Second order approximation for the customer time in queue
distribution under the FIFO service discipline

Łukasz Kruk\textsuperscript{a*}, John Lehoczky\textsuperscript{b}, Steven Shreve\textsuperscript{b}

\textsuperscript{a}Department of Computer Science, Marie Curie-Skłodowska University
Pl. M. Curie-Skłodowskiej 1, 20-031 Lublin, Poland
\textsuperscript{b}Carnegie Mellon University, Pittsburgh, USA

Abstract

A single server with one customer class, serviced by the FIFO protocol, is considered and the
instantaneous time in the queue profile of the customers is investigated. We provide the second
order approximation for the random measure describing the customer time in the queue distribu-
tion under heavy traffic conditions.

1. Introduction

The real-time queueing theory is devoted to the study of systems that service
customers with individual timing requirements. Such systems arise naturally in
manufacturing in which orders have due dates, or in real-time computer and
communication networks. To study queueing systems in which the customers
have deadlines, we must attach a lead-time variable to each customer in the
system. It is convenient to model the vector of customer lead times at any time \( t \)
as a counting measure on \( \mathbb{R} \) with a unit atom at the current lead-time of each
customer and total mass equal to to the number of customers in the system at
that time. Doytchinov, Lehoczky and Shreve \cite{1} investigated the single queue
case under the Earliest-Deadline-First (EDF) queue discipline. They proved that
under heavy traffic conditions, a suitably scaled random lead time measure
converges to a non-random function of the limit of the scaled workload process.
Kruk, Lehoczky, Shreve and Yeung \cite{2} gave the corresponding results for the
First-In-First-Out (FIFO) queue discipline and generalized both the EDF and the
FIFO results to the case of a single station with \( K \) input streams, queued in
separate buffers and served by the head-of-the-line processor sharing (HOL-PS)
policy across streams. Yeung and Lehoczky \cite{3} generalized the single server,

\footnote{Corresponding author: \textit{e-mail address}: lkruk@golem.umcs.lublin.pl}
single customer class analysis for EDF and FIFO to multi-class feedforward
of [2] to multi-class acyclic EDF networks. Monte-Carlo simulations show that
the approximations obtained in [1-4] are surprisingly accurate.

It is natural to ask about the rate of convergence in the above-mentioned
results. One way to address this question is to find the second order
approximation for the (scaled) random lead time measure, i.e., to give a
Functional Central Limit Theorem (FCLT) for the (suitably magnified)
difference between the empirical and the theoretical instantaneous lead time
profiles. This corresponds, roughly speaking, to the identification of the second
term in the Taylor expansion for the random lead time measure. This paper
presents the first step in this direction. We consider a single server with one
customer class, serviced by the FIFO protocol and investigate the instantaneous
time in the queue profile of the customers. This can be thought of as a special
case of the lead time profile for the EDF or FIFO service discipline, with all the
customer deadlines (initial lead times) equal to zero. Our main result provides
the second order approximation for the random measure describing the customer
time in queue distribution.

2. The model, assumptions and notation

We have a sequence of single-station queueing systems, each serving one
class of customers. The queueing systems are indexed by the superscript \((n)\).
Each queue is empty at time zero. The inter-arrival times for the customer arrival
process are \(\{u_j^{(n)}\}_{j=1}^\infty\), a sequence of strictly positive, independent, identically
distributed random variables with mean \(1/\lambda^{(n)}\) and standard deviation \(\alpha^{(n)}\). The
service times are \(\{v_j^{(n)}\}_{j=1}^\infty\), another sequence of positive, independent, identically
distributed random variables with mean \(1/\mu^{(n)}\) and standard deviation \(\beta^{(n)}\). We
assume that the sequences \(\{u_j^{(n)}\}_{j=1}^\infty\) and \(\{v_j^{(n)}\}_{j=1}^\infty\) are independent. We define the
customer arrival times

\[
S_0^{(n)} \oplus 0, \quad S_k^{(n)} \oplus \sum_{i=1}^{k} u_i^{(n)}, \quad k \geq 1, \tag{2.1}
\]

the customer arrival process

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

and the work arrival process

\[
V^{(n)}(t) \oplus \sum_{j=1}^{[t]} v_j^{(n)}, \quad t \geq 0. \tag{2.3}
\]
The work which has arrived to the queue by time \( t \) is then \( V^{(n)} \left( A^{(n)}(t) \right) \). We assume that customers are served using the FIFO queue discipline, i.e., the server always services the customer with the longest time in queue. The netput process \( N^{(n)}(t) @ A^{(n)}(t) - t \) measures the amount of work in queue at time \( t \) provided that the server is never idle up to time \( t \). The cumulative idleness process \( I^{(n)}(t) @ \min_{0 \leq s \leq t} N^{(n)}(s) \) gives the amount of time the server is idle, and adding this to the netput process, we obtain the workload process \( W^{(n)}(t) @ N^{(n)}(t) + I^{(n)}(t) \), which records the amount of work in the queue, taking server idleness into account. Let us also define the queue length processes \( Q^{(n)}(t) \), as the number of customers in the queue at time \( t \). All these processes are right-continuous with left-hand limits (RCLL). We assume that the following limits exist and are all positive:

\[
\lim_{n \to \infty} \lambda^{(n)} = \lambda, \quad \lim_{n \to \infty} \mu^{(n)} = \mu, \quad \lim_{n \to \infty} \alpha^{(n)} = \alpha, \quad \lim_{n \to \infty} \beta^{(n)} = \beta. \tag{2.4}
\]

Define the traffic intensity \( \rho^{(n)} @ \lambda^{(n)} / \mu^{(n)} \). We make the heavy traffic assumption

\[
\lim_{n \to \infty} \sqrt{n} \left( 1 - \rho^{(n)} \right) = \gamma \tag{2.5}
\]

for some \( \gamma \in (0) \). We also impose the usual Lindeberg condition on the inter-arrival and service times:

\[
\lim_{n \to \infty} \mathbb{E} \left[ \left( u_j^{(n)} - \left( \lambda^{(n)} \right)^{-1} \right)^2 I \left( \frac{1}{n} \lambda^{(n)} - \frac{1}{\sqrt{n}} \right) \right] = \lim_{n \to \infty} \mathbb{E} \left[ \left( v_j^{(n)} - \left( \mu^{(n)} \right)^{-1} \right)^2 I \left( \frac{1}{n} \mu^{(n)} - \frac{1}{\sqrt{n}} \right) \right] = 0 \quad \forall c > 0. \tag{2.6}
\]

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

\[
\Phi^{(n)}(t) = \frac{1}{\sqrt{n}} I^{(n)}(nt), \quad \Psi^{(n)}(t) = \frac{1}{\sqrt{n}} W^{(n)}(nt), \quad \Theta^{(n)}(t) = \frac{1}{\sqrt{n}} Q^{(n)}(nt),
\]

and the centered heavy traffic scaling for the arrival processes

\[
\Pi^{(n)}(t) = \frac{1}{\sqrt{n}} \left[ A^{(n)}(nt) - \lambda^{(n)} nt \right], \quad \Psi^{(n)}(t) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left( v_j^{(n)} - \frac{1}{\mu^{(n)}} \right).
\]

We define also

\[
\mathcal{M}^{(n)}(t) = \frac{1}{\sqrt{n}} \left[ V^{(n)} \left( A^{(n)}(nt) \right) - nt \right].
\]
Note that $\tilde{\mu}(n)(t) = \bar{\mu}(n)(t) + \tilde{\rho}(n)(t)$. Theorem 3.1 in [5] and Theorem 14.6 in [6] imply that

$$\tilde{\mu}(n) \Rightarrow A^*,$$

where $A^*$ is a Brownian motion with no drift and variance $\alpha^2 \lambda^2$ per unit time. It is also a standard result [7] that

$$\left( \bar{\mu}(n), \bar{\rho}(n), \tilde{\rho}(n) \right) \Rightarrow \left( N^*, I^*, W^* \right),$$

where $N^*$ is a Brownian motion with the variance $(\alpha^2 + \beta^2) \lambda$ per unit time and drift $-\gamma$, $I^*(t) \equiv \min_{0 < s \leq t} N^*(s)$, and $W^*(t) = N^*(t) + I^*(t)$. In other words, $W^*$ is a reflected Brownian motion with drift, and $I^*$ causes the reflection. Here and elsewhere, the symbol $\Rightarrow$ denotes weak convergence of measures on the space $D_{\mathbb{S}} [0, \infty)$ (or $D_{\mathbb{S}} (-\infty, 0]$) of RCLL functions from $[0, \infty)$ ($(-\infty, 0]$) to a Polish space $\mathbb{S}$. The topology of this space is a generalization of the topology introduced by Skorohod for $D_{\mathbb{S}} [0, 1]$. See [6] for details. We usually take $\mathbb{S} = \mathbb{S}^d$ (as in (2.7)) or $\mathbb{S}^d$, with appropriate dimension $d$ (e.g., in (2.8), $d = 3$) for vector-valued functions, unless explicitly stated otherwise.

We shall now define a collection of measure-valued processes which will be useful in the analysis of the instantaneous lead time profile of the customers. We shall follow the convention from [1-4] and consider the instantaneous lead time profile of the customers, where (in the FIFO case) the lead time is the negative of the time spent in queue, i.e.,

lead time = arrival time – current time.

Queue length measure:

$$Q^{(n)}(t)(C) \equiv \left\{ \text{Number of customers in the queue at time } t \text{ having lead times at time } t \text{ in } C \subset \mathbb{S} \right\}.$$

Workload measure:

$$W^{(n)}(t)(C) \equiv \left\{ \text{Work in the queue at time } t \text{ associated with customers in this queue having lead times at time } t \text{ in } C \subset \mathbb{S} \right\}.$$

Customer arrival measure:

$$A^{(n)}(t)(C) \equiv \left\{ \text{Number of all arrivals by time } t \text{ having lead times at time } t \text{ in } C \subset \mathbb{S} \right\}.$$

Workload arrival measure:

$$V^{(n)}(t)(C) \equiv \left\{ \text{Work associated with all arrivals by time } t \text{ having lead times at time } t \text{ in } C \subset \mathbb{S} \right\}.$$

We define the frontier
The negative of the time in queue of the customer currently in service or $S_{(a)}^{(n)}(t) - t$ if the queue is empty.

For the processes just defined, we use the following heavy traffic scalings:

\[
\begin{align*}
\mathcal{Q}^{(n)}(t)(C) & \circledast \frac{1}{\sqrt{n}} \mathcal{Q}^{(n)}(nt)(\sqrt{n}C), \\
\mathcal{W}^{(n)}(t)(C) & \circledast \frac{1}{\sqrt{n}} \mathcal{W}^{(n)}(nt)(\sqrt{n}C), \\
\mathcal{A}^{(n)}(t)(C) & \circledast \frac{1}{\sqrt{n}} \mathcal{A}^{(n)}(nt)(\sqrt{n}C), \\
\mathcal{V}^{(n)}(t)(C) & \circledast \frac{1}{\sqrt{n}} \mathcal{V}^{(n)}(nt)(\sqrt{n}C), \\
\mu^{(n)}(t) & \circledast \frac{1}{\sqrt{n}} F^{(n)}(nt).
\end{align*}
\]

3. First order analysis

We set $H(y) = (-y)^+$. The function $H$ is a 1:1 mapping of $(-\infty, 0]$ onto $[0, \infty)$ and $H^{-1}(y) = -y$ on $[0, \infty)$. We define the limiting scaled frontier process

\[
F^+(t) \circledast H^{-1}(W(t)) = -W^+(t), \quad t \geq 0,
\]

where $W^+$ is as in (2.8). Denote by $\mathcal{M}$ the set of all finite, nonnegative measures on $\mathcal{B}(\mathbb{R})$, the Borel subsets of $\mathbb{R}$. Under the weak topology, $\mathcal{M}$ is a separable, metrizable topological space. The following results are contained in Proposition 3.10 and Theorem 3.1 of [1] (or Theorem 6.1 of [5]).

**Proposition 3.1.** We have $\mu^{(n)} \Rightarrow F^+$ as $n \to \infty$.

**Theorem 3.2.** Let $\hat{W}^+$ and $\hat{Q}^+$ be the measure-valued processes defined by

\[
\hat{W}^+(t)(C) \circledast \int_{C \cap \left( F^+(t), \infty \right]} \left( 1 - 1_{[0, \infty)}(y) \right) dy = m(C \cap \left( F^+(t), 0 \right]),
\]

for all Borel sets $C \subset \circ$, where $m$ denotes Lebesgue measure, and $\hat{Q}^+(t) \circledast \hat{W}^+(t)$. The processes $\mathcal{W}^{(n)}$ and $\mathcal{Q}^{(n)}$ converge weakly in $D([0, \infty), \mathcal{M})$ to $\hat{W}^{(n)}$ and $\hat{Q}^{(n)}$, respectively.

The aim of this paper is to investigate the rate of convergence in Theorem 3.2 by finding the corresponding second order approximations.

4. Second order approximations

In all what follows, we fix $t > 0$. The main result of this paper is the following.
Theorem 4.1 Let $B$ be a Brownian motion with the variance $\lambda (\alpha^2 + \beta^2)$ per unit time and no drift, independent of $F^*(t)$. Then, as $n \to \infty$,
\[
\left( n^\frac{1}{2} \left[ W^{(n)}(t)(y,\infty) - H\left(y \lor H^{-1}(\hat{W}^{(n)}(t))\right) \right] , \ y \leq 0 \right) \Rightarrow \left( B\left(y \lor F^*(t)\right) - B\left(F^*(t)\right)1_{(-\infty,F^*(t))}(y), \ y \leq 0 \right).
\]
(4.1)

Moreover, if $\lim_{n \to \infty} n^\frac{1}{2} (\lambda^{(n)} - \lambda) = \nu$ for some $\nu \in \mathbb{R}$, then
\[
\left( n^\frac{1}{2} \left[ Q^{(n)}(t)(y,\infty) - \lambda H\left(y \lor H^{-1}(\hat{W}^{(n)}(t))\right) \right] , \ y \leq 0 \right) \Rightarrow \left( \hat{B}\left(y \lor F^*(t)\right) - \lambda B\left(F^*(t)\right)1_{(-\infty,F^*(t))}(y), \ y \leq 0 \right).
\]
(4.2)

where $\hat{B}$ is another Brownian motion, independent of $F^*(t)$, with the variance $\alpha^2 \lambda^3$ per unit time and drift $\nu$, and such that the correlation between $B$ and $\hat{B}$ equals $\frac{\alpha}{\sqrt{\alpha^2 + \beta^2}}$.

The convergence in both (4.1) and (4.2) takes place in $D_{(-\infty,0]}$.

Remark 4.2 By (2.8), (3.1) and Theorem 3.2, for each $y \leq 0$,
\[
\hat{W}^{(n)}(t)(y,\infty) \Rightarrow \hat{W}^*(t)(y,\infty) = H\left(y \lor F^*(t)\right) = H\left(y \lor H^{-1}(W^*(t))\right)
\]
\[
\approx H\left(y \lor H^{-1}\left(\hat{W}^{(n)}(t)\right)\right),
\]
and, similarly, $Q^{(n)}(t)(y,\infty) = \lambda H\left(y \lor H^{-1}(\hat{W}^{(n)}(t))\right)$, i.e., the empirical measure $\hat{W}^{(n)}(t)$ ($Q^{(n)}(t)$) of any half-line $(y,\infty)$ can be estimated by plugging the rescaled workload $\hat{W}^{(n)}(t)$, instead of the limiting workload $\hat{W}^*(t)$, into the theoretical profile $H\left(y \lor H^{-1}(W^*(t))\right)$ ($\lambda H\left(y \lor H^{-1}(W^*(t))\right)$). Theorem 4.1 characterizes the accuracy of these approximations.

We shall prove only (4.1) (the proof of (4.2) is similar). To this end, we need two auxiliary results. For all $y \leq 0$, we have
\[
\Psi^{(n)}(t)(y,\infty) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \gamma_j^{(n)} 1_{[\sqrt{n}y \leq S_j^{(n)} \leq \sqrt{n}y]}.
\]

Let us define the process
Proposition 4.3 We have, as \( n \to \infty \),
\[
\begin{bmatrix}
\mathbf{y}^{(n)}(t)(y, \infty) \\
\mathbf{y}^{(n)}(t)(y, \infty)
\end{bmatrix} \overset{\mathcal{D}}{\to} \sqrt{n} \left[ \mathbf{y}^{(n)}(t)(y, \infty) + H\left(y + \sqrt{nt}\right) - H(y) \right], \quad y \le 0. \quad (4.3)
\]

where \( B \) is a Brownian motion with no drift and variance \( \lambda(\alpha^2 + \beta^2) \) per unit time, independent of \( F^*(t) \) and \( W^*(t) \).

**PROOF:** By ordinary and renewal FCLTs for triangular arrays (see, e.g., [5, 6, 8], we have, with \( m = \sqrt{n} \),
\[
\begin{bmatrix}
\mathbf{y}^{(n)}(t)(y, \infty) \\
\mathbf{y}^{(n)}(t)(y, \infty)
\end{bmatrix} = \frac{1}{\sqrt{n}} \left\{ \sum_{k=1}^{A^{(n)}(nt)} \mathbf{u}^{(n)}_{k} + \sqrt{n}H\left(y + \sqrt{nt}\right) - \sqrt{n}H(y) \right\}
\]
\[
= \frac{1}{\sqrt{n}} \left\{ \sum_{k=1}^{A^{(n)}(m^2)} \mathbf{u}^{(m^2)}_{k} - m\left[\mathbf{y}\{m^2 + my \ge 0\} - m\left[\mathbf{y}\{m^2 + my < 0\}\right] \right] \right\}
\]
\[
\Rightarrow B\left(|y|\right), \quad (4.5)
\]
where \( B \) is a Brownian motion with no drift and variance \( \lambda(\alpha^2 + \beta^2) \) per unit time.

Define, for \( s \ge 0 \),
\[
K^{(n)}(s) \@ V^{(n)}(nt)(-\infty, -nt + s] = V^{(n)}(nt)[-nt, \infty) - V^{(n)}(nt)(-nt + s, \infty)
\]
\[
= \sqrt{n} \mathbf{y}^{(n)}(t)[-\sqrt{nt}, \infty) - \sqrt{n} \mathbf{y}^{(n)}(t)(-\sqrt{nt} + s, \infty) \quad (4.6)
\]
In other words, for \( 0 \le s \le nt \), \( K^{(n)}(s) \) is the work arrived to the server in the time interval \([0, s]\). Let
\[
L^{(n)}(s) \@ \min_{0 \le u \le s} L^{(n)}(u), \quad M^{(n)}(s) \@ \min_{0 \le u \le nt} L^{(n)}(u). \quad (4.7)
\]
By (4.6) and (4.7), we have
\[
M^{(n)}(nt) = \min_{0 \le u \le nt} \sqrt{n} \left[ \mathbf{y}^{(n)}(t)[-\sqrt{nt}, \infty) - \mathbf{y}^{(n)}(t)(-\sqrt{nt} + \frac{u}{\sqrt{n}, \infty}) \right]
\]
\[
= \min_{-\sqrt{nt} \le u \le 0} \sqrt{n} \left[ \mathbf{y}^{(n)}(t)[-\sqrt{nt}, \infty) - \mathbf{y}^{(n)}(t)(\eta, \infty) - \eta - \sqrt{nt} \right]. \quad (4.8)
\]
It is easy to see (compare the definitions of the netput, cumulative idleness and workload processes), that for $0 \le s \le nt$ we have $W^{(n)}(s) = L^{(n)}(s) - M^{(n)}(s)$. In particular, by (4.6)-(4.8) and the fact that $\Psi^{(n)}(0,\infty) = 0$, we have

$$
\Psi^{(n)}(t) = \frac{1}{\sqrt{n}}(L^{(n)}(nt) - M^{(n)}(nt)) = \Psi^{(n)}(t)[-\sqrt{nt}, \infty) - \sqrt{nt} - \min_{-\sqrt{nt} \le \eta \le 0} \left[\Psi^{(n)}(t)[-\sqrt{nt}, \infty) - \Psi^{(n)}(\eta, \infty) - \eta - \sqrt{nt}\right]
$$

(4.9)

Using this fact, together with (4.9), after some algebraic manipulations, we get, for every fixed $y_0 < 0$,

$$
\Psi^{(n)}(t) = \max_{-\sqrt{nt} \le \eta \le 0} \left\{ \frac{1}{\sqrt{n}} \sum_{j=A^{(n)}(nt+\sqrt{nt})}^{A^{(n)}(nt+\sqrt{nt})} \nu_j^{(n)} + \rho^{(n)} + \eta o\left(\frac{1}{\sqrt{n}}\right) \right\}
$$

$$= \max_{0 \le s \le 1} \left\{ \frac{1}{\sqrt{n}} \sum_{j=A^{(n)}(nt+\sqrt{nt})}^{A^{(n)}(nt+\sqrt{nt})} \nu_j^{(n)} - \rho^{(n)}(ntu) - \gamma tu \right\} + o(1)
$$

(4.10)

$$= \max_{0 \le s \le 1} \left\{ \frac{1}{\sqrt{n}} \sum_{j=A^{(n)}(nt+\sqrt{nt})}^{A^{(n)}(nt+\sqrt{nt})} \nu_j^{(n)} - \rho^{(n)}(\sqrt{ny_0 + ny}) \right\} + o(1)
$$

$$+ \frac{1}{\sqrt{n}} \sum_{j=A^{(n)}(nt+\sqrt{nt})}^{A^{(n)}(nt+\sqrt{nt})} \nu_j^{(n)} - \rho^{(n)}(\sqrt{ny_0 + ny}) \right\} + o(1)
$$

where $X^{(n)}(u)$ ($Y^{(n)}(u)$) is the first (second) term inside the curled brackets in the RHS of the third equation in (4.10). Let us notice that, by the definition of $X^{(n)}$ and the first equality in (4.5), the process $X^{(n)}$ is independent on $(\Psi^{(n)}(t)(\rho, \infty), \; y_0 \le \rho \le 0)$. Moreover, by FCLT, $X^{(n)}(u) \Rightarrow X(u)$, where $X$ is a driftless Brownian motion, and $Y^{(n)} \Rightarrow 0$ by the Differencing Theorem (see, e.g., Theorem A.3 of [1]). Therefore, by (4.10), for every $y_0$, $\Psi^{(n)}(t)$ is
asymptotically independent on \((\mathcal{Y}^{(n)}_y(t), y_0 \leq y \leq 0)\). This, together with (2.8) and (4.5), yields
\[
\left( \mathcal{Y}^{(n)}_y(t), y_0 \leq y \leq 0, \mathcal{W}^{(n)}_y(t) \right) \Rightarrow \left( B(|y|), y_0 \leq y \leq 0, W^*(t) \right), \tag{4.11}
\]
where \(B\) and \(W^*(t)\) are independent. By the definition of the topology on \(D(-\infty, 0]\), (4.11) can be upgraded to \(y \leq 0\). By (3.1) and Proposition 3.1,
\[
\mathcal{Y}^{(n)}_y(t) = F^*(t) + o(1) = H^{-1}(W^*) + o(1) = H^{-1}\left( \mathcal{W}^{(n)}_y \right) + o(1), \tag{4.12}
\]
and, by (3.1) and (the upgraded) (4.11),
\[
\left( \mathcal{Y}^{(n)}_y(t), y \leq 0 \right), \mathcal{W}^{(n)}_y(t), H^{-1}\left( \mathcal{W}^{(n)}_y \right) \Rightarrow \left( B(|y|), y \leq 0, W^*(t), F^*(t) \right) \tag{4.13}
\]
where \(B\) and \((W^*(t), F^*(t))\) are independent. Finally, (4.4) follows immediately from (4.12) and (4.13).

\[\Box\]

**Lemma 4.4** For every \(y_0 < 0\), as \(n \to \infty\),
\[
n^{\frac{1}{4}} \sup_{y \leq y_0} \mathcal{Y}^{(n)}_y(t) \{y\} \to 0. \tag{4.14}
\]

**PROOF:** Fix \(y_0 < 0\). For every \(y_0 \leq y \leq 0\), we have
\[
n^{\frac{1}{4}} \left[ \mathcal{Y}^{(n)}_y(t)(y, \infty) + H\left(y + \sqrt{nt}\right) - H(y) \right]
\]
\[
\leq n^{\frac{1}{4}} \left[ \mathcal{Y}^{(n)}_y(t)(y, \infty) + H\left(y + \sqrt{nt}\right) - H(y) \right]
\]
\[
\leq n^{\frac{1}{4}} \left[ \mathcal{Y}^{(n)}_y(t)(y - \frac{1}{n}, \infty) + H\left(y - \frac{1}{n} + \sqrt{nt}\right) - H\left(y - \frac{1}{n}\right) + \frac{1}{n} \right].
\]

So
\[
n^{\frac{1}{4}} \mathcal{Y}^{(n)}_y(t) \{y\} \leq \sup_{y_0 \leq y \leq 0} \left\{ n^{\frac{1}{4}} \left[ \mathcal{Y}^{(n)}_y(t)(y - \frac{1}{n}, \infty) + H\left(y - \frac{1}{n} + \sqrt{nt}\right) - H\left(y - \frac{1}{n}\right) \right] \right. \\
- n^{\frac{1}{4}} \left[ \mathcal{Y}^{(n)}_y(t)(y, \infty) + H\left(y + \sqrt{nt}\right) - H(y) \right] + n^{-\frac{3}{4}} \Rightarrow 0
\]
by (4.3), (4.5), the Differencing Theorem and the Continuous Mapping Theorem. \[\Box\]
PROOF OF (4.1): For any $0 \leq y$, the LHS of (4.1) equals

\[
\begin{align*}
&\frac{1}{n^3} \left[ \hat{\mathcal{H}}^{(n)}(t)(y, \infty) - H \left( y \vee H^{-1}\left( \mu^{(n)}(t) \right) \right) \right] \\
&= \frac{1}{n^3} \left[ \mathcal{H}^{(n)}(t)(y, \infty) - \mu^{(n)}(t) \left( y \vee \mu^{(n)}(t), \infty \right) \right] \\
&\quad + n^2 \left[ \mathcal{H}^{(n)}(t) \left( y \vee \mu^{(n)}(t), \infty \right) - H \left( y \vee \mu^{(n)}(t) \right) \right] \\
&\quad + \frac{1}{n^3} \left[ H \left( y \vee \mu^{(n)}(t) \right) - H \left( y \vee H^{-1}\left( \mu^{(n)}(t) \right) \right) \right].
\end{align*}
\]

(4.15)

We shall analyze each term on the RHS of (4.15) separately.

By the definition of the frontier and the FIFO discipline, none of the customers in queue at time $t$ with lead time greater than $F^{(n)}(t)$ has ever been in service by time $t$. Thus, if $y \geq \mu^{(n)}(t)$, then the first term on the RHS of (4.15) is zero and, moreover, if $y < \mu^{(n)}(t)$, then

\[
0 \leq \frac{1}{n^3} \left[ \hat{\mathcal{H}}^{(n)}(t)(y, \infty) - \mu^{(n)}(t) \left( y \vee \mu^{(n)}(t), \infty \right) \right] \\
= \frac{1}{n^3} \left[ \mathcal{H}^{(n)}(t)(\mu^{(n)}(t), \infty) - \mathcal{H}^{(n)}(t) \left( \mu^{(n)}(t), \infty \right) \right] \\
\leq \frac{1}{n^3} \mathcal{H}^{(n)}(t) \left\{ \mu^{(n)}(t) \right\}.
\]

(4.16)

By Proposition 3.1 and Lemma 4.4, we have

\[
\mathcal{H}^{(n)}(t) \left\{ \mu^{(n)}(t) \right\} \Rightarrow 0,
\]

(4.17)

and thus, by (4.16) and (4.17),

\[
\frac{1}{n^3} \left[ \hat{\mathcal{H}}^{(n)}(t)(y, \infty) - \mathcal{H}^{(n)}(t) \left( y \vee \mu^{(n)}(t), \infty \right) \right] \Rightarrow 0.
\]

(4.18)

Let us notice that, by definition, $\mu^{(n)}(t) + \sqrt{n} t \geq 0$, so

\[
H \left( y \vee \mu^{(n)}(t) \right) + \sqrt{n} t = 0.
\]

Thus,
\[ n^{\frac{1}{2}} \left[ \mathcal{V}^{(n)}(t) \mathcal{V}^{H^{(n)}}(t) - H \left( y \mathcal{V}^{H^{(n)}}(t) \right) \right] = n^{\frac{1}{2}} \left[ \mathcal{V}^{(n)}(t) \mathcal{V}^{H^{(n)}}(t) + H \left( y \mathcal{V}^{H^{(n)}}(t) + \sqrt{nt} \right) - H \left( y \mathcal{V}^{H^{(n)}}(t) \right) \right] \] (4.19)

\[ \Rightarrow B \left( y \mathcal{V}^{H^{*}}(t) \right) \]

by Proposition 4.3. Finally, by a similar argument,

\[ n^{\frac{1}{2}} \left( \mathcal{V}^{(n)}(t) + \mathcal{V}^{H^{(n)}}(t) \right) = \mathcal{W}^{(n)} \left\{ \mathcal{V}^{H^{(n)}}(t) \right\} + n^{\frac{1}{2}} \left[ \mathcal{V}^{(n)}(t) \mathcal{V}^{H^{(n)}}(t) + H \left( \mathcal{V}^{H^{(n)}}(t) + \sqrt{nt} \right) - H \left( \mathcal{V}^{H^{(n)}}(t) \right) \right] \] (4.20)

\[ \Rightarrow B \left( \mathcal{V}^{H^{*}}(t) \right) \]

by Proposition 4.3, (4.17) and the inequality \( 0 \leq \mathcal{W}^{(n)} \left\{ \mathcal{V}^{H^{(n)}}(t) \right\} \leq \mathcal{V}^{(n)} \left\{ \mathcal{V}^{H^{(n)}}(t) \right\} \).

Thus, by (2.8), (3.1) and Proposition 3.1, we have

\[ n^{\frac{1}{2}} \left[ H \left( y \mathcal{V}^{H^{(n)}}(t) \right) - H \left( y \mathcal{V}^{H^{*}(n)}(t) \right) \right] \]

\[ = n^{\frac{1}{2}} \left[ y \mathcal{V}^{H^{(n)}}(t) - y \mathcal{V}^{H^{(n)}}(t) \right] \] (4.21)

\[ \Rightarrow - B \left( \mathcal{V}^{H^{*}(t)} \right) I_{\left[ -\infty, F^{*}(t) \right]}(y). \]

It is easy to see that the convergence in (4.18), (4.19) and (4.21) is, in fact, joint, so, by (4.15), (4.1) follows.

**Corollary 4.5** As \( n \to \infty \),

\[ n^{\frac{1}{2}} \left[ \mathcal{W}^{(n)}(t)(y, \infty) - H \left( y \mathcal{V}^{H^{(n)}}(t) \right) \right], \quad y \leq 0 \] \( \Rightarrow \left( B \left( y \mathcal{V}^{H^{*}(t)} \right), \quad y \leq 0 \right) \). (4.22)

Moreover, if \( \lim_{n \to \infty} n^{\frac{1}{2}} \left( \lambda^{(n)} - \lambda \right) = \nu \) for some \( \nu \in \mathcal{O} \), then

\[ n^{\frac{1}{2}} \left[ \mathcal{Q}^{(n)}(t)(y, \infty) - \lambda H \left( y \mathcal{V}^{H^{(n)}}(t) \right) \right], \quad y \leq 0 \] \( \Rightarrow \left( B \left( y \mathcal{V}^{H^{*}(t)} \right), \quad y \leq 0 \right) \). (4.23)

The convergence in (4.22)-(4.23) takes place in \( D_{\left[ -\infty, 0 \right]} \).

(4.22) follows immediately from the fact that \( n^{\frac{1}{2}} \left[ \mathcal{W}^{(n)}(t)(y, \infty) - H \left( y \mathcal{V}^{H^{(n)}}(t) \right) \right] \) is the sum of the first two terms on the RHS of (4.15), together with (4.18) and (4.19). The proof of (4.23) is similar.
Acknowledgements

This work was Supported by the State Committee for Scientific Research of Poland Grant No. 2 PO3A 012 23, ONR N00014-92-J-1524, DARPA F33615-00-C-1729 and by NSF Grant No. DMS-01-03814.

References