k(y) \left(1 - G_k(y)\right). \tag{3.12}$$

We use the notation $\widehat{V}_k^{(n)} = \widehat{V}_{k,y^*}^{(n)}$ and $\widehat{Z}_k^{(n)} = \widehat{Z}_{k,y^*}^{(n)}$. Note that $V_k^{(n)} = V_{k,y^*}^{(n)}$ and $\beta_k^{(n)} = \beta_{k,y^*}^{(n)}$.

The following theorem is due to Prokhorov [16], Theorem 3.1 (see also [1], Theorem 14.6). Here and elsewhere, the symbol $\Rightarrow$ denotes weak convergence of measures on the space $D[0, \infty)$ of functions from $[0, \infty)$ to $\mathbb{R}$ which are right-continuous with left limits. The topology of this space is a generalization of the topology introduced by Skorohod for $D[0, 1]$. See [1] for details.

Theorem 3.1 *For every k=1, ... ,K and every $y \le y_k^*$,*

$$\big(\widehat{A}_k^{(n)}, \widehat{V}_{k,y}^{(n)}\big) \Rightarrow \big(A_k^*, V_{k,y}^*\big), \tag{3.13}$$

where A_k^*, $V_{k,y}^*$ are independent Brownian motions with mean 0 and variances $\alpha_k^2\lambda_k^3$ and $\beta_{k,y}^2$ respectively.

This theorem has the following corollaries.

Corollary 3.2 *For every $k = 1, \dots, K$ and $y \le y_k^*$*

$$\widehat{Z}_{k,y}^{(n)} \Rightarrow V_{k,y}^* \circ \lambda_k e + \frac{G_k(y)}{\mu_k} A_k^*, \tag{3.14}$$

where e is the identify function $e(t) = t$ for all $t \ge 0$.

PROOF: We compute

$$\begin{aligned}
&\widehat{Z}_{k,y}^{(n)}(t) \\
&= \frac{1}{\sqrt{n}} \left[V_{k,y}^{(n)}\big(A_k^{(n)}(nt)\big) - G_k(y)\rho_k^{(n)} nt \right] \qquad (3.15) \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^{A_k^{(n)}(nt)} \left[v_{k,j}^{(n)} \mathbb{I}_{\{L_{k,j}^{(n)} \le \sqrt{n}\, y\}} - \frac{G_k(y)}{\mu_k^{(n)}} \right] + \frac{G_k(y)}{\mu_k^{(n)}\sqrt{n}} \left[A_k^{(n)}(nt) - \lambda_k^{(n)} nt \right] \\
&= \left[\widehat{V}_{k,y}^{(n)}\left(\frac{1}{n} A_k^{(n)}(nt)\right) - \widehat{V}_{k,y}^{(n)}(\lambda_k t) \right] + \widehat{V}_{k,y}^{(n)}(\lambda_k t) + \frac{G_k(y)}{\mu_k^{(n)}} \widehat{A}_k^{(n)}(t).
\end{aligned}$$

The term in square brackets converges to zero because $\frac{1}{n} A_k^{(n)}(nt) \Rightarrow \lambda_k t$ (see, e.g., Theorem A.3 of [9]). □

Corollary 3.3 *The process $\sum_{k=1}^K \widehat{W}_k^{(n)}$ converges weakly to a reflected Brownian motion with drift.*

PROOF: Since $\sum_{k=1}^K \widehat{W}_k^{(n)}$ is the total workload, a quantity which is invariant under work-conserving disciplines, this is a classical result due to Iglehart and Whitt [11]. One can also obtain it immediately from Corollary 3.2 and the equation

$$\begin{aligned}
\sum_{k=1}^K \widehat{W}_k^{(n)}(t) &= \sum_{k=1}^K \widehat{Z}_k^{(n)}(t) - \sqrt{n}\Big(1 - \sum_{k=1}^K \rho_k^{(n)}\Big) t \qquad (3.17) \\
&- \min_{0 \le s \le t} \left[\sum_{k=1}^K \widehat{Z}_k^{(n)}(s) - \sqrt{n}\Big(1 - \sum_{k=1}^K \rho_k^{(n)}\Big) s \right].
\end{aligned}$$

□

For the remainder of this paper, we make the following assumption.

Assumption 3.4 *We have*

$$\widehat{W}^{(n)} \triangleq \left(\widehat{W}_1^{(n)}, \ldots, \widehat{W}_K^{(n)}\right) \Rightarrow W^*, \tag{3.18}$$

where $W^* = (W_1^*, \ldots, W_K^*)$ *is an RCLL process on* $\mathbb{R}^K$*.*

In Appendix A we will show that this assumption holds when the weight processes $w_k^{(n)}(t)$ converge sufficiently fast to positive constants w_k and the traffic intensities are *unbalanced*, i.e.,

$$0 < \frac{\rho_1}{w_1} \leq \frac{\rho_2}{w_2} \leq \cdots \leq \frac{\rho_{K-1}}{w_{K-1}} < \frac{\rho_K}{w_K}. \tag{3.19}$$

Actually, in this case $W_k^* \equiv 0$ for all $k < K$. In the case of *balanced traffic intensities*

$$\frac{\rho_1}{w_1} = \frac{\rho_2}{w_2} = \cdots = \frac{\rho_{K-1}}{w_{K-1}} = \frac{\rho_K}{w_K}, \tag{3.20}$$

Assumption 3.4 is verified in a still unpublished work of M. Reiman and K. Ramanan.

Corollary 3.5 *The processes* W_k^**,* $k = 1, \ldots, K$*, are continuous.*

PROOF: For simplicity, we prove continuity of W_k^* on $[0,1]$. For $x \in D[0,1]$, let $j(x) = \sup_{0 \leq t \leq 1} |x(t) - x(t-)|$. By the Continuous Mapping Theorem $j(\widehat{Z}_k^{(n)}) \Rightarrow 0$. But $j(\widehat{W}_k^{(n)}) = j(\widehat{Z}_k^{(n)})$, and so $j(\widehat{W}_k^{(n)}) \Rightarrow 0$. According to Theorem 13.4 of [1], this implies continuity of the limiting process W_k^*. □

4 Measure-Valued Processes

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

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

In this section, we define a collection of measure-valued processes which will be useful in the analysis of the instantaneous lead-time profile of the customers.

Queue length measures:

$$\mathcal{Q}_k^{(n)}(t)(B) \triangleq \left\{ \begin{array}{l} \text{Number of customers in queue } k \text{ at time } t \\ \text{having lead-times at time } t \text{ in } B \subset \mathbb{R} \end{array} \right\}. \tag{4.1}$$

Workload measures:

$$\mathcal{W}_k^{(n)}(t)(B) \triangleq \left\{ \begin{array}{l} \text{Work in queue } k \text{ at time } t \text{ associated with customers} \\ \text{in this queue having lead-times at time } t \text{ in } B \subset \mathbb{R} \end{array} \right\}. \tag{4.2}$$

Customer arrival measures:

$$\mathcal{A}_k^{(n)}(t)(B) \triangleq \left\{ \begin{array}{l} \text{Number of all arrivals in queue } k \text{ by time } t \\ \text{having lead-times at time } t \text{ in } B \subset \mathbb{R} \end{array} \right\}. \tag{4.3}$$

Workload arrival measures:

$$\mathcal{V}_k^{(n)}(t)(B) \triangleq \left\{ \begin{array}{l} \text{Work in queue } k \text{ associated with all arrivals by} \\ \text{time } t \text{ having lead-times at time } t \text{ in } B \subset \mathbb{R} \end{array} \right\}. \tag{4.4}$$

The following relationships easily follow:

$$Q_k^{(n)}(t) = \mathcal{Q}_k^{(n)}(t)(\mathbb{R}), \qquad W_k^{(n)}(t) = \mathcal{W}_k^{(n)}(t)(\mathbb{R}), \tag{4.5}$$

$$A_k^{(n)}(t) = \mathcal{A}_k^{(n)}(t)(\mathbb{R}), \qquad V_k^{(n)}\big(A_k^{(n)}(t)\big) = \mathcal{V}_k^{(n)}(t)(\mathbb{R}), \tag{4.6}$$

$$\begin{aligned} \mathcal{A}_k^{(n)}(t)(B) &= \sum_{j=1}^{A_k^{(n)}(t)} \mathbb{I}_{\{L_{k,j}^{(n)}-(t-S_{k,j}^{(n)})\in B\}} \qquad (4.7) \\ &= \sum_{j=1}^{\infty} \mathbb{I}_{\{S_{k,j}^{(n)}\in B+t-L_{k,j}^{(n)},\ S_{k,j}^{(n)}\le t\}}, \\ \mathcal{V}_k^{(n)}(t)(B) &= \sum_{j=1}^{A_k^{(n)}(t)} v_{k,j}^{(n)} \mathbb{I}_{\{L_{k,j}^{(n)}-(t-S_{k,j}^{(n)})\in B\}} \qquad (4.8) \\ &= \sum_{j=1}^{\infty} v_{k,j}^{(n)} \mathbb{I}_{\{S_{k,j}^{(n)}\in B+t-L_{k,j}^{(n)},\ S_{k,j}^{(n)}\le t\}}. \end{aligned}$$

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 in each queue k and the largest lead-time of all customers from queue k, whether present or departed, who have ever been in service. For $k = 1, \ldots, K$, we define the *frontier* for queue k

$$F_k^{(n)}(t) \triangleq \left\{ \begin{array}{l} \text{largest lead-time of all customers from queue } k \\ \text{who have ever been in service, whether still} \\ \text{present or not, or } \sqrt{n}\, y_k^* - t, \text{ if this quantity is} \\ \text{larger than the former one} \end{array} \right\}, \tag{4.9}$$

and the *current lead time*

$$C_k^{(n)}(t) \triangleq \left\{ \begin{array}{l} \text{lead-time of the customer in service in} \\ \text{queue } k \text{ or } F_k^{(n)}(t) \text{ if the queue is empty} \end{array} \right\}. \tag{4.10}$$

Prior to arrival of the first customer, $F_k^{(n)}(t) = \sqrt{n}\, y_k^* - t$. Under the EDF queue discipline, there is no customer in queue k with lead-time smaller than $C_k^{(n)}(t)$, and there has never been a customer in service in this queue whose lead-time, if the customer were still present, would exceed $F_k^{(n)}(t)$. Furthermore, $C_k^{(n)}(t) \le F_k^{(n)}(t)$ for all $t \ge 0$, $k = 1, \ldots, K$. Both $F_k^{(n)}$ and $C_k^{(n)}$ are RCLL (right-continuous processes with left-hand limits).

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

$$\widehat{F}_k^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} F_k^{(n)}(nt), \qquad \widehat{C}_k^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} C_k^{(n)}(nt), \tag{4.11}$$

$$\widehat{\mathcal{Q}}_k^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} \mathcal{Q}_k^{(n)}(nt)(\sqrt{n}B), \qquad \widehat{\mathcal{W}}_k^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} \mathcal{W}_k^{(n)}(nt)(\sqrt{n}B). \tag{4.12}$$

We define also

$$\begin{aligned} \widehat{\mathcal{A}}_k^{(n)}(t)(B) &\triangleq \frac{1}{\sqrt{n}} A_k^{(n)}(nt)(\sqrt{n}B), \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^{A_k^{(n)}(nt)} 1_{\{L_{k,j}^{(n)} - (nt - S_{k,j}^{(n)}) \in \sqrt{n}B\}} \qquad (4.13) \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^{\infty} 1_{\{S_{k,j}^{(n)} \in \sqrt{n}B + nt - L_{k,j}^{(n)},\ S_{k,j}^{(n)} \le nt\}}, \\ \widehat{\mathcal{V}}_k^{(n)}(t)(B) &\triangleq \frac{1}{\sqrt{n}} \mathcal{V}_k^{(n)}(nt)(\sqrt{n}\, B) \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^{A_k^{(n)}(nt)} v_{k,j}^{(n)} 1_{\{L_{k,j}^{(n)} - (nt - S_{k,j}^{(n)}) \in \sqrt{n}B\}} \qquad (4.14) \\ &= \frac{1}{\sqrt{n}} \sum_{j=1}^{\infty} v_{k,j}^{(n)} 1_{\{S_{k,j}^{(n)} \in \sqrt{n}B + nt - L_{k,j}^{(n)},\ S_{k,j}^{(n)} \le nt\}}. \end{aligned}$$

5 Earliest-Deadline-First Limits

We begin with an informal discussion of the limiting lead-time profile when customers are served according to the earliest-deadline-first discipline. At each time t, there are potentially two types of customers in queue k: those whose lead times are in $\big[C_k^{(n)}(t), F_k^{(n)}(t)\big)$ and those whose lead times are in $\big[F_k^{(n)}(t), \infty\big)$. Customers of the former type are present only under unusual circumstances. Either there must be a customer present with lead time less than $\sqrt{n}\, y_k^* - t$, an unlikely event for t of order n, or else a customer in service, the customer we call $\mathcal{C}$, whose lead time is $F_k^{(n)}(t)$ at time t, must have been pre-empted by a customer with smaller lead time. When $\mathcal{C}$ is pre-empted, the pre-empting customer is the only customer with lead time in $\big[C_k^{(n)}(t), F_k^{(n)}(t)\big)$, but additional customers may later arrive, sustaining a period of time before service of $\mathcal{C}$ is resumed. These later customers must have lead times less than or equal to $F_k^{(n)}(t)$, which in turn is strictly less than $\sqrt{n}\, y_k^*$. Hence, the arrival intensity for these customers is strictly less than $\lambda_k^{(n)}$, although they have the full attention of the server. This causes the period of time before service of $\mathcal{C}$ is resumed to be brief and suggests that few customers ever have lead times in $\big[C_k^{(n)}, F_k^{(n)}\big)$. This intuition is made precise by Lemma 5.4 below.

Ignoring customers with lead times in $\big[C_k^{(n)}(t), F_k^{(n)}\big)$, we work with the lead-time distribution of customers in queue k by examing those with lead times in $\big[F_k^{(n)}(t), \infty\big)$. Except possibly for a customer with lead time $F_k^{(n)}(t)$, these customers have never been in service. In order for one of these customers to have lead time $y > F_k^{(n)}(t)$, it must arrive at some time $t-s$ prior to t and be assigned lead time $y+s$ upon arrival. If G_k has a density, then the density of the assigned lead time distribution is $\frac{1}{\sqrt{n}} G_k'\big(\frac{y+s}{\sqrt{n}}\big)$ (see (2.4)), and multiplying by the arrival rate $\lambda_k^{(n)}$, we obtain the density of customers with lead time y:

$$\frac{\lambda_k^{(n)}}{\sqrt{n}} \int_0^t G_k'\Big(\frac{y+s}{\sqrt{n}}\Big) ds = \lambda_k^{(n)} \left[G_k\Big(\frac{y+t}{\sqrt{n}}\Big) - G_k\Big(\frac{y}{\sqrt{n}}\Big)\right].$$

The heavy traffic scaling considers the density of $\frac{1}{\sqrt{n}}$ times the actual number of customers whose lead times are $\hat{y} = \frac{y}{\sqrt{n}} > \frac{1}{\sqrt{n}} F_k^{(n)}(t) = \widehat{F}_k^{(n)}(\hat{t})$ at scaled time $\hat{t} = \frac{t}{n}$. This density is

$$\frac{\lambda_k^{(n)}}{\sqrt{n}} [G_k(\hat{y} + \sqrt{n}\,\hat{t}) - G_k(\hat{y})] \frac{dy}{d\hat{y}} = \lambda_k^{(n)} [G_k(\hat{y} + \sqrt{n}\,\hat{t}) - G_k(\hat{y})],$$

and as $n \to \infty$, it converges to $\lambda_k[1 - G_k(\hat{y})]$. The scaled number of customers in queue k at time t (scaled time $\hat{t}$) is thus approximately equal to

$$\lambda_k \int_{\widehat{F}_k^{(n)}(\hat{t})}^{\infty} \big(1 - G_k(\eta)\big)\, d\eta,$$

and since the average work per customer is $\frac{1}{\mu_k^{(n)}}$, the scaled work in queue can be approximated by

$$\rho_k \int_{\widehat{F}_k^{(n)}(\hat{t})}^{\infty} \big(1 - G_k(\eta)\big)\, d\eta.$$

As $n \to \infty$, this should converge to $W_k^*(t)$ of Assumption 3.4, and this gives us an equation which characterizes $F_k^* = \lim_{n\to\infty} \widehat{F}_k^{(n)}$:

$$W_k^* = \rho_k \int_{F_k^*}^{\infty} \big(1 - G_k(\eta)\big)\, d\eta.$$

These considerations prompt the following definitions.

For $k = 1, \ldots, K$, we set

$$H_k(y) \triangleq \int_y^{\infty} \big(1 - G_k(\eta)\big)\, d\eta = \begin{cases} \int_y^{y_k^*} \big(1 - G_k(\eta)\big)\, d\eta, & \text{if } y \le y_k^*, \\ 0, & \text{if } y > y_k^*. \end{cases} \tag{5.1}$$

The function H_k maps $(-\infty, y_k^*]$ onto $[0,\infty)$ and is strictly decreasing and Lipschitz continuous with Lipschitz constant 1 on $(-\infty, y_k^*]$. Therefore, there exists a continuous inverse function H_k^{-1} which maps $[0,\infty)$ onto $(-\infty, y_k^*]$. We next define what we shall show is the *limiting scaled frontier process*

$$F_k^*(t) \triangleq H_k^{-1}\Big(\frac{W_k^*(t)}{\rho_k}\Big), \quad t \geq 0, \tag{5.2}$$

where W_k^* is as in Assumption 3.4. Let $\widehat{F}^{(n)} = \big(\widehat{F}_1^{(n)}, \dots, \widehat{F}_K^{(n)}\big)$ and $F^* = (F_1^*, \dots, F_K^*)$. Denote by $\mathcal{M}$ the set of all finite, nonnegative measures on $\mathcal{B}(\mathbb{R})$, the Borel subsets of $\mathbb{R}$, and denote by $\mathcal{M}^K$ the K-fold product of $\mathcal{M}$. Under the weak topology, $\mathcal{M}$ is a separable, metrizable topological space, and so is $\mathcal{M}^K$.

Under Assumptions 2.1 and 3.4, we have the following generalizations of Proposition 3.10, Theorem 3.1 and Corollary 3.2 of [9]. The proofs are straightforward generalizations of those given in [9], with the exception of Lemma 5.4 below.

Proposition 5.1 *Under the Earliest-Deadline-First queue discipline, we have* $\widehat{F}^{(n)} \Rightarrow F^*$.

Theorem 5.2 *For $k = 1, \dots, K$, define measure-valued processes $\widehat{\mathcal{W}}_k^*$ and $\widehat{\mathcal{Q}}_k^*$ by*

$$\widehat{\mathcal{W}}_k^*(t)(B) \triangleq \rho_k \int_{B\cap[F_k^*(t),\infty)} \big(1 - G_k(y)\big)\, dy, \quad \widehat{\mathcal{Q}}_k^*(t)(B) \triangleq \mu_k \widehat{\mathcal{W}}_k^*(t) \tag{5.3}$$

for all Borel sets $B \subset \mathbb{R}$. Under the Earliest-Deadline-First queue discipline, the processes $\widehat{\mathcal{W}}^{(n)} = \big(\widehat{\mathcal{W}}_1^{(n)}, \dots, \widehat{\mathcal{W}}_K^{(n)}\big)$ and $\widehat{\mathcal{Q}}^{(n)} = \big(\widehat{\mathcal{Q}}_1^{(n)}, \dots, \widehat{\mathcal{Q}}_K^{(n)}\big)$ converge weakly in $D\big([0,\infty), \mathcal{M}^K\big)$ to $\widehat{\mathcal{W}}^* = \big(\widehat{\mathcal{W}}_1^*, \dots, \widehat{\mathcal{W}}_K^*\big)$ and $\widehat{\mathcal{Q}}^* = \big(\widehat{\mathcal{Q}}_1^*, \dots, \widehat{\mathcal{Q}}_K^*\big)$, respectively.

Corollary 5.3 *Under the Earliest-Deadline-First queue discipline, the K-dimensional scaled queue length process $\widehat{Q}^{(n)} = \big(\widehat{Q}_1^{(n)}, \dots, \widehat{Q}_K^{(n)}\big)$ converges weakly to the process $\big(\mu_1 W_1^*, \dots, \mu_K W_K^*\big)$.*

To extend the arguments of [9] to prove the above results, only the generalization of Proposition 3.6 of [9] needs more attention. We state and prove the relevant result.

Lemma 5.4 *Under the Earliest-Deadline-First queue discipline, we have for $k = 1, \dots, K$,*

$$\widehat{\mathcal{W}}_k^{(n)}[\widehat{C}_k^{(n)}, \widehat{F}_k^{(n)}) \Rightarrow 0, \quad \widehat{\mathcal{Q}}_k^{(n)}[\widehat{C}_k^{(n)}, \widehat{F}_k^{(n)}) \Rightarrow 0. \tag{5.4}$$

Proof: We fix k and prove convergence on $[0,1]$. For $0 \leq t \leq 1$, we define

$$\tau_k^{(n)}(t) \triangleq \sup\{s \in [0,t]; \widehat{C}_k^{(n)}(s) = \widehat{F}_k^{(n)}(s)\}. \tag{5.5}$$

We further define

$$\widehat{D}_k^{(n)}(t) \tag{5.6}$$
$$= \frac{1}{\sqrt{n}} \sum_{j=1}^{\infty} v_{k,j}^{(n)} \mathbb{I}_{\{n\tau_k^{(n)}(t) < S_{k,j}^{(n)} \le nt\}} \mathbb{I}_{\{L_{k,j}^{(n)} - (nt - S_{k,j}^{(n)}) < F_k^{(n)}(n\tau_k^{(n)}(t)) - n(t - \tau_k^{(n)}(t))\}}.$$

This is the scaled work brought by customers who arrive in queue k within the time interval $(n\tau_k^{(n)}(t), nt]$ with lead times at arrival smaller than the current frontier. Because there is work with lead time in $\left[C_k^{(n)}(s), F_k^{(n)}(s)\right)$ for all $s \in \left[n\tau_k^{(n)}, nt\right]$, the effort expended by the server on customers with lead times smaller than the frontier during the period $n\tau_k^{(n)}(t)$ to nt is $T_k^{(n)}(nt) - T_k^{(n)}\left(n\tau_k^{(n)}\right)$, the full entitlement of queue k during this period. The work at time nt with lead time smaller than the frontier is thus the work at time $n\tau_k^{(n)}(t)$ with lead time smaller than the fronter plus the work which arrives with lead time smaller than the frontier minus $T_k^{(n)}(nt) - T_k^{(n)}\left(n\tau_k^{(n)}\right)$. It follows that

$$\begin{aligned} 0 &\le \widehat{\mathcal{W}}_k^{(n)}(t)\left[\widehat{C}_k^{(n)}(t), \widehat{F}_k^{(n)}(t)\right) \\ &= \widehat{\mathcal{W}}_k^{(n)}\left(\tau_k^{(n)}(t)\right)\left[\widehat{C}_k^{(n)}(\tau_k^{(n)}(t)), \widehat{F}_k^{(n)}(\tau_k^{(n)}(t))\right) + \widehat{D}_k^{(n)}(t) \\ &\quad - \frac{1}{\sqrt{n}}\left[T_k^{(n)}(nt) - T_k^{(n)}\left(n\tau_k^{(n)}(t)\right)\right]. \end{aligned} \tag{5.7}$$

We examine each of the three terms on the right-hand side of (5.7).

By definition $\widehat{\mathcal{W}}_k^{(n)}\left(\tau_k^{(n)}(t) - \right)\left[\widehat{C}_k^{(n)}(\tau_k^{(n)}(t)-), \widehat{F}_k^{(n)}(\tau_k^{(n)}(t)-)\right) = 0$, and so

$$\widehat{\mathcal{W}}_k^{(n)}\left(\tau_k^{(n)}(t)\right)\left[\widehat{C}_k^{(n)}(\tau_k^{(n)}(t)), \widehat{F}_k^{(n)}(\tau_k^{(n)}(t))\right) \le j(\widehat{W}_k^{(n)}), \tag{5.8}$$

where $j(x) = \sup_{0 \le s \le 1} |x(s) - x(s-)|$ for $x \in D[0,1]$. In light of Corollary 3.5, we conclude that

$$\widehat{\mathcal{W}}_k^{(n)}\left(\tau_k^{(n)}(t)\right)\left[\widehat{C}_k^{(n)}(\tau_k^{(n)}(t)), \widehat{F}_k^{(n)}(\tau_k^{(n)}(t))\right) \Rightarrow 0, \tag{5.9}$$

where the convergence is that of RCLL processes defined on $[0,1]$.

In contrast to the proof of Proposition 3.6 in [9], we cannot assert that the workload in queue k is being decreased at rate 1 on the time interval $(n\tau_k^{(n)}(t), nt]$. Indeed, $\dot{T}_k^{(n)}$ is, in general, not equal to 1. However, on the interval under consideration, the k-th queue is never empty, so

$$U_k^{(n)}(nt) = U_k^{(n)}\left(n\tau_k^{(n)}(t)\right). \tag{5.10}$$

By (2.15) and (2.18), we have

$$T_k^{(n)}(s) = V_k^{(n)}\left(A_k^{(n)}(s)\right) - W_k^{(n)}(s) + U_k^{(n)}(s), \quad s \ge 0. \tag{5.11}$$

Thus, the amount by which the workload in queue k has been decreased on $(n\tau_k^{(n)}(t), nt]$ due to the activity of the server, divided by the scaling factor $\sqrt{n}$, is

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\left[T_k^{(n)}(nt)-T_k^{(n)}\left(n\tau_k^{(n)}(t)\right)\right]\\
&\qquad= \frac{1}{\sqrt{n}}\left[V_k^{(n)}\left(A_k^{(n)}(nt)\right)-V_k^{(n)}\left(A_k^{(n)}\left(n\tau_k^{(n)}(t)\right)\right)\right] \qquad (5.12)\\
&\qquad\qquad -\frac{1}{\sqrt{n}}\left[W_k^{(n)}(nt)-W_k^{(n)}\left(n\tau_k^{(n)}(t)\right)\right]\\
&\qquad= \widehat{Z}_k^{(n)}(t)-\widehat{Z}_k^{(n)}\left(\tau_k^{(n)}(t)\right)-\left[\widehat{W}_k^{(n)}(t)-\widehat{W}_k^{(n)}\left(\tau_k^{(n)}(t)\right)\right]\\
&\qquad\qquad +\rho_k^{(n)}\sqrt{n}\left(t-\tau_k^{(n)}(t)\right).
\end{aligned}
$$

Because $0 \le \tau_k^{(n)}(t) \le t$, we have

$$
\begin{aligned}
\min_{0\le s\le t}\widehat{Z}_k^{(n)}(s) \;&\le\; \widehat{Z}_k^{(n)}\left(\tau_k^{(n)}(t)\right) \;\le\; \max_{0\le s\le t}\widehat{Z}_k^{(n)}(s), && (5.13)\\
\min_{0\le s\le t}\widehat{Z}_{k,y}^{(n)}(s) \;&\le\; \widehat{Z}_{k,y}^{(n)}\left(\tau_k^{(n)}(t)\right) \;\le\; \max_{0\le s\le t}\widehat{Z}_{k,y}^{(n)}(s), && (5.14)\\
\min_{0\le s\le t}\widehat{W}_k^{(n)}(s) \;&\le\; \widehat{W}_k^{(n)}\left(\tau_k^{(n)}(t)\right) \;\le\; \max_{0\le s\le t}\widehat{W}_k^{(n)}(s). && (5.15)
\end{aligned}
$$

We conclude from Corollary 3.2, Assumption 3.4 and the Continuity Theorem that the upper and lower bounds in (5.13), (5.14) and (5.15) have weak limits, and hence

$$
\frac{1}{\sqrt{n}}\left[T_k^{(n)}(nt)-T_k^{(n)}\left(n\tau_k^{(n)}(t)\right)\right]=\rho_k^{(n)}\sqrt{n}\left(t-\tau_k^{(n)}(t)\right)+O(1). \qquad (5.16)
$$

Finally, to study $\widehat{D}_k^{(n)}(t)$, we choose $y \le y_k^*$ and consider two cases.
Case I: $n\tau_k^{(n)}(t)+\sqrt{n}\,(y_k^*-y) \le nt$.

In this case, on the event $\{n\tau_k^{(n)}(t)+\sqrt{n}\,(y_k^*-y) < S_{k,j}^{(n)} \le nt\}$, the inequality $L_{k,j}^{(n)}-(nt-S_{k,j}^{(n)}) < F_k^{(n)}\left(n\tau_k^{(n)}(t)\right)-n(t-\tau_k^{(n)}(t))$ implies the inequality $L_{k,j}^{(n)} \le \sqrt{n}\,y$. Therefore,

$$
\begin{aligned}
&\widehat{D}_k^{(n)}(t)\\
&\quad\le \frac{1}{\sqrt{n}}\sum_{j=1}^{\infty} v_{k,j}^{(n)}\mathbb{I}_{\{n\tau_k^{(n)}(t)<S_{k,j}^{(n)}\le n\tau_k^{(n)}(t)+\sqrt{n}\,(y_k^*-y)\}}\\
&\qquad +\frac{1}{\sqrt{n}}\sum_{j=1}^{\infty} v_{k,j}^{(n)}\mathbb{I}_{\{n\tau_k^{(n)}(t)+\sqrt{n}\,(y_k^*-y)<S_{k,j}^{(n)}\le nt\}}\mathbb{I}_{\{L_{k,j}^{(n)}\le\sqrt{n}\,y\}} \qquad (5.17)\\
&\quad= \frac{1}{\sqrt{n}}\left[V_k^{(n)}\left(A_k^{(n)}\left(n\tau_k^{(n)}(t)+\sqrt{n}\,(y_k^*-y)\right)\right)-V_k^{(n)}\left(A_k^{(n)}\left(n\tau_k^{(n)}(t)\right)\right)\right]\\
&\qquad +\frac{1}{\sqrt{n}}\left[V_{k,y}^{(n)}\left(A_k^{(n)}(nt)\right)-V_{k,y}^{(n)}\left(A_k^{(n)}\left(n\tau_k^{(n)}(t)+\sqrt{n}\,(y_k^*-y)\right)\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
= \quad & \widehat{Z}_k^{(n)}\big(\tau_k^{(n)}(t)+\frac{1}{\sqrt{n}}(y_k^*-y)\big)-\widehat{Z}_k^{(n)}\big(\tau_k^{(n)}(t)\big)+\rho_k^{(n)}(y_k^*-y)\\
& +\widehat{Z}_{k,y}^{(n)}(t)-\widehat{Z}_{k,y}^{(n)}\big(\tau_k^{(n)}(t)+\frac{1}{\sqrt{n}}(y_k^*-y)\big)\\
& +G_k(y)\rho_k^{(n)}\sqrt{n}\,\big(t-\tau_k^{(n)}(t)\big)-G_k(y)\rho_k^{(n)}(y_k^*-y).
\end{aligned}
$$

Thus, by (5.13), (5.14),

$$
\widehat{D}_k^{(n)}(t)\le G_k(y)\rho_k^{(n)}\sqrt{n}\,\big(t-\tau_k^{(n)}(t)\big)+O(1). \tag{5.18}
$$

We now use (5.18) and (5.16) to conclude from (5.7) that

$$
\begin{aligned}
0 \quad \le \quad & \widehat{\mathcal{W}}_k^{(n)}\big(\tau_k^{(n)}(t)\big)\big[\widehat{C}_k^{(n)}(\tau_k^{(n)}(t)),\widehat{F}_k^{(n)}(\tau_k^{(n)}(t))\big)\\
& -\big(1-G_k(y)\big)\rho_k^{(n)}\sqrt{n}\,\big(t-\tau_k^{(n)}(t)\big)+O(1),
\end{aligned} \tag{5.19}
$$

which implies because of (5.9), that

$$
\big(1-G_k(y)\big)\rho_k^{(n)}\big(t-\tau_k^{(n)}(t)\big)\le O\left(\frac{1}{\sqrt{n}}\right). \tag{5.20}
$$

Case II: $n\tau_k^{(n)}(t)+\sqrt{n}\,(y_k^*-y)>nt$.
In this case, $t-\tau_k^{(n)}(t)\le\frac{1}{\sqrt{n}}(y_k^*-y)$, and again (5.20) holds.

We choose $y<y_k^*$ so that $1-G_k(y)>0$ and conclude that

$$
\tau_k^{(n)}(t)\Rightarrow t, \tag{5.21}
$$

where the convergence is that of RCLL processes on $[0,1]$.

We now use (5.8), (5.12) and (5.17) in (5.7) to obtain the more precise estimate

$$
\begin{aligned}
&\big(1-G_k(y)\big)\rho_k^{(n)}\sqrt{n}\,\big(t-\tau_k^{(n)}(t)\big)\\
&\le \quad j\big(\widehat{W}_k^{(n)}\big)+\Big[\widehat{Z}_k^{(n)}\big(\tau_k^{(n)}(t)+\frac{1}{\sqrt{n}}(y_k^*-y)\big)-\widehat{Z}_k^{(n)}\big(\tau_k^{(n)}(t)\big)\Big]\\
&\qquad +\Big[\widehat{Z}_{k,y}^{(n)}(t)-\widehat{Z}_{k,y}^{(n)}\big(\tau_k^{(n)}(t)+\frac{1}{\sqrt{n}}(y_k^*-y)\big)\Big]+\Big[\widehat{W}_k^{(n)}(t)-\widehat{W}_k^{(n)}\big(\tau_k^{(n)}(t)\big)\Big]\\
&\qquad -\Big[\widehat{Z}_k^{(n)}(t)-\widehat{Z}_k^{(n)}\big(\tau_k^{(n)}(t)\big)\Big]+\big(1-G_k(y)\big)\rho_k^{(n)}(y_k^*-y)
\end{aligned} \tag{5.22}
$$

in Case I, and the companion estimate $\sqrt{n}\big(t-\tau_k^{(n)}(t)\big)\le(y_k^*-y)$ in Case II. According to the Differencing Theorem (e.g., Theorem A.3 of [9]), every term on the right-hand side of (5.22), save the last, converges weakly to 0. Consequently,

$$
\Big(\sqrt{n}\,\big(t-\tau_k^{(n)}(t)\big)-(y_k^*-y)\Big)^+\Rightarrow 0. \tag{5.23}
$$

Because $y < y_k^*$ is arbitrary and $t - \tau_k^{(n)}(t) \geq 0$, we conclude

$$\sqrt{n}\left(t - \tau_k^{(n)}(t)\right) \Rightarrow 0. \tag{5.24}$$

Armed with (5.24), we return once again to (5.17), this time taking $y = y_k^*$ so that Case II is vacuous and concluding that $\widehat{D}_k^{(n)} \Rightarrow 0$. ¿From (5.7), using (5.8), we see that

$$\widehat{\mathcal{W}}_k^{(n)}(t)\left[\widehat{C}_k^{(n)}(t), \widehat{F}_k^{(n)}(t)\right) \leq j\left(W_k^{(n)}\right) + \widehat{D}_k^{(n)}(t), \tag{5.25}$$

and hence $\widehat{\mathcal{W}}_k^{(n)}[\widehat{C}_k^{(n)}, \widehat{F}_k^{(n)}) \Rightarrow 0$. For the second part of (5.4), we observe that

$$\begin{aligned} &\widehat{\mathcal{Q}}_k^{(n)}(t)\left[\widehat{C}_k^{(n)}(t), \widehat{F}_k^{(n)}(t)\right) \\ &\quad \leq \frac{1}{\sqrt{n}}\left[1 + A_k^{(n)}(nt) - A_k^{(n)}\left(n\tau_k^{(n)}(t)\right)\right] \\ &\quad = \frac{1}{\sqrt{n}} + \left[\widehat{A}_k^{(n)}(t) - \widehat{A}_k^{(n)}\left(\tau_k^{(n)}(t)\right)\right] + \lambda_k^{(n)}\sqrt{n}\left(t - \tau_k^{(n)}(t)\right), \end{aligned} \tag{5.26}$$

and all terms on the right-hand side have limit zero. □

We may now follow the proofs of Section 3 of [9], obtaining

$$\left(\widehat{W}_1^{(n)}[\widehat{F}_1^{(n)}, \infty), \ldots, \widehat{W}_K^{(n)}[\widehat{F}_K^{(n)}, \infty)\right) \Rightarrow (W_1^*, \ldots, W_K^*) \tag{5.27}$$

in place of (3.27) of [9], using the continuity of the mapping

$$H^{-1}\colon [0,\infty)^K \overset{onto}{\longrightarrow} \prod_{k=1}^{K}(-\infty, y_k^*] \tag{5.28}$$

defined by

$$H^{-1}(x_1, \ldots, x_K) \triangleq \left(H_1^{-1}(x_1), \ldots, H_K^{-1}(x_K)\right) \tag{5.29}$$

to modify the proof of Proposition 3.10 of [9], and replacing $\psi\colon \mathbb{R} \to \mathcal{M}$ in the proof of Theorem 3.1 of [9] by $\psi\colon \mathbb{R}^K \to \mathcal{M}^K$ given by

$$\begin{aligned} &\psi(x_1, \ldots, x_K)(B_1, \ldots, B_K) \\ &\quad \triangleq \left(\int_{B_1 \cap [x_1, \infty)} \left(1 - G_1(\eta)\right) d\eta, \ldots, \int_{B_K \cap [x_K, \infty)} \left(1 - G_K(\eta)\right) d\eta\right). \end{aligned} \tag{5.30}$$

6 First-In-First-Out Limits

In this section we assume that customers within each of the K queues are served using the First-In-First-Out (FIFO) queue discipline, i.e., the server always services the customer with longest time in queue. This discipline is equivalent to EDF if all customers in each queue have initial lead times 0.

As we remarked at the end of Section 2, all the processes introduced in that section, with the exception of the queue lengths, are independent of the service discipline within each queue, provided that the server always allocates its total capacity in the same way among all queues which have work to do (i.e., the processes $w_k^{(n)}$ defined in Section 2 remain unchanged). Thus, in this section we adopt all the notation and the assumptions of Sections 2 and 3. We still assume that initial lead times are distributed as in (2.4), although we shall sometimes use the results of the previous section in the case that all initial lead times are zero as a way of justifying assertions about the limiting process under FIFO. Assumptions 2.1 and 3.4 are still in force.

The aim of this section is to give a counterpart of Theorem 5.2 for the FIFO scheduling policy. We will start with the case of a single queue ($K = 1$) in which the notation somewhat simplifies. In particular, we drop all the subscripts k in G_k, y_k^*, $v_{k,j}^{(n)}$, etc., when a single queue is considered. It turns out that the ideas developed for this case can be easily generalized to $K \geq 2$ arrival streams.

Denote by W^* the limiting (real-valued) workload in the queue, i.e., the limit of $\widehat{W}^{(n)}$. W^* is the Brownian motion with drift described in Corollary 3.3 (with $K = 1$).

The lead time of a customer in queue is the sum of two independent random variables, the lead time assigned upon arrival and the negative of the time since arrival. If the former is always zero, then the FIFO queue discipline coincides with EDF, and Theorem 5.2 implies that

$$\widehat{\mathcal{W}}^*(t)(B) = \int_{B\cap[F^*(t),\infty)} \big(1 - \mathbb{I}_{[0,\infty)}(y)\big)\, dy = m\big(B \cap [F^*(t), 0]\big),$$

where m denotes Lebesgue measure. In particular, $W^*(t) = \widehat{\mathcal{W}}^*(t)(\mathbb{R}) = -F^*(t)$. In other words, $\widehat{\mathcal{W}}^*(t)$ is Lebesgue measure restricted to $[-W^*(t), 0]$.

When the lead times assigned upon arrival are nontrivial, it is reasonable to expect $\widehat{\mathcal{W}}^*(t)$ to be the convolution of Lebesgue measure on $[-W^*(t), 0]$ with dG, the distribution of scaled lead times. The following theorem confirms this conjecture.

Theorem 6.1 *Let $\widehat{\mathcal{W}}^*$ be the measure-valued process defined by*

$$\widehat{\mathcal{W}}^*(t)(B) \triangleq \int_B (G(\eta + W^*(t)) - G(\eta))d\eta,$$

i.e., the process whose value is the convolution of dG with the uniform distribution on $[-W^*(t), 0]$. Let also $\widehat{\mathcal{Q}}^* \triangleq \lambda\widehat{\mathcal{W}}^*$. In the case of a single queue, under the First-In-First-Out queue discipline, we have

$$\widehat{\mathcal{W}}^{(n)} \Rightarrow \widehat{\mathcal{W}}^*, \quad \widehat{\mathcal{Q}}^{(n)} \Rightarrow \widehat{\mathcal{Q}}^*. \tag{6.31}$$

Let us remark that this result was conjectured in [9].

PROOF: For $y \le 0$, $l \in \mathbb{R}$, define

$$\mathcal{M}^{(n)}(t)([y,0]\times(-\infty,l]) \triangleq \sum_{j=1}^{A^{(n)}(t)} v_j^{(n)} \mathbb{I}_{\{y \le S_j^{(n)} - t\}} \mathbb{I}_{\{L_j^{(n)} \le l\}}.$$

This is the arrived work (including departed work) whose time in queue is less than or equal to $-y$ and whose lead time upon arrival was less than or equal to l. Similarly, for $y \le 0$, $l \in \mathbb{R}$, let $\mathcal{N}^{(n)}(t)([y,0]\times(-\infty,l])$ denote the work still present in queue whose time in queue is less than or equal to $-y$ and whose lead time upon arrival was less than or equal to l. The scaled versions of these two processes are

$$\widehat{\mathcal{N}}^{(n)}(t)([y,0]\times(-\infty,l]) \triangleq \frac{1}{\sqrt{n}}\mathcal{N}^{(n)}(nt)([\sqrt{n}y,0]\times(-\infty,\sqrt{n}l])$$

and

$$\begin{aligned}\widehat{\mathcal{M}}^{(n)}(t)([y,0]\times(-\infty,l]) &\triangleq \frac{1}{\sqrt{n}}\mathcal{M}^{(n)}(nt)([\sqrt{n}y,0]\times(-\infty,\sqrt{n}l]) \\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^{\infty} v_j^{(n)} \mathbb{I}_{\{\sqrt{n}y+nt \le S_j^{(n)} \le nt\}} \mathbb{I}_{\{L_j^{(n)} \le \sqrt{n}l\}}.\end{aligned}$$

We have

$$\widehat{\mathcal{M}}^{(n)}(t)([y,0]\times(-\infty,l]) = \widehat{Y}_m^{(n)}(t) + \widehat{U}_m^{(n)}(t),$$

where

$$\begin{aligned}\widehat{Y}_m^{(n)}(t) &= \frac{1}{\sqrt{n}}\sum_{j=1}^{\infty}\left[v_j^{(n)}\mathbb{I}_{\{L_j^{(n)} \le \sqrt{n}l\}} - \frac{1}{\mu^{(n)}}G(l)\right]\mathbb{I}_{\{\sqrt{n}y+nt \le S_j^{(n)} \le nt\}}, \\ \widehat{U}_m^{(n)}(t) &= \frac{G(l)}{\mu^{(n)}\sqrt{n}}\sum_{j=1}^{\infty}\mathbb{I}_{\{\sqrt{n}y+nt \le S_j^{(n)} \le nt\}}.\end{aligned}$$

Recall that, by Theorem 3.1, the process $\widehat{V}_l^{(n)}$ defined by (3.9) (recall that we drop the subscript $k=1$) converges weakly to a Brownian motion V_l^*. Furthermore,

$$\widehat{Y}^{(n)}(t) = \widehat{V}_l^{(n)}\left(\frac{1}{n}A^{(n)}(nt)\right) - \widehat{V}_l^{(n)}\left(\frac{1}{n}A^{(n)}((nt+\sqrt{n}y)^+-)\right).$$

But

$$\widehat{A}^{(n)}(t) = \frac{1}{\sqrt{n}}[A^{(n)}(nt) - \lambda^{(n)}nt],$$

so

$$\frac{1}{n}A^{(n)}(nt) = \frac{1}{\sqrt{n}}\widehat{A}^{(n)}(t) + \lambda^{(n)}t \Rightarrow \lambda t,$$

$$\frac{1}{n}A^{(n)}((nt+\sqrt{n}y)^+-) = \frac{1}{\sqrt{n}}\widehat{A}^{(n)}\left(\left(t+\frac{y}{\sqrt{n}}\right)^+-\right) + \lambda^{(n)}\left(t+\frac{y}{\sqrt{n}}\right)^+ \Rightarrow \lambda t.$$

The Differencing Theorem implies that $\widehat{Y}^{(n)} \Rightarrow 0$.

On the other hand,

$$\begin{aligned}
\widehat{U}_m^{(n)}(t) &= \frac{G(l)}{\mu^{(n)}\sqrt{n}}\left[A^{(n)}(nt) - A^{(n)}((nt+\sqrt{n}y)^+-)\right] \\
&= \frac{G(l)}{\mu^{(n)}}\left[\widehat{A}^{(n)}(nt) + \lambda^{(n)}\sqrt{n}t\right] \\
&- \frac{G(l)}{\mu^{(n)}}\left[\widehat{A}^{(n)}\left(\left(t+\frac{y}{\sqrt{n}}\right)^+-\right) + \lambda^{(n)}\sqrt{n}\left(t+\frac{y}{\sqrt{n}}\right)^+\right] \\
&= \frac{G(l)}{\mu^{(n)}}\left[\widehat{A}^{(n)}(t) - \widehat{A}^{(n)}\left(\left(t+\frac{y}{\sqrt{n}}\right)^+-\right)\right] \\
&+ G(l)\left(\frac{\lambda^{(n)}}{\mu^{(n)}}-1\right)\left[\sqrt{n}t - (\sqrt{n}t+y)^+\right] \qquad (6.32) \\
&+ G(l)\left[\sqrt{n}t - (\sqrt{n}t+y)^+\right].
\end{aligned}$$

The first term on the right-hand side of (6.32) has limit zero. Indeed,

$$\widehat{A}^{(n)}(t) - \widehat{A}^{(n)}\left(\left(t+\frac{y}{\sqrt{n}}\right)^+\right) \Rightarrow 0$$

by the Differencing Theorem and

$$\widehat{A}^{(n)}\left(\left(t+\frac{y}{\sqrt{n}}\right)^+\right) - \widehat{A}^{(n)}\left(\left(t+\frac{y}{\sqrt{n}}\right)^+-\right) \Rightarrow 0$$

because $\widehat{A}^{(n)}$ converges to a continuous limit and hence the maximal jump of $A^{(n)}$ on any finite time horizon converges to 0 (compare the proof of Corollary 3.5). The second term on the right-hand side of (6.32) has limit zero because $\frac{\lambda^{(n)}}{\mu^{(n)}} \to 1$ and $0 \leq \sqrt{n}t - (\sqrt{n}t+y)^+ \leq -y$.

Thus, we have proved that, for every $T > 0$, $y \leq 0$ and $l \in \mathbb{R}$, we have

$$\sup_{0\leq t\leq T}\left|\mathcal{M}^{(n)}(t)([y,0]\times(-\infty,l]) + \left((\sqrt{n}t+y)^+ - \sqrt{n}t\right)G(l)\right| \xrightarrow{P} 0. \qquad (6.33)$$

Clearly, $(\sqrt{n}t+y)^+ - \sqrt{n}t$ is Lipschitz continuous in y (with the Lipschitz constant 1). Moreover, both $\mathcal{M}^{(n)}(t)([y,0]\times(-\infty,l])$ and $((\sqrt{n}t+y)^+ - \sqrt{n}t)\,G(l)$ are monotone in y. Thus, by the same argument as in Proposition 3.4 of [9], we can upgrade (6.33) to

$$\sup_{y_0\leq y\leq 0}\sup_{0\leq t\leq T}\left|\mathcal{M}^{(n)}(t)([y,0]\times(-\infty,l]) + \left((\sqrt{n}t+y)^+ - \sqrt{n}t\right)G(l)\right| \xrightarrow{P} 0 \qquad (6.34)$$

for every $T > 0$, $y_0 \le 0$ and $l \in \mathbb{R}$. In particular, taking $l = y^*$ in (6.34) we get, by (2.4), (2.5),

$$\sup_{y_0 \le y \le 0} \sup_{0 \le t \le T} \left| \mathcal{M}^{(n)}(t)([y,0] \times \mathbb{R}) + (\sqrt{n}t + y)^+ - \sqrt{n}t \right| \xrightarrow{P} 0. \tag{6.35}$$

A slight modification of the above argument yields

$$\sup_{y_0 \le y \le 0} \sup_{0 \le t \le T} \left| \mathcal{M}^{(n)}(t)([y,0] \times (-\infty, l)) + \left((\sqrt{n}t + y)^+ - \sqrt{n}t \right) G(l-) \right| \xrightarrow{P} 0 \tag{6.36}$$

for every $T > 0$, $y_0 \le 0$ and $l \in \mathbb{R}$. Using (6.36) and proceeding as in the proof of the Glivenko-Cantelli theorem (see e.g. Theorem 20.6 in [2]), we can finally upgrade (6.34) to

$$\sup_{l \in \mathbb{R}} \sup_{y_0 \le y \le 0} \sup_{0 \le t \le T} \left| \mathcal{M}^{(n)}(t)([y,0] \times (-\infty, l]) + \left((\sqrt{n}t + y)^+ - \sqrt{n}t \right) G(l) \right| \xrightarrow{P} 0 \tag{6.37}$$

for every $T > 0$, $y_0 \le 0$. Using (6.35) and proceeding as in the proof of Corollary 3.5 of [9], we get

$$\sup_{y_0 \le y \le 0} \sup_{0 \le t \le T} \mathcal{M}^{(n)}(t)(\{y\} \times \mathbb{R}) \xrightarrow{P} 0$$

for every $T > 0$, $y_0 \le 0$.

Let us introduce the notation

$$F^{(n)}(t) \triangleq C^{(n)}(t) \triangleq \left\{ \begin{array}{l} \text{The negative of the time in queue} \\ \text{of the customer currently in service} \\ \text{or } S^{(n)}_{A^{(n)}(t)} - t \text{ if the queue is empty} \end{array} \right\}.$$

Let us notice that $C^{(n)}$, $F^{(n)}$ defined above would coincide with those introduced in Section 4 (and used to study the EDF queue discipline) if all the initial lead-times of the customers in queue were zero. Denote also by $\widehat{C}^{(n)}$, $\widehat{F}^{(n)}$ the rescaled versions of $C^{(n)}$, $F^{(n)}$:

$$\widehat{C}^{(n)}(t) \triangleq \widehat{F}^{(n)}(t) \triangleq \frac{1}{\sqrt{n}} F^{(n)}(nt).$$

By Corollary 3.8 in [9] (applied to the case of zero initial lead times), we have

$$\widehat{\mathcal{N}}^{(n)}(\{\widehat{F}^{(n)}\} \times \mathbb{R}) = \widehat{\mathcal{N}}^{(n)}([\widehat{C}^{(n)}, \widehat{F}^{(n)}] \times \mathbb{R}) \Rightarrow 0.$$

Also, Proposition 3.10 in [9] (or Proposition 5.1 of the present paper) yields

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

By a slight modification of the proof of Proposition 3.11 in [9], we can show that

$$\sup_{l\in\mathbb{R}}\sup_{y\le 0}\sup_{0\le t\le T}\left|\widehat{\mathcal{N}}^{(n)}(t)([y,0]\times(-\infty,l])+(y\vee\widehat{F}^{(n)}(t))G(l)\right|\xrightarrow{P}0.$$

But

$$(y\vee\widehat{F}^{(n)}(t))\Rightarrow(y\vee -W^*(t)).$$

Thus, we have proved that $\widehat{\mathcal{N}}^{(n)}(t)$ converges weakly to the product measure with the uniform distribution on $[-W^*(t),0]$ in one coordinate and the cumulative distribution function G in the other one. Let us recall that the first coordinate indicates the *negative* of the time in queue and the second one the initial lead-times of customers.

The lead time of a customer is equal to the sum of its initial lead time and the negative of the time it has already spent in queue. The mapping from $(-\infty,0]\times\mathbb{R}$ into $\mathbb{R}$ given by $(y,l)\to l+y$ induces a continuous function from the space of measures on $(-\infty,0]\times\mathbb{R}$ to measures on the real line. Thus, we get the first part of (6.31). The proof of the second part is analogous, with $v_j^{(n)}$ and $\mu^{(n)}$ replaced by 1. □

As in the case of the EDF service discipline, the above result can be easily generalized to the case of K arrival processes. The corresponding result is

Theorem 6.2 *Let $\widehat{\mathcal{W}}_k^*$ be the measure-valued process defined by*

$$\widehat{\mathcal{W}}_k^*(t)(B)\triangleq\rho_k\int_B\left(G_k\left(\eta+\frac{W_k^*(t)}{\rho_k}\right)-G_k(\eta)\right)d\eta,\qquad(6.38)$$

i.e., the process whose distribution is the convolution of dG_k and the uniform distribution on $\left[-\frac{W_k^*(t)}{\rho_k},0\right]$ with density ρ_k. Also, let $\widehat{\mathcal{Q}}_k^*\triangleq\mu_k\widehat{\mathcal{W}}_k^*$, $\widehat{\mathcal{W}}^*\triangleq(\widehat{\mathcal{W}}_1^*,...,\widehat{\mathcal{W}}_K^*)$, $\widehat{\mathcal{Q}}^*\triangleq(\widehat{\mathcal{Q}}_1^*,...,\widehat{\mathcal{Q}}^*)$. We have, in the case of the First-In-First-Out queue discipline,

$$(\widehat{\mathcal{W}}_1^{(n)},...,\widehat{\mathcal{W}}_K^{(n)})\Rightarrow\widehat{\mathcal{W}}^*,\qquad(\widehat{\mathcal{Q}}_1^{(n)},...,\widehat{\mathcal{Q}}_K^{(n)})\Rightarrow\widehat{\mathcal{Q}}^*.$$

The proof of Theorem 6.2 is an easy generalization of the argument proving Theorem 6.1, where the necessary results from [9] are replaced by their K-dimensional counterparts discussed in Section 5. In particular, Proposition 5.1 should be used in place of Proposition 3.10 from [9].

Corollary 6.3 *Under the First-In-First-Out queue discipline, the K-dimensional scaled queue length process $(\widehat{Q}_k^{(n)})_{k=1,\dots,K}$ converges weakly to $(\mu_k W_k^*)_{k=1,\dots,K}$.*

Corollary 6.3 is an example of "state space collapse," a notion introduced by Reiman [17] to capture the idea that one limiting process (scaled queue length) is a deterministic function of another limiting process (scaled workload). Corollary 6.3 is close to the results of Bramson [3] and Williams [19], who obtain the relationship between individual queue lengths and total workload under head of the line proportional processor sharing, but not the weighted processor sharing of this paper.

7 Simulations

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 Sections 2 and 3. We denoted the queue length for the k-th input stream in the n-th queueing system by $Q_k^{(n)}$, and the scaled queue length by $\widehat{Q}_k^{(n)} \triangleq \frac{1}{\sqrt{n}} Q_k^{(n)}$. For large values of n, under both the earliest-deadline-first (EDF) and the first-in-first-out (FIFO) discipline, $\widehat{Q}_k^{(n)}$ is approximately equal to $Q_k^* = \mu_k W_k^*$ (Corollaries 5.3, 6.3), where W_k^* is the limit of $\widehat{W}_k^{(n)}$ as in Assumption 3.4. Similarly, the measure-valued processes $\widehat{\mathcal{Q}}_k^{(n)}$ converges to $\mu_k \widehat{\mathcal{W}}_k^*$, where (recalling equations from earlier sections)

$$\widehat{\mathcal{W}}_k^*(t)(B) \triangleq \rho_k \int_{B \cap [F_k^*(t), \infty)} \big(1 - G_k(y)\big) dy, \tag{5.3}$$

under EDF (Theorem 5.2), and

$$\widehat{\mathcal{W}}_k^*(t)(B) \triangleq \rho_k \int_B \left(G_k \left(\eta + \frac{W_k^*(t)}{\rho_k} \right) - G_k(\eta) \right) d\eta \tag{6.38}$$

under FIFO (Theorem 6.2). In (5.3), the process F_k^* is given by

$$F_k^*(t) \triangleq H_k^{-1} \left(\frac{W_k^*(t)}{\rho_k} \right). \tag{5.2}$$

Recall that customers arrive with initial lead-time (deadline) distribution

$$\mathbb{P}\left\{ L_{k,j}^{(n)} \leq \sqrt{n}\, y \right\} = G_k(y). \tag{2.4}$$

We define $G_k^{(n)}(x) = G_k(\frac{x}{\sqrt{n}})$, so that $\mathbb{P}\{L_{k,j}^{(n)} \leq x\} = G_k^{(n)}(x)$ is the cumulative distribution function of initial lead times in the k-th input stream of the n-th queueing system. The process $F_k^*(t)$ of (5.2) is characterized in terms of the function

$$H_k(y) \triangleq \int_y^\infty \big(1 - G_k(\eta)\big) d\eta. \tag{5.1}$$

We introduce the function

$$H_k^{(n)}(x) \triangleq \sqrt{n} H_k \left(\frac{x}{\sqrt{n}} \right) = \int_x^\infty \big(1 - G_k^{(n)}(\xi)\big) d\xi. \tag{7.1}$$

Under EDF, for large values of n, $\widehat{\mathcal{Q}}_k^{(n)}(y,\infty) \approx \lambda_k H_k(y \vee F_k^*)$, and because H_k is non-increasing,

$$\begin{aligned}
\lambda_k H_k(y \vee F_k^*) &= \lambda_k H_k \left(y \vee H_k^{-1}\left(\frac{W_k^*}{\rho_k}\right)\right) \\
&= \lambda_k H_k(y) \wedge \mu_k W_k^* \qquad (7.2)\\
&= \left(\lambda_k \frac{1}{\sqrt{n}} H_k^{(n)}(\sqrt{n}\, y)\right) \wedge Q_k^* \\
&\approx \left(\lambda_k \frac{1}{\sqrt{n}} H_k^{(n)}(\sqrt{n}\, y)\right) \wedge \frac{1}{\sqrt{n}} Q_k^{(n)}.
\end{aligned}$$

Multiplying (7.2) by $\sqrt{n}$, we obtain

$$\mathcal{Q}_k^{(n)}(x,\infty) = \sqrt{n}\, \widehat{\mathcal{Q}}_k^{(n)}\left(\frac{x}{\sqrt{n}}, \infty\right) \approx \left(\lambda_k H_k^{(n)}(x)\right) \wedge Q_k^{(n)}. \qquad (7.3)$$

Relation (7.3) connects the unscaled queue length $Q_k^{(n)}$ with the number of customers whose unscaled lead times exceed x, and the function $H_k^{(n)}$ appearing in this relation can be computed from the cumulative distribution function $G_k^{(n)}$ for the unscaled lead times. Relation (7.3) can be tested by simulation under the earliest-deadline-first discipline without knowledge of the parameter n.

Under FIFO, we have from (6.38) that

$$\begin{aligned}
\widehat{\mathcal{Q}}_k^{(n)}(y,\infty) &\approx \lambda_k \int_y^\infty \left(G_k\left(\eta + \frac{W_k^*}{\rho_k}\right) - G_k(\eta)\right) d\eta \qquad (7.4)\\
&= \frac{\lambda_k}{\sqrt{n}} \int_{\sqrt{n}\, y}^\infty \left(G_k^{(n)}\left(\xi + \frac{\sqrt{n}\, W_k^*}{\rho_k}\right) - G_k^{(n)}(\xi)\right) d\xi.
\end{aligned}$$

Replacing $\sqrt{n}\, W_k^*$ in (7.4) by the approximation $\frac{1}{\mu_k} Q_k^{(n)}$ and multiplying by $\sqrt{n}$, we obtain

$$\begin{aligned}
\mathcal{Q}_k^{(n)}(x,\infty) &= \sqrt{n}\, \widehat{\mathcal{Q}}_k^{(n)}\left(\frac{x}{\sqrt{n}}, \infty\right) \qquad (7.5)\\
&\approx \lambda_k \int_x^\infty \left(G_k^{(n)}\left(\xi + \frac{1}{\lambda_k} Q_k^{(n)}\right) - G_k^{(n)}(\xi)\right) d\xi.
\end{aligned}$$

Once again, we have a relation which connects the unscaled queue length $Q_k^{(n)}$ with the number of customers whose unscaled lead times exceed x, and the the only function appearing in (7.5) is the unscaled lead-time distribution $G_k^{(n)}$. Relation (7.5) can be tested by simulation under the first-in-first-out discipline without knowledge of the parameter n.

In each of Figures 1–8 below, there are eight input streams creating an overall traffic intensity of 0.96. In Figures 1, 3, 5 and 7, this traffic intensity is divided equally among the eight streams, whereas in Figures 2, 4, 6 and 8, the eighth stream is dominant. The traffic intensities $\rho_k^{(n)}$ are indicated in the captions; the service rates in all cases are $\mu_k^{(n)} = 1$, so that $\lambda_k^{(n)} = \rho_k^{(n)}$. In Figures 1–6, each simulation run is initiated with an empty queue and a local time is accumulated when the queue length of the eighth queue is exactly equal to 60. In order to avoid excessive dependence between successive samples, at the instant the local time reaches the value 10, a sample is taken and the local time counter is reset to zero. The simulation continues until 100 samples are recorded. In particular, in the formulas (7.3) and (7.5), we used the value $Q_8^{(n)} = 60$. In Figures 1–4, $G_8(x) = (1 - e^{-x/80})\mathbb{I}_{\{x \geq 0\}}$. In Figures 5 and 6, $G_8(x) = \mathbb{I}_{\{x \geq 80\}}$; in this case, EDF and FIFO coincide, as do the right-hand sides of (7.3) and (7.5).

The left-most dots indicate the pointwise minimum empirical cumulative distribution function of the lead-time profile for these 100 samples, the right-most dots indicate the pointwise maximum, and the central dots are the average. As a function of x, the right-hand side of (7.3) is plotted as a dashed curve in Figures 1, 2, 5 and 6. In Figures 3, 4, 5 and 6, the right-hand side of (7.5) is plotted as a dashed curve.

We obtained the solid curves in Figures 1–6 by replacing $\lambda_8^{(n)}$ in the right-hand sides of (7.3) and (7.5) by $\lambda_8^{(n)}/0.96$. This normalization by the total traffic intensity causes the theory to have better predictive value. Indeed, with this normalization the theoretical cumulative distribution functions and the pointwise average empirical cumulative distribution functions are in almost perfect agreement, except in Figure 5. The cause of ths disagreement in this figure, and indeed the reasonableness of multiplying the approximations (7.2) and (7.4) by $\sqrt{n}$ in order to obtain the relations (7.3) and (7.5) is a subject for future research.

Finally, in Appendix A, it is shown that when there is a dominant traffic intensity, the scaled workloads and queue lengths in the non-dominant queues converge to zero. Figures 7 and 8 graph the empirical marginal tail probabilities of the queue length processes. In Figure 7, one of the identical eight streams with common traffic intensity 0.12 is shown. In Figure 8, there are eight streams of three types, having traffic intensities 0.05, 0.15 or 0.31. The empirical marginal tail probabilities are shown for one stream of each type. The behavior in these two cases is quite different. Figure 8 shows that the non-dominant traffic flows become a negligible part of the workload. Figure 7 shows that in the equal traffic intensity case, the workload will be evenly spread across the different streams, none of which become negligible.

A Appendix

This appendix provides a sufficient condition for Assumption 3.4 to hold, the condition described immediately following that assumption. For this appendix, we adopt the notation and assumptions of Sections 2 and 3, except of course Assumption 3.4 and its consequence Corollary 3.5. In its place we impose Assumption A.2 below. A technical "Crushing Lemma" needed for this appendix is contained in Appendix B.

Recall that, for $k = 1, ..., K$, $R_k^{(n)}(t)$ defined by (2.17) is the work performed by the server on queue k up to time t. Recall also the notation $e(t) = t$ for all $t \geq 0$.

Lemma A.1 *For every k, we have*

$$\frac{1}{n} R_k^{(n)} \circ ne \Rightarrow \rho_k e.$$

PROOF: ¿From Corollary 3.3 we see that $\frac{1}{n}\sum_{k=1}^{K} W_k^{(n)} \circ ne \Rightarrow 0$, and since each term in the sum is nonnegative, we have $\frac{1}{n} W_k^{(n)} \circ ne \Rightarrow 0$ for each k. Equation (2.18) implies

$$\frac{1}{n} R_k^{(n)}(nt) \quad = \quad -\frac{1}{n} W_k^{(n)}(nt) + \frac{1}{n}\left[V_k^{(n)}(A_k^{(n)}(nt)) - \rho_k^{(n)} nt\right] + \rho_k^{(n)} t,$$

and the right-hand side converges to $\rho_k t$ because of Corollary 3.2. □

To proceed further, we make the following assumption.

Assumption A.2 *There exist positive constants w_k, $k = 1, ..., K$, satisfying (3.19) and such that for every $T > 0$ the sequence of random variables*

$$\sqrt{n} \sup_{0 \leq t \leq nT} |w_k^{(n)}(t) - w_k|,$$

$k = 1, ..., K$, $n = 1, 2, ...$, is tight.

This, of course, implies that $w_k^{(n)} \Rightarrow w_k$ as $n \to \infty$ for $k = 1, ..., K$.

Proposition A.3 *Under Assumption A.2, we have $\widehat{W}_k^{(n)} \Rightarrow 0$ for all $k < K$.*

Corollary A.4 *$\widehat{W}_K^{(n)}$ converges weakly to a reflected Brownian motion with drift.*

Corollary A.4 follows immediately from Proposition A.3 and Corollary 3.3. We devote the remainder of the appendix to the proof of the proposition.

PROOF OF PROPOSITION A.3:
For simplicity, we shall prove convergence of the workloads $\widehat{W}_k^{(n)}$, $k < K$, on $[0, 1]$. Equations (2.14) and (2.19) imply

$$\dot{R}_k^{(n)}(t) = \sum_{m=0}^{K-1} \sum_{\alpha \in A_k^c(m)} \frac{w_k^{(n)}(t)}{1 - \sum_{\ell \in \alpha} w_\ell^{(n)}(t)} \dot{J}_\alpha^{(n)}(t). \tag{A.1}$$

Choose $0 < c_0 < \min_{j=1,\ldots,K} w_j$ and let

$$E_n = \left\{ w_k^{(n)}(t) > c_0, \;\; t \in [0, n], \;\; k = 1, ...K \right\}. \tag{A.2}$$

By Assumption A.2,

$$\mathbb{P}(E_n) \to 1. \tag{A.3}$$

On E_n we have, for $k < K$,

$$\begin{aligned}
\frac{\dot{R}_K^{(n)}(t)}{w_K^{(n)}(t)} - \frac{\dot{R}_k^{(n)}(t)}{w_k^{(n)}(t)} &= \sum_{m=1}^{K-1} \sum_{\alpha \in A_K^c(m) \cap A_k(m)} \frac{1}{1 - \sum_{l \in \alpha} w_l^{(n)}(t)} \dot{J}_\alpha^{(n)}(t) \\
&\quad - \sum_{m=1}^{K-1} \sum_{\alpha \in A_K(m) \cap A_k^c(m)} \frac{1}{1 - \sum_{l \in \alpha} w_l^{(n)}(t)} \dot{J}_\alpha^{(n)}(t) \\
&\leq \sum_{m=1}^{K-1} \sum_{\alpha \in A_k(m)} \frac{1}{1 - \sum_{l \in \alpha} w_l^{(n)}(t)} \dot{J}_\alpha^{(n)}(t) \\
&\leq \frac{1}{c_0} \dot{I}_k^{(n)}(t).
\end{aligned}$$

But then

$$\begin{aligned}
\dot{U}_k^{(n)}(t) &= \sum_{m=1}^{K} \sum_{\alpha \in A_k(m)} \frac{w_k^{(n)}(t)}{w_k^{(n)}(t) + 1 - \sum_{l \in \alpha} w_l^{(n)}(t)} \dot{J}_\alpha^{(n)}(t) \\
&\geq w_k^{(n)}(t) \sum_{m=1}^{K} \sum_{\alpha \in A_k(m)} \dot{J}_\alpha^{(n)}(t) \\
&= w_k^{(n)}(t)\, \dot{I}_k^{(n)}(t) \\
&\geq c_0\, w_k^{(n)}(t) \left(\frac{\dot{R}_K^{(n)}(t)}{w_K^{(n)}(t)} - \frac{\dot{R}_k^{(n)}(t)}{w_k^{(n)}(t)} \right),
\end{aligned}$$

and

$$\begin{aligned}
\dot{T}_k^{(n)}(t) &= \dot{R}_k^{(n)}(t) + \dot{U}_k^{(n)}(t) \\
&\geq \dot{R}_k^{(n)}(t) + c_0\, w_k^{(n)}(t) \left(\frac{\dot{R}_K^{(n)}(t)}{w_K^{(n)}(t)} - \frac{\dot{R}_k^{(n)}(t)}{w_k^{(n)}(t)} \right).
\end{aligned}$$

We define

$$\widetilde{T}_k^{(n)}(t) \triangleq \begin{cases} R_k^{(n)}(t) + c_0 \int_0^t w_k^{(n)}(s) \left(\frac{\dot{R}_K^{(n)}(s)}{w_K^{(n)}(s)} - \frac{\dot{R}_k^{(n)}(s)}{w_k^{(n)}(s)} \right) ds & \text{on } E_n, \\ T_k^{(n)}(t) & \text{on } E_n^c. \end{cases} \tag{A.4}$$

Then

$$\dot{T}_k^{(n)}(t) \geq \frac{d}{dt} \widetilde{T}_k^{(n)}(t) \geq 0 \tag{A.5}$$

for all $t \geq 0$ and, after integration, we get $T_k^{(n)} \geq \widetilde{T}_k^{(n)}$. By Assumption A.2, (A.3) and Lemma A.1,

$$\frac{1}{n}\widetilde{T}_k^{(n)} \circ ne \Rightarrow \rho_k e + c_0 w_k \left(\frac{\rho_K}{w_K} - \frac{\rho_k}{w_k}\right) e > \rho_k e.$$

We define

$$\widetilde{N}_k^{(n)}(t) = V_k^{(n)}(A_k^{(n)}(t)) - \widetilde{T}_k^{(n)}(t),$$

so that $N_k^{(n)} \leq \widetilde{N}_k^{(n)}$. We set $\widetilde{U}_k^{(n)}(t) = -\min_{0 \leq s \leq t} \widetilde{N}_k^{(n)}(s)$ and $\widetilde{W}_k^{(n)}(t) = \widetilde{N}_k^{(n)}(t) + \widetilde{U}_k^{(n)}(t)$. We show that

$$W_k^{(n)} \leq \widetilde{W}_k^{(n)}. \tag{A.6}$$

Indeed, let us define

$$\tau = \sup\{t \geq 0; W_k^{(n)}(s) \leq \widetilde{W}_k^{(n)}(s) \text{ for all } s \in [0,t]\}.$$

We wish to prove $\tau = \infty$, so assume on the contrary that $\tau < \infty$. Note that for every t, we have

$$\begin{aligned} N_k^{(n)}(t) - N_k^{(n)}(t-) &= \widetilde{N}_k^{(n)}(t) - \widetilde{N}_k^{(n)}(t-) \\ &= V_k^{(n)}\big(A_k^{(n)}(t)\big) - V_k^{(n)}\big(A_k^{(n)}(t-)\big) \geq 0, \end{aligned}$$

and consequently,

$$W_k^{(n)}(t) - W_k^{(n)}(t-) = \widetilde{W}_k^{(n)}(t) - \widetilde{W}_k^{(n)}(t-) \geq 0.$$

By the definition of τ, we have $W_k^{(n)}(\tau-) \leq \widetilde{W}_k^{(n)}(\tau-)$ and hence $W_k^{(n)}(\tau) \leq \widetilde{W}_k^{(n)}(\tau)$. If $W_k^{(n)}(\tau) > 0$, then for some $\epsilon > 0$, we have $W_k^{(n)}(t) > 0$ for $\tau \leq t \leq \tau + \epsilon$. It follows that $U_k^{(n)}(t) = U_k^{(n)}(\tau)$ for $\tau \leq t \leq \tau + \epsilon$ and so, for this range of t, we have

$$\begin{aligned} W_k^{(n)}(t) &= W_k^{(n)}(t) - W_k^{(n)}(\tau) + W_k^{(n)}(\tau) \\ &\leq N_k^{(n)}(t) - N_k^{(n)}(\tau) + \widetilde{W}_k^{(n)}(\tau) \\ &\leq \widetilde{N}_k^{(n)}(t) - \widetilde{N}_k^{(n)}(\tau) + \widetilde{W}_k^{(n)}(\tau) \\ &\leq \widetilde{W}_k^{(n)}(t) \end{aligned}$$

(the second inequality follows from (A.5)). This contradicts the definition of τ. It follows that we must have $W_k^{(n)}(\tau) = 0$, and in particular, $W_k^{(n)}(\tau) - W_k^{(n)}(\tau-) = 0$, i.e., there is no arrival at time τ. Thus there is an $\epsilon > 0$ with the property that there is no arrival in the interval $[\tau, \tau+\epsilon]$, and in particular, $W_k^{(n)}(t) = 0$ for $\tau \leq t \leq \tau + \epsilon$. Again the definition of τ is contradicted. We conclude that $\tau = \infty$, which is equivalent to (A.6).

Now we want to show that for $k = 1, ...K - 1$

$$\widehat{W}_k^{(n)}(t) \Rightarrow 0, \tag{A.7}$$

i.e., that the workloads corresponding to lower traffic intensities get crushed under the heavy traffic scaling. We prove this by induction on k. Let $k_0 \leq K-1$ be given and suppose that (A.7) is true for all $k = 1, ..., k_0 - 1$ (in particular, no assumption is necessary to consider $k_0 = 1$). The Crushing Lemma B.3 (see Appendix B) applied to

$$\overline{Z}^{(n)} \triangleq \widehat{Z}_{k_0}^{(n)},$$

$$Z^* \triangleq V_{k_0}^* \circ \lambda_{k_0} e + \frac{1}{\mu_{k_0}} A_{k_0}^*,$$

$$\overline{T}^{(n)}(t) \triangleq \frac{1}{n} \widetilde{T}_{k_0}^{(n)}(nt) - \rho_{k_0}^{(n)} t$$

(which is Lipschitz continuous with constant $L := 1$ because $0 \leq \frac{d}{dt} \widetilde{T}_{k_0}^{(n)}(t) \leq 1$ for all n, t) and

$$T^*(t) \triangleq c_0 w_{k_0} \left(\frac{\rho_K}{w_K} - \frac{\rho_{k_0}}{w_{k_0}} \right) t$$

yields $\widetilde{\tau}_{k_0}^{(n)}(t) \Rightarrow t$, where $\widetilde{\tau}_{k_0}^{(n)}(t) \triangleq \sup\{s \in [0,t] : \frac{1}{\sqrt{n}} \widetilde{W}_{k_0}^{(n)}(ns) = 0\}$. Thus, because $0 \leq W_k^{(n)} \leq \widetilde{W}_k^{(n)}$,

$$\tau_{k_0}^{(n)}(t) \Rightarrow t, \tag{A.8}$$

where $\tau_{k_0}^{(n)}(t) \triangleq \sup\{s \in [0,t] : \frac{1}{\sqrt{n}} W_{k_0}^{(n)}(ns) = 0\}$. By the definition of $\tau_{k_0}^{(n)}$, we also have

$$\frac{1}{\sqrt{n}} W_{k_0}^{(n)}(n\tau_{k_0}^{(n)}(t)-) = 0. \tag{A.9}$$

For all $k = 1, ..., K$ we show that

$$\widehat{Z}_k^{(n)}(t) - \widehat{Z}_k^{(n)}(\tau_{k_0}^{(n)}(t)-) \Rightarrow 0. \tag{A.10}$$

Indeed,

$$\begin{aligned} \widehat{Z}_k^{(n)}(t) - \widehat{Z}_k^{(n)}(\tau_{k_0}^{(n)}(t)-) &= \left[\widehat{Z}_k^{(n)}(t) - \widehat{Z}_k^{(n)}(\tau_{k_0}^{(n)}(t)) \right] \\ &+ \left[\widehat{Z}_k^{(n)}(\tau_{k_0}^{(n)}(t)) - \widehat{Z}_k^{(n)}(\tau_{k_0}^{(n)}(t)-) \right]. \end{aligned}$$

The first process on the right-hand side converges weakly to zero by the Differencing Theorem, and the second one converges weakly to zero because, by

Corollary 3.2, $\widehat{Z}_k^{(n)}$ has a continuous limit and thus the maximal jump of $\widehat{Z}_k^{(n)}$ on every finite time horizon $[0,T]$ converges weakly to zero. This proves (A.10).

Thus, for $k \neq k_0$, the process $c_k^{(n)}$ defined by

$$c_k^{(n)}(t) \triangleq \left(\widehat{Z}_k^{(n)}(t) - \widehat{Z}_k^{(n)}(\tau_{k_0}^{(n)}(t)-)\right) - \frac{w_k}{w_{k_0}}\left(\widehat{Z}_{k_0}^{(n)}(t) - \widehat{Z}_{k_0}^{(n)}(\tau_{k_0}^{(n)}(t)-)\right) \tag{A.11}$$

converges weakly to 0. We have

$$\begin{aligned} c_k^{(n)}(t) &= \frac{1}{\sqrt{n}}\left(V_k^{(n)}(A_k^{(n)}(nt)) - V_k^{(n)}(A_k^{(n)}(n\tau_{k_0}^{(n)}(t)-))\right) \\ &\quad - \frac{w_k}{\sqrt{n}w_{k_0}}\left(V_{k_0}^{(n)}(A_{k_0}^{(n)}(nt)) - V_{k_0}^{(n)}(A_{k_0}^{(n)}(n\tau_{k_0}^{(n)}(t)-))\right) \\ &\quad - \sqrt{n}\,(\rho_k^{(n)} - \frac{w_k}{w_{k_0}}\rho_{k_0}^{(n)})(t - \tau_{k_0}^{(n)}(t)). \end{aligned}$$

For every $s \in [n\tau_{k_0}^{(n)}(t), nt]$, queue k_0 is nonempty, and hence every nonzero term $\frac{1}{1-\sum_{l\in\alpha} w_l^{(n)}(t)}\dot{J}_\alpha^{(n)}(s)$ in the sum

$$\frac{\dot{R}_k^{(n)}(s)}{w_k^{(n)}(s)} = \sum_{m=0}^{K-1}\sum_{\alpha\in A_k^c(m)}\frac{1}{1-\sum_{l\in\alpha} w_l^{(n)}(t)}\dot{J}_\alpha^{(n)}(s)$$

also appears in the sum

$$\frac{\dot{R}_{k_0}^{(n)}(s)}{w_{k_0}^{(n)}(s)} = \sum_{m=0}^{K-1}\sum_{\alpha\in A_{k_0}^c(m)}\frac{1}{1-\sum_{l\in\alpha} w_l^{(n)}(t)}\dot{J}_\alpha^{(n)}(s).$$

In other words,

$$\frac{\dot{R}_k^{(n)}(s)}{w_k^{(n)}(s)} \le \frac{\dot{R}_{k_0}^{(n)}(s)}{w_{k_0}^{(n)}(s)} \tag{A.12}$$

for all $k = 1, ...K$ and all $s \in [n\tau_{k_0}^{(n)}(t), nt]$ for which $w_k^{(n)}(s) \neq 0$ and $w_{k_0}^{(n)}(s) \neq 0$.

As a final preparatory step, we let $\epsilon_0 > 0$ be given and choose $\overline{C} > 0$ so large that $\mathbb{P}(G_n) \ge 1 - \epsilon_0$ for all n, where

$$G_n = \left\{\sqrt{n}\sup_{0\le t\le n}\max_{j=1,\dots,K}|w_k^{(n)}(t) - w_k| \le \overline{C}\right\}. \tag{A.13}$$

(Such a choice is possible by Assumption A.2.) Let n_1 be so large that $\min_{j=1,\dots,K} w_j > \frac{\overline{C}}{\sqrt{n_1}}$. In particular, on G_n we have $w_k^{(n)}(t) > 0$ for all $n \ge n_1$, $k = 1, ..., K$ and $0 \le t \le n$. Then there exists a constant C such that on G_n for all $n \ge n_1$, $k = 1, ..., K$ and $0 \le t \le n$ we have

$$\sqrt{n}\left|\frac{w_k^{(n)}(t)}{w_{k_0}^{(n)}(t)} - \frac{w_k}{w_{k_0}}\right| \le C. \tag{A.14}$$

Consider an arbitrary $k > k_0$. By (A.12) and (A.14), we have on G_n

$$
\begin{aligned}
\widehat{W}_k^{(n)}(t) - \widehat{W}_k^{(n)}(\tau_{k_0}^{(n)}(t)-) &= \frac{1}{\sqrt{n}}\left(V_k^{(n)}(A_k^{(n)}(nt)) - V_k^{(n)}(A_k^{(n)}(n\tau_{k_0}^{(n)}(t)-))\right) \\
&\quad - \frac{1}{\sqrt{n}}\left(R_k^{(n)}(nt) - R_k^{(n)}(n\tau_{k_0}^{(n)}(t))\right) \\
&\geq \frac{1}{\sqrt{n}}\left(V_k^{(n)}(A_k^{(n)}(nt)) - V_k^{(n)}(A_k^{(n)}(n\tau_{k_0}^{(n)}(t)-))\right) \\
&\quad - \frac{1}{\sqrt{n}}\int_{n\tau_{k_0}^{(n)}(t)}^{nt} \frac{w_k^{(n)}(s)}{w_{k_0}^{(n)}(s)}\dot{R}_{k_0}^{(n)}(s)\,ds \\
&\geq \frac{1}{\sqrt{n}}\left(V_k^{(n)}(A_k^{(n)}(nt)) - V_k^{(n)}(A_k^{(n)}(n\tau_{k_0}^{(n)}(t)-))\right) \\
&\quad - \frac{w_k}{\sqrt{n}w_{k_0}}\left(R_{k_0}^{(n)}(nt) - R_{k_0}^{(n)}(n\tau_{k_0}^{(n)}(t))\right) \\
&\quad - C(t - \tau_{k_0}^{(n)}(t)) \\
&= \frac{w_k}{\sqrt{n}w_{k_0}}\left(V_{k_0}^{(n)}(A_{k_0}^{(n)}(nt)) - V_{k_0}^{(n)}(A_{k_0}^{(n)}(n\tau_{k_0}^{(n)}(t)-))\right) \\
&\quad + \sqrt{n}\left(\rho_k^{(n)} - \frac{w_k}{w_{k_0}}\rho_{k_0}^{(n)}\right)\,(t - \tau_{k_0}^{(n)}(t)) \;+\; c_k^{(n)}(t) \\
&\quad - \frac{w_k}{\sqrt{n}w_{k_0}}\left(R_{k_0}^{(n)}(nt) - R_{k_0}^{(n)}(n\tau_{k_0}^{(n)}(t))\right) \\
&\quad - C(t - \tau_{k_0}^{(n)}(t)) \qquad\qquad\qquad\qquad\qquad\qquad \text{(A.15)} \\
&= \frac{w_k}{w_{k_0}}\left(\widehat{W}_{k_0}^{(n)}(t) - \widehat{W}_{k_0}^{(n)}(\tau_{k_0}^{(n)}(t)-)\right) \\
&\quad + \sqrt{n}\,\left(\rho_k^{(n)} - \frac{w_k}{w_{k_0}}\rho_{k_0}^{(n)}\right)(t - \tau_{k_0}^{(n)}(t)) \;+\; c_k^{(n)}(t) \\
&\quad - C(t - \tau_{k_0}^{(n)}(t)) \\
&\geq \frac{w_k}{w_{k_0}}\left(\widehat{W}_{k_0}^{(n)}(t) - \widehat{W}_{k_0}^{(n)}(\tau_{k_0}^{(n)}(t)-)\right) \;+\; d_k^{(n)}(t),
\end{aligned}
$$

where

$$
d_k^{(n)}(t) = c_k^{(n)}(t) - \left((1 + \frac{w_k}{w_{k_0}})c + C\right)(t - \tau_{k_0}^{(n)}(t)) \Rightarrow 0
$$

with constants c and C entering the definition of $d_k^{(n)}$ being the same as in (3.4) and (A.14) respectively. The last inequality in (A.15) follows from (3.19) and

(3.4) and uses the fact that $k_0 \le K-1$, to wit,

$$\begin{aligned}
\sqrt{n}\,\left(\rho_k^{(n)} - \frac{w_k}{w_{k_0}}\rho_{k_0}^{(n)}\right) &= \sqrt{n}w_k\left(\frac{\rho_k^{(n)}}{w_k} - \frac{\rho_{k_0}^{(n)}}{w_{k_0}}\right)\\
&= \sqrt{n}w_k\left(\frac{\rho_k^{(n)}-\rho_k}{w_k} - \frac{\rho_{k_0}^{(n)}-\rho_{k_0}}{w_{k_0}}\right)\\
&+ \sqrt{n}w_k\left(\frac{\rho_k}{w_k} - \frac{\rho_{k_0}}{w_{k_0}}\right)\\
&\ge -(1+\frac{w_k}{w_{k_0}})c.
\end{aligned}$$

By Corollary 3.3 and the Differencing Theorem once again,

$$\sum_{k=1}^{K}\left(\widehat{W}_k^{(n)}(t) - \widehat{W}_k^{(n)}(\tau_{k_0}^{(n)}(t)-)\right) \Rightarrow 0.$$

Using our induction assumption we actually have

$$p_k^{(n)}(t) \triangleq \sum_{k=k_0}^{K}\left(\widehat{W}_k^{(n)}(t) - \widehat{W}_k^{(n)}(\tau_{k_0}^{(n)}(t)-)\right) \Rightarrow 0.$$

However, by our previous considerations, on G_n

$$\begin{aligned}
p_k^{(n)}(t) &\ge \left(\sum_{k=k_0}^{K}\frac{w_k}{w_{k_0}}\right)\left(\widehat{W}_{k_0}^{(n)}(t) - \widehat{W}_{k_0}^{(n)}(\tau_{k_0}^{(n)}(t)-)\right) + \sum_{k=k_0+1}^{K} d_k^{(n)}(t)\\
&= \left(\sum_{k=k_0}^{K}\frac{w_k}{w_{k_0}}\right)\widehat{W}_{k_0}^{(n)}(t) + \sum_{k=k_0+1}^{K} d_k^{(n)}(t)
\end{aligned}$$

thanks to (A.9). Thus, on the set G_n with $\mathbb{P}(G_n) \ge 1-\epsilon_0$, we have the following bound on $\widehat{W}_{k_0}^{(n)}$:

$$0 \le \widehat{W}_{k_0}^{(n)}(t) \le \frac{1}{\sum_{k=k_0}^{K}\frac{w_k}{w_{k_0}}}\left(p_k^{(n)}(t) - \sum_{k=k_0+1}^{K} d_k^{(n)}(t)\right) \Rightarrow 0.$$

This proves (A.7) for $k = k_0$, because $\epsilon_0 > 0$ is arbitrary. □

Corollary A.5 *If, instead of the dominance assumption (3.19), we have*

$$0 < \frac{\rho_1}{w_1} \le \frac{\rho_2}{w_2} \le \ldots \le \frac{\rho_{l_0}}{w_{l_0}} < \frac{\rho_{l_0+1}}{w_{l_0+1}} = \ldots = \frac{\rho_K}{w_K},$$

then $\widehat{W}_k^{(n)}(t) \Rightarrow 0$ *for all* $k \le k_0$.

Indeed, the proof given above applies, without any modifications, also to this more general situation.

B Appendix

The purpose of this appendix is to prove the "Crushing Lemma" B.4 used in the proof of Proposition A.3.

B.1 Convergence of inverse processes.

Let $\phi : [0,1) \to [0,\infty)$ be a strictly increasing, continuous function with $\phi(0) = 0$, $\phi(1) \triangleq \lim_{x\uparrow 1} \phi(x) = \infty$. For $x, y \in [0,\infty]$, define

$$\rho(x,y) = |\phi^{-1}(x) - \phi^{-1}(y)|.$$

Then ρ is a metric on $[0,\infty]$ and $([0,\infty],\rho)$ is separable and compact.

Let $D_{[0,\infty]}[0,\infty)$ be the space of RCLL functions on $[0,\infty)$, taking values in $([0,\infty],\rho)$. Let $D_{\mathbb{R}}[0,\infty)$ be the space of RCLL functions on $[0,\infty)$ taking values in $\mathbb{R}$. On both these spaces we impose the Skorohod topology discussed in Chapter 3, Section 5 of [10] and briefly reviewed in Appendix A of [9].

For $f \in D_{\mathbb{R}}[0,\infty)$ and $t \ge 0$, define

$$\Lambda f(t) = \inf\{s \ge 0; f(s) > t\},$$

where $\inf \emptyset = \infty$. It is clear from this definition that $\Lambda f(t)$ is a nondecreasing function of t.

Lemma B.1 *Λ maps $D_{\mathbb{R}}[0,\infty)$ into $D_{[0,\infty]}[0,\infty)$.*

PROOF: Let $f \in D_{\mathbb{R}}[0,\infty)$ be given. Since Λf is nondecreasing, it has left limits. We need to show that Λf is right-continuous.

Let $t \ge 0$ be given and define $s = \Lambda f(t)$. If $s = \infty$, then Λf is right-continuous at t.

Assume $s < \infty$. By the definition of $\Lambda f(t)$, there is a sequence $\{s_n\}$ with $s < s_n$ for every n, $s_n \downarrow s$, and $f(s_n) > t$ for all n. Choose a sequence $\{t_n\}$ with $t_n \downarrow t$ and $t < t_n < f(s_n)$ for every n. Then $\Lambda f(t_n) \le s_n$ and consequently

$$\liminf_{t'\downarrow t} \Lambda f(t') \le \liminf_{n\to\infty} \Lambda f(t_n) \le s = \Lambda f(t).$$

Because Λf is nondecreasing,

$$\liminf_{t'\downarrow t} \Lambda f(t') = \lim_{t'\downarrow t} \Lambda f(t') \ge \Lambda f(t),$$

and we have established the right-continuity of Λf. □

Lemma B.2 *Let $\{f_n\}$ be a sequence of functions in $D_{\mathbb{R}}[0,\infty)$ converging to a continuous, strictly increasing function f. Then Λf is continuous (taking values in $[0,\infty]$) and Λf_n converges to Λf in $D_{[0,\infty]}[0,\infty)$.*

PROOF: We first show that Λf is continuous. Since Λf is nondecreasing and right-continuous, it suffices to show

$$\limsup_{t' \uparrow t} \Lambda f(t') \geq \Lambda f(t)$$

for every $t > 0$. Let $t > 0$ be given and set $s = \Lambda f(t)$.

Case I: $s = \infty$.

Define $M = \sup_{u \geq 0} f(u)$. We have $M \leq t < \infty$. Since f is strictly increasing, f cannot attain the value M. We may choose a sequence $0 < s_1 < s_2 < ...$ such that $s_n \to \infty$, $t_n = f(s_n) < M$, and $t_n \to M$. Then $s_n = \Lambda f(t_n)$ and

$$\limsup_{t' \uparrow t} \Lambda f(t') \geq \lim_{n \to \infty} \Lambda f(t_n) = \lim_{n \to \infty} s_n = \infty = \Lambda f(t).$$

Case II: $0 < s < \infty$.

Since f is strictly increasing, $f(s) = t$. Let $0 < s_1 < s_2 < ...$ be such that $s_n \uparrow s$ and set $t_n = f(s_n)$. Then $t_n \uparrow t$, $\Lambda f(t_n) = s_n$ and

$$\limsup_{t' \uparrow t} \Lambda f(t') \geq \lim_{n \to \infty} \Lambda f(t_n) = \lim_{n \to \infty} s_n = s = \Lambda f(t).$$

Case III: $s = 0$.

Then $\Lambda f \equiv 0$ on $[0, t]$.

We next show that Λf_n converges to Λf. It suffices to show that for every T the restriction of Λf_n to $[0, T]$ converges in the Skorokhod metric to the restriction of Λf to $[0, T]$.

Fix $T > 0$. If $\Lambda f(T) < \infty$, set $S = \Lambda f(T)$, so $T = f(S)$. If $\Lambda f(T) = \infty$, let $\epsilon > 0$ be given and choose $S < \infty$, so that $\rho(S - 2, \infty) < \frac{\epsilon}{2}$.

Next, choose $\delta_1 \in (0, 1)$ so that $|x - y| \leq 2\delta_1$ implies $\rho(x, y) \leq \frac{\epsilon}{2}$ for every $x, y \in [0, S + 2]$.

The continuous positive function $f(s + \delta_1) - f(s)$ has a positive minimum δ_2 over $[0, S]$, i.e., for all $s \in [0, S]$ we have

$$f(s + \delta_1) - f(s) \geq \delta_2.$$

Because $f_n \to f$ in $D_{\mathbb{R}}[0, \infty)$, we may choose N so that for every $n \geq N$ there is a strictly increasing mapping λ_n mapping $[0, S + 2\delta_1]$ onto itself with

$$\sup_{0 \leq s \leq S + 2\delta_1} |f_n(s) - f(\lambda_n(s))| + \sup_{0 \leq s \leq S + 2\delta_1} |s - \lambda_n(s)| < \delta_1 \wedge \delta_2.$$

Let $t \in [0, f(S)]$ be given and set $s = \Lambda f(t)$, so $0 \leq s \leq S$ and $t = f(s)$. We have

$$f_n(s + 2\delta_1) > f(\lambda_n(s + 2\delta_1)) - \delta_2 \geq f(s + \delta_1) - \delta_2 \geq f(s) = t,$$

which shows that

$$\Lambda f_n(t) \leq s + 2\delta_1 = \Lambda f(t) + 2\delta_1.$$

On the other hand, for $0 \le u \le s - 2\delta_1$,

$$f_n(u) < f(\lambda_n(u)) + \delta_2 \le f(u + \delta_1) + \delta_2 \le f(s - \delta_1) + \delta_2 \le f(s) = t,$$

which shows that

$$\Lambda f_n(t) \ge s - 2\delta_1 = \Lambda f(t) - 2\delta_1.$$

We conclude that, for every $t \in [0, f(S)]$,

$$|\Lambda f_n(t) - \Lambda f(t)| \le 2\delta_1.$$

By the choice of δ_1 and the fact that $\Lambda f(t) \in [0, S]$ for $t \in [0, f(S)]$, we have

$$\rho(\Lambda f_n(t), \Lambda f(t)) \le \frac{\epsilon}{2}$$

for all $t \in [0, f(S)]$.

If $\Lambda f(T) < \infty$, so $f(S) = T$, we are done. In the event that $\Lambda f(T) = \infty$, we must consider $\rho(\Lambda f_n(t), \Lambda f(t))$ for $t \in (f(S), T]$. But for $t > f(S)$,

$$\Lambda f(t) \ge \Lambda f(f(S)) = S$$

and

$$\Lambda f_n(t) \ge \Lambda f_n(f(S)) \ge \Lambda f(f(S)) - 2\delta_1 = S - 2\delta_1 > S - 2.$$

Therefore, for $t > f(S)$,

$$\begin{aligned}
\rho(\Lambda f_n(t), \Lambda f(t)) &= |\phi^{-1}(\Lambda f_n(t)) - \phi^{-1}(\Lambda f(t))| \\
&\le |\phi^{-1}(\Lambda f_n(t)) - 1| + |1 - \phi^{-1}(\Lambda f(t))| \\
&\le |1 - \phi^{-1}(S - 2)| + |1 - \phi^{-1}(S)| \\
&= \rho(S - 2, \infty) + \rho(S, \infty) < \epsilon.
\end{aligned}$$

□

Remark B.3 *Assume the hypotheses of Lemma B.2 and let $s \ge 0$ be given. Fix n. If $f_n(s)$ is positive, we may choose a sequence $t_m \uparrow f_n(s)$ with $0 \le t_m < f_n(s)$ for all m. Then $\Lambda f_n(t_m) \le s$. Letting $m \to \infty$, we obtain*

$$\Lambda f_n(f_n(s)-) \le s$$

for all $s \ge 0$ such that $f_n(s) > 0$. Fix $S > 0$. Set $T = f(S)$ and define

$$M_n(T) = \sup_{0 \le t \le T+1} |\Lambda f_n(t) - \Lambda f_n(t-)|.$$

Because $\Lambda f_n \to \Lambda f$, which is finite and continuous, and $M_n(T)$ is an upper semi-continuous function of Λf_n, we have

$$0 \le \limsup_{n\to\infty} M_n(T) \le \sup_{0 \le t \le T+1} |\Lambda f(t) - \Lambda f(t-)| = 0. \qquad \text{(B.1)}$$

Let $0 \leq s \leq S$ be given and suppose that $f_n(s) > 0$. Then assuming, as we can, that n is large enough to guarantee $\sup_{0\leq s\leq S} f_n(s) \leq T+1$, we have

$$s \geq \Lambda f_n(f_n(s)-) \geq \Lambda f_n(f_n(s)) - M_n(T) = \Lambda f_n(\max\{f_n(s), 0\}) - M_n(T).$$

In fact, under an additional assumption that $\Lambda f_n(0) = 0$, for every $0 \leq s \leq S$ we have

$$s \geq \Lambda f_n(\max\{f_n(s), 0\}) - M_n(T) \tag{B.2}$$

also in the event that $f_n(s) \leq 0$.

B.2 The Crushing Lemma

Lemma B.4 *Let $\{\overline{Z}^{(n)}\}_{n=1}^{\infty}$ be a sequence of RCLL processes converging weakly to a continuous limit Z^*. Assume $\overline{Z}^{(n)}(0) = 0$ for every n. Let $\{\overline{T}^{(n)}\}_{n=1}^{\infty}$ be a sequence of RCLL processes satisfying*

$$\overline{T}^{(n)}(t) \geq \overline{T}^{(n)}(t-) \tag{B.3}$$

for all $t \geq 0$, $\overline{T}^{(n)}(0) = 0$ for every n. Assume also that there exists a finite (possibly negative) constant L and a sequence of strictly positive numbers $\{\epsilon_n\}_{n=1}^{\infty}$ such that

$$\overline{T}^{(n)}(t) \geq Lt \tag{B.4}$$

for all $t \in [0, \epsilon_n]$.

Assume $\overline{T}^{(n)} \Rightarrow T^*$, where T^* is nonrandom, continuous and strictly increasing. Define

$$\overline{N}^{(n)}(t) = \overline{Z}^{(n)}(t) - \sqrt{n}\,\overline{T}^{(n)}(t),$$
$$\overline{I}^{(n)}(t) = -\min_{0\leq s\leq t} \overline{N}^{(n)}(s),$$
$$\overline{W}^{(n)} = \overline{N}^{(n)} + \overline{I}^{(n)}$$

and let, for $t \geq 0$,

$$\tau^{(n)}(t) = \sup\{s \in [0,t]; \overline{W}^{(n)}(s) = 0\}. \tag{B.5}$$

Then $\tau^{(n)}(t) \Rightarrow t$.

Proof: First, let us remark that $\overline{W}^{(n)}(0) = 0$, so the supremum in (B.5) is not over the empty set. We shall prove convergence of $\tau^{(n)}(t)$ for $0 \leq t \leq 1$.

We can always assume that the constant L entering (B.4) is actually 1. Indeed, if $L < 1$ (this is the only case we need to analyze), consider, for each n, an auxiliary function $g_n(t)$ which is equal to t on $[0, \frac{1}{n}]$, $\frac{2}{n} - t$ on $[\frac{1}{n}, \frac{2}{n}]$ and is identically zero for $t > \frac{2}{n}$. Then, define

$$\overline{S}^{(n)}(t) = \overline{T}^{(n)}(t) + (1 - L)g_n(t),$$

$$\delta_n = \frac{1}{n} \wedge \epsilon_n.$$

It is clear that $\overline{S}^{(n)}$ satisfies all the assumptions about $\overline{T}^{(n)}$ with L and ϵ_n changed to 1 and δ_n respectively. Moreover, $\overline{S}^{(n)} \geq \overline{T}^{(n)}$ and $\overline{S}^{(n)}(t) = \overline{T}^{(n)}(t)$ for $t \geq \frac{2}{n}$. Thus, if we put

$$\overline{M}^{(n)}(t) = \overline{Z}^{(n)}(t) - \sqrt{n}\,\overline{S}^{(n)}(t),$$

$$\overline{J}^{(n)}(t) = -\min_{0 \leq s \leq t} \overline{M}^{(n)}(s),$$

then $\overline{M}^{(n)} \leq \overline{N}^{(n)}$, which implies $\overline{J}^{(n)} \geq \overline{I}^{(n)}$. This yields, for $t \geq \frac{2}{n}$,

$$0 \leq \overline{W}^{(n)}(t) \leq \overline{M}^{(n)}(t) + \overline{J}^{(n)}(t).$$

For $0 \leq t \leq \frac{2}{n}$, we have

$$0 \leq \tau^{(n)}(t) \leq \tau^{(n)}\left(\frac{2}{n}\right) \leq \frac{2}{n},$$

so

$$|t - \tau^{(n)}(t)| \leq \frac{2}{n}.$$

Thus, the validity of our lemma for $\overline{S}^{(n)}$ clearly implies its validity for $\overline{T}^{(n)}$. Therefore, without loss of generality, we assume $L = 1$.

On $(\tau^{(n)}(t), t]$, the process $\overline{W}^{(n)}$ is strictly positive and so $\overline{I}^{(n)}$ is constant. Therefore,

$$\begin{aligned}
\overline{W}^{(n)}(\tau^{(n)}(t)) &= \overline{Z}^{(n)}(\tau^{(n)}(t)) - \sqrt{n}\,\overline{T}^{(n)}(\tau^{(n)}(t)) + \overline{I}^{(n)}(\tau^{(n)}(t)) \\
&= \overline{Z}^{(n)}(\tau^{(n)}(t)) - \sqrt{n}\,\overline{T}^{(n)}(\tau^{(n)}(t)) + \overline{I}^{(n)}(t) \\
&\geq \overline{Z}^{(n)}(\tau^{(n)}(t)) - \sqrt{n}\,\overline{T}^{(n)}(\tau^{(n)}(t)) - \overline{N}^{(n)}(t) \qquad \text{(B.6)} \\
&= \overline{Z}^{(n)}(\tau^{(n)}(t)) - \overline{Z}^{(n)}(t) \qquad \text{(B.7)} \\
&\quad + \sqrt{n}\left(\overline{T}^{(n)}(t) - \overline{T}^{(n)}(\tau^{(n)}(t))\right).
\end{aligned}$$

We may rewrite this as

$$\overline{T}^{(n)}(t) - \overline{T}^{(n)}(\tau^{(n)}(t)) \leq \frac{1}{\sqrt{n}}\overline{W}(\tau^{(n)}(t)) + \frac{1}{\sqrt{n}}\left[\overline{Z}^{(n)}(t) - \overline{Z}^{(n)}(\tau^{(n)}(t))\right].$$

Now, by (B.3),

$$\frac{1}{\sqrt{n}}\overline{W}(\tau^{(n)}(t)) \leq \frac{1}{\sqrt{n}} \max_{0\leq s\leq 1} \left[\overline{Z}^{(n)}(s) - \overline{Z}^{(n)}(s-)\right],$$

and the right-hand side has limit zero. Furthermore,

$$\frac{1}{\sqrt{n}} \left[\overline{Z}^{(n)}(t) - \overline{Z}^{(n)}(\tau^{(n)}(t))\right] \leq \frac{1}{\sqrt{n}} \left[\overline{Z}^{(n)}(t) - \min_{0\leq s\leq 1} \overline{Z}^{(n)}(s)\right],$$

and again the right-hand side has limit zero. On the other hand, by the Continuous Mapping Theorem,

$$\begin{aligned} \overline{T}^{(n)}(t) - \overline{T}^{(n)}(\tau^{(n)}(t)) &\geq \overline{T}^{(n)}(t) - \sup_{0\leq s\leq t} \overline{T}^{(n)}(s) \\ &\Rightarrow T^*(t) - \sup_{0\leq s\leq t} T^*(s) \\ &= T^*(t) - T^*(t) = 0, \end{aligned}$$

because T^* is increasing. We conclude that

$$\overline{T}^{(n)} - \overline{T}^{(n)} \circ \tau^{(n)} \Rightarrow 0, \tag{B.8}$$

and hence

$$\overline{T}^{(n)} \circ \tau^{(n)} = \overline{T}^{(n)} + \left(\overline{T}^{(n)} \circ \tau^{(n)} - \overline{T}^{(n)}\right) \Rightarrow T^*. \tag{B.9}$$

We want to show that

$$\tau^{(n)}(t) \Rightarrow t. \tag{B.10}$$

Lemma B.2 and the Continuous Mapping Theorem imply

$$\Lambda\overline{T}^{(n)} \Rightarrow \Lambda T^* = (T^*)^{-1}. \tag{B.11}$$

Set

$$M_n = \sup_{0\leq t\leq T^*(1)+1} |\Lambda\overline{T}^{(n)}(t) - \Lambda\overline{T}^{(n)}(t-)|.$$

Define, for a given n,

$$F_n = \left\{ \sup_{0\leq t\leq 1} \overline{T}^{(n)}(t) \leq T^*(1) + 1 \right\}.$$

Our assumptions imply

$$\lim_{n\to\infty} P(F_n) = 1. \tag{B.12}$$

By (B.4) with $L = 1$, we see that $\Lambda\overline{T}^{(n)}(0) = 0$ for every n. Thus, according to Remark B.3, on F_n we have, for $0 \le t \le 1$,

$$t \ge \tau^{(n)}(t) \ge \Lambda\overline{T}^{(n)}(\max\{\overline{T}^{(n)}(\tau^{(n)}(t)), 0\}) - M_n \tag{B.13}$$

(compare (B.2)). By the same remark (see especially (B.1)) and the Continuous Mapping Theorem, we have $M_n \Rightarrow 0$. Thus, if we show

$$\Lambda\overline{T}^{(n)}(\max\{\overline{T}^{(n)}(\tau^{(n)}(t)), 0\}) \Rightarrow t, \tag{B.14}$$

then (B.12) and (B.13) imply (B.10) by the Squeeze Theorem.

Consider arbitrary constants $\eta > 0$, $\epsilon > 0$. Let $c \in (0, 1]$ be chosen in such a way that the value $A \triangleq T^*(1)+c$ is attained by the function T^*, i.e. $A = T^*(S_1)$ for some $S_1 > 1$. Choose $\delta > 0$ such that

$$w_{(T^*)^{-1}}(\delta, [0, A]) < \frac{\epsilon}{4}, \tag{B.15}$$

where the last symbol represents the modulus of continuity of $(T^*)^{-1}$ on $[0, A]$: $w_{(T^*)^{-1}}(\delta, [0, A]) \triangleq \max_{0\le s,t\le A, |s-t|\le\delta} |(T^*)^{-1}(s) - (T^*)^{-1}(t)|$. Let, for any real function f on [0,A],

$$||f||_{[0,A]} \triangleq \sup_{s\in[0,A]} |f(s)|$$

and let

$$A_n = \left\{||\Lambda\overline{T}^{(n)} - (T^*)^{-1}||_{[0,A]} \le \frac{\epsilon}{4}\right\}. \tag{B.16}$$

By (B.11) and the continuity of $(T^*)^{-1}$, we can choose n_0 such that for all $n \ge n_0$

$$P(A_n) \ge 1 - \frac{\eta}{3}. \tag{B.17}$$

Indeed, convergence in the Skorohod metric to a continuous limit implies uniform convergence (see [1], p. 124). Let

$$B_n = \left\{w_{\Lambda\overline{T}^{(n)}}(\delta, [0, A]) \le \frac{3\epsilon}{4}\right\}. \tag{B.18}$$

By (B.15), $A_n \subseteq B_n$ for each n. Now let

$$C_n = \left\{||\overline{T}^{(n)}||_{[0,1]} \le A\right\} \tag{B.19}$$

and suppose that n_1 is such that

$$P(C_n) \ge 1 - \frac{\eta}{3} \tag{B.20}$$

for $n \geq n_1$ (see the comment after (B.17)). Define also

$$D_n = \left\{ || \max\{\overline{T}^{(n)}(\tau^{(n)}), 0\} - T^*||_{[0,1]} \leq \delta \right\}. \tag{B.21}$$

Choose n_2 such that for all $n \geq n_2$ we have

$$P(D_n) \geq 1 - \frac{\eta}{3}. \tag{B.22}$$

Such a choice is possible because, by (B.9) and the Continuous Mapping Theorem,

$$\max\{\overline{T}^{(n)}(\tau^{(n)}(t)), 0\} \Rightarrow \max\{T^*(t), 0\} = T^*(t)$$

and T^* (hence $\max\{T^*(t), 0\}$) is continuous. Take

$$E_n = A_n \cap C_n \cap D_n, \qquad n_3 = n_0 \vee n_1 \vee n_2.$$

By (B.17), (B.20) and (B.22), for $n \geq n_3$ we have $P(E_n) \geq 1 - \eta$. Moreover, on E_n we have, for all $t \in [0, 1]$,

$$\begin{aligned}
|\Lambda\overline{T}^{(n)}(\max\{\overline{T}^{(n)}(\tau^{(n)}(t)), 0\}) - t| &= |\Lambda\overline{T}^{(n)}(\max\{\overline{T}^{(n)}(\tau^{(n)}(t)), 0\}) - (T^*)^{-1}(T^*(t))| \\
&\leq |\Lambda\overline{T}^{(n)}(\max\{\overline{T}^{(n)}(\tau^{(n)}(t)), 0\}) - \Lambda\overline{T}^{(n)}(T^*(t))| \\
&\quad + |\Lambda\overline{T}^{(n)}(T^*(t)) - (T^*)^{-1}(T^*(t))| \\
&\leq \frac{3\epsilon}{4} + \frac{\epsilon}{4} \\
&= \epsilon
\end{aligned}$$

by (B.16), (B.18), (B.19) and (B.21). This proves (B.14), and (B.10) follows. □

References

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

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

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

[4] Chang, C.-S., Stability, queue length, and delay of deterministic and stochastic queueing networks, *IEEE Transactions on Automatic Control* **39**, May 1994, 913–931.

[5] Coffman, E. G., Jr., Puhalskii, A. A. and Reiman, M. I., Polling systems with zero switchover times: a heavy traffic-traffic averaging principle, *Annals of Applied Probability* **5**, 1995, 681–719.

[6] Cruz, R. L., A calculus for network delay, Part I: Network elements in isolation, *IEEE Transactions on Information Theory* **37**, January 1991, 114–131.

[7] Cruz, R. L., A calculus for network delay, Part II: Network analysis, *IEEE Transactions on Information Theory* **37**, January 1991, 132–141.

[8] Demers, A., Keshav, S. and Shenker, S., Design and analysis of a fair queueing system, *Proceedings of ACM SIGCOMM 89*, September 1989.

[9] Doytchinov, B., Lehoczky, J. P. and Shreve, S. E., Real-time queues in heavy traffic with earliest-deadline-first queue discipline, *Annals of Applied Probability*, to appear.

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

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

[12] Kershav, S., *An Engineering Approach to Computer Networking*, Addison-Wesley, Reading, MA, 1997.

[13] Lehoczky, J. P., Scheduling communication networks carrying real-time traffic, *Proceedings of the IEEE Real-Time Systems Symposium*, Madrid, 1998.

[14] Parekh, A. K. and Gallager, R. G., A generalized processor sharing approach to flow control in integrated services networks: The single node case, *IEEE/ACM Transactions on Networking* **1**, June 1993, 344–357.

[15] Parekh, A. K. and Gallager, R. G., A generalized processsor sharing approch to flow control in integrated services networks: The multiple node case, *IEEE/ACM Transactions on Networking* **2**, April 1994, 137–150.

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

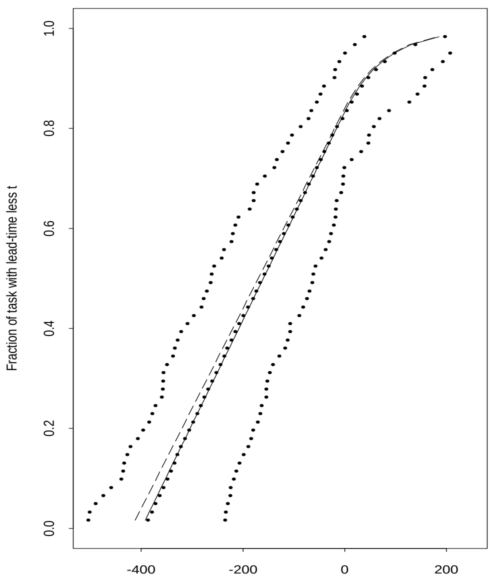
[17] Reiman, M., Some diffusion approximations with state space collapse, *Modelling and Performance Evaluation Methodology*, *Lectures Notes in Control and Information Systems* **60**, 209–240, Springer, New York, 1984.

[18] Yeung, S. N. and Lehoczky, J. P., Real-time queueing networks in heavy traffic with EDF and FIFO queue discipline, working paper, Department of Statistics, Carnegie Mellon University.

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

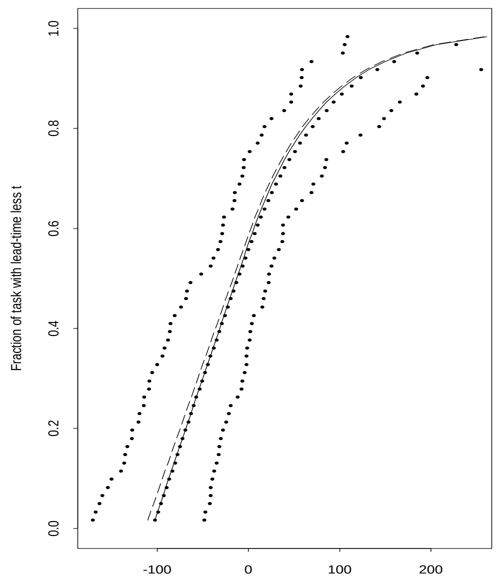
[20] Zhang, H., Service disciplines for guaranteed performance service in packet-switching networks, *Proceedings of the IEEE*, October, 1995.

Figure 1: M/M/1, EDF, Heavy Traffic Queue



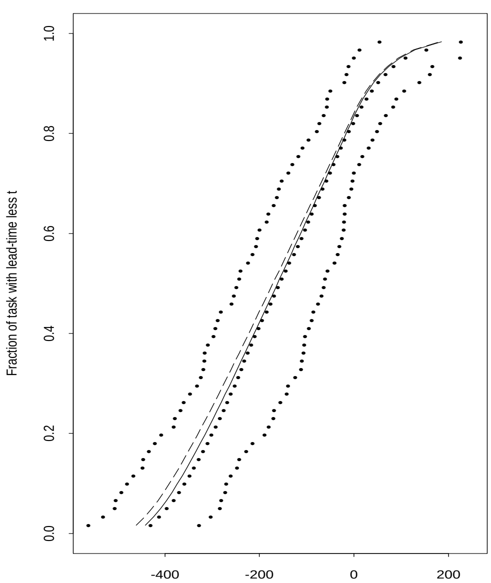
Deadline = 80Exp(1), rho = (.12,.12,.12,.12,.12,.12,.12,.12),
local time = 10, sample size = 100, level = 60.

Figure 2: M/M/1, EDF, Heavy Traffic Queue



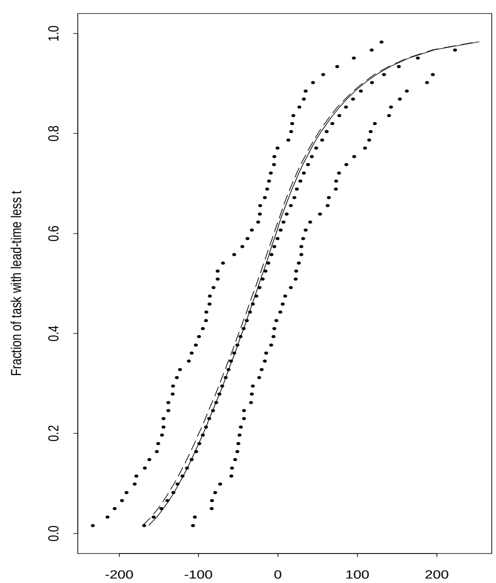
Deadline = 80Exp(1), rho = (.05,.05,.05,.05,.15,.15,.15,.31),
local time = 10, sample size = 100, level =60.

Figure 3: M/M/1, FIFO, Heavy Traffic Queue



Deadline = 80Exp(1), rho = (.12,.12,.12,.12,.12,.12,.12,.12),
local time = 10, sample size = 100, level = 60.

Figure 4: M/M/1, FIFO, Heavy Traffic Queue



Deadline = 80Exp(1), rho = (.05,.05,.05,.05,.15,.15,.15,.31),
local time = 10, sample size = 100, level =60.

Figure 5: M/M/1, Heavy Traffic Queue

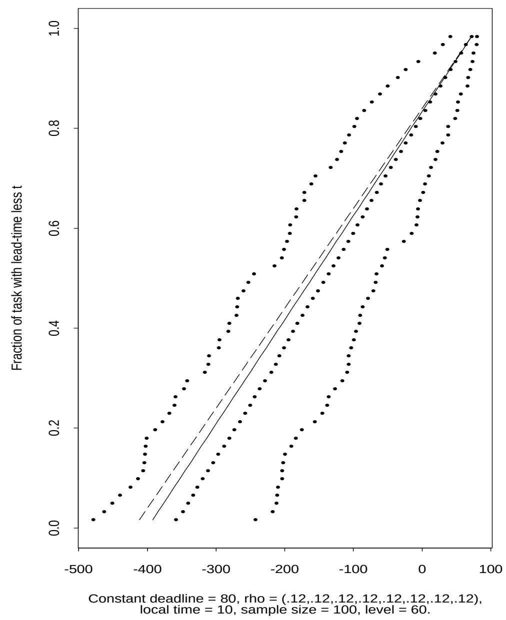


Figure 6: M/M/1, Heavy Traffic Queue

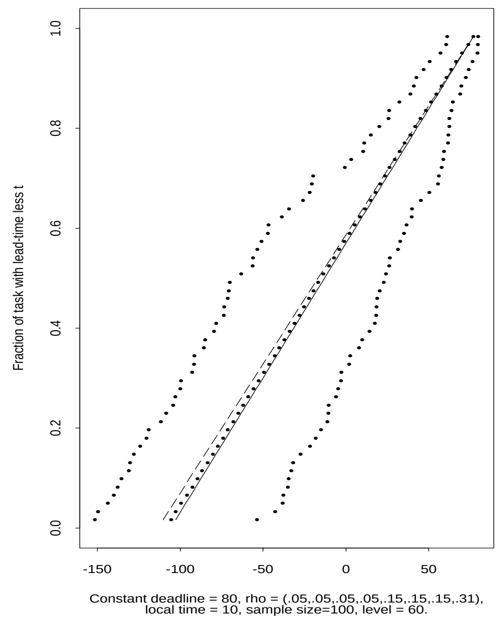


Figure 7: M/M/1, EDF

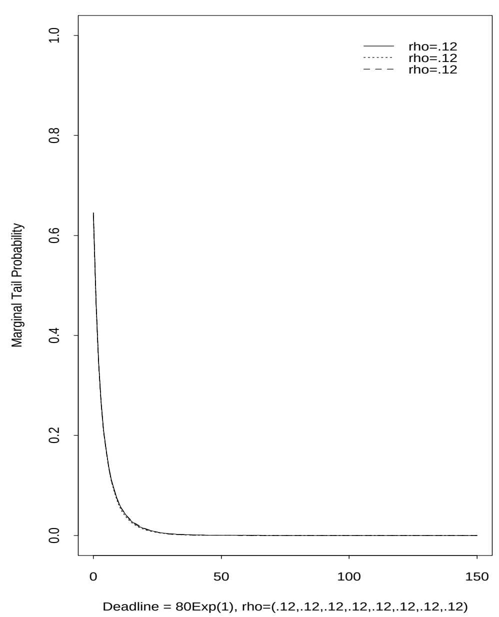


Figure 8: M/M/1, EDF