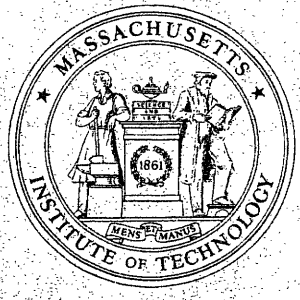
OPERATIONS RESEARCH CENTER

working paper



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

A Multiclass Hybrid Production Center in Heavy Traffic

Vien Nguyen

OR 305-95 May 1995

A Multiclass Hybrid Production Center in Heavy Traffic

Viên Nguyen
Sloan School of Management, M.I.T., Cambridge, MA 02139

Abstract

This paper presents an analysis of a single-stage hybrid production system that makes multiple types of products, some of which are made to-order while others are made to-stock. The analysis begins with a formal heavy traffic limit theorem of the production system, which is modeled as a mixed queueing network. Taking insights from the limit theorem, the analysis continues with the development of an approximation procedure. Numerical experiments indicate that this procedure provides good estimates for performance measures and bounds such as fill rates and average inventory levels.

KEYWORDS: multiclass queueing networks, mixed queueing networks, make-to-order production, make-to-stock production, diffusion approximation, reflected Brownian motion, performance analysis.

Contents:

August, 1994

This paper presents an analysis of a "hybrid" production system that makes multiple types of products, some of which are produced to inventory (make-to-stock), while others are produced in response to actual customer demands (make-to-order). We develop a procedure for performance analysis of the production system depicted in Figure 1. In particular, we envision the production process as a single aggregate operation with first-in-first-out (FIFO) service discipline. Production of make-to-stock items follows a policy of one-for-one replenishment. That is, a base stock level is specified for each type of products; demand is filled from finished-goods-inventory; and each item pulled from inventory triggers a replenishment order to restore the finished-goods-inventory to the desired base stock level. Demands that cannot be met due to insufficient inventory will be considered lost.

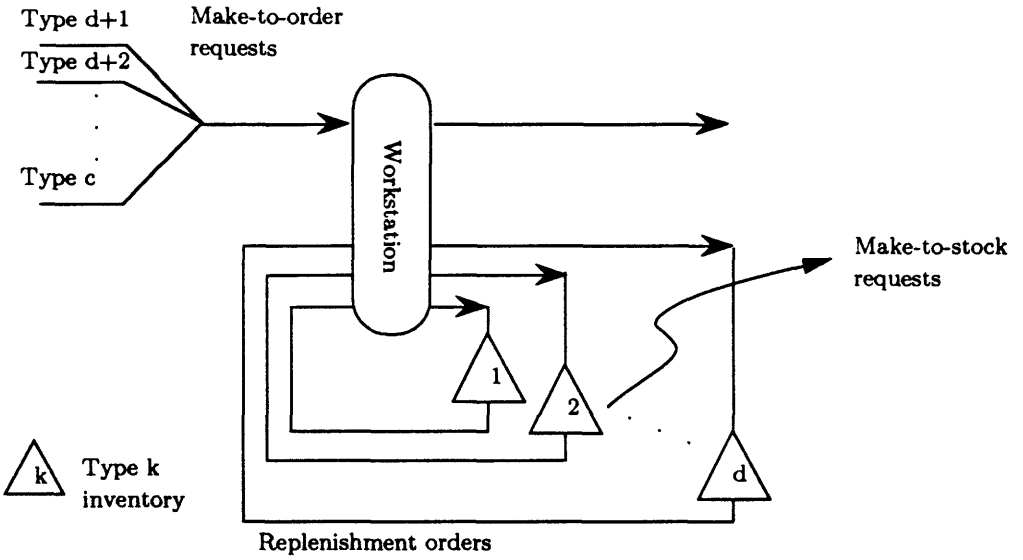


Figure 1: A workstation with mixed jobs of multiple types

We propose to study the production system depicted in Figure 1 via the "mixed" queueing network shown in Figure 2. Station 0 represents the workstation (hereafter interchangeably referred to as "workcenter"): an arrival at station 0 signals a production request and each service completion at station 0 corresponds to the production of an item. Stations 1 to d model the finished-goods-inventories (FGI) for make-to-stock products: items in queue k represent the FGI of type k ($1 \leq k \leq d$) and service durations at station k correspond to intervals between demands (i.e., inter-demand times) of product k. Each filled demand triggers a corresponding replenishment order, so jobs that "depart" from station k are routed to station 0. Because demands that cannot be filled from inventory are simply lost, the number of items in FGI

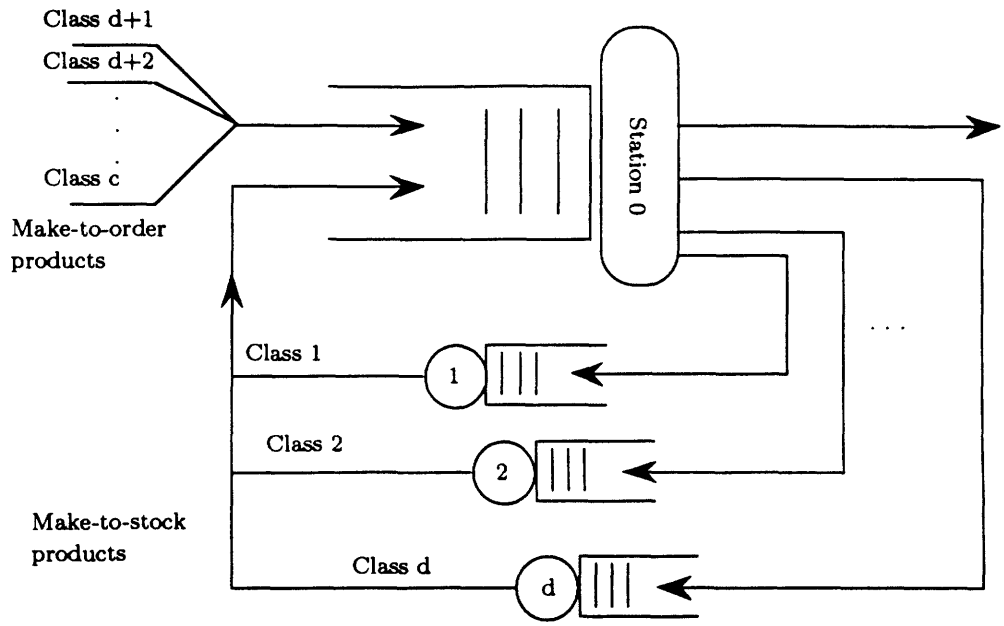


Figure 2: A multiclass, mixed queueing network

remains nonnegative; moreover, the number of items in FGI summed with the number of replenishment orders at the workcenter for each product type is *constant* at all times and equals the pre-specified base stock level. In the language of queueing networks, make-to-order products are "open" jobs whereas make-to-stock products are "closed' jobs.

In some settings, the nature of the product line may require that some items be made to customer orders (for example, items with a high level of customization), and others be made to inventory (for example, commodity items). Even in a nominally make-to-stock scenario, however, a recent study by Carr, Güllü, Jackson and Muckstadt (1993) suggests that a hybrid mode of operation can provide more efficient service. Although there is an abundance of results regarding analysis and optimization of production systems that are either make-to-order or make-to-stock, there are few results for systems that employ *both* types of production (see Basket, Chandy, Muntz and Palacios 1975). A previous paper by Nguyen (1994) developed an approximation procedure, based on heavy traffic theory, for a single-stage hybrid production system in which each make-to-order and make-to-stock category contains exactly one product type (that is, $d = 1$ and $c = 2$ in Figures 1 and 2). In the present paper, we extend the analysis to the case in which each make-to-order/make-to-stock category may contain several types of products.

We will label make-to-stock products as types 1 to d. This convention conveniently corresponds to the numbering of stations representing finished-goods-inventories, so that make-to-

stock products of type k alternately visit stations 0 and k. The base-stock level for products of type k is n_k. Make-to-order products will be designated by types $d+1$ to c. Let λ_k be the demand rate and m_{0k} be the mean production time of type k products; the squared coefficient of variation (SCV) of type k demands is c_k^2 and the SCV of type k processing times is c_{0k}^2.

Figure 2 suggests that we model the demand processes of make-to-stock products by the service processes at stations 1 to d. Strictly speaking, a service process characterized by independent and identically distributed service times may not be a faithful representation of the demand process because the first inter-demand time following a period of inventory depletion (during which demands are lost) typically is not statistically similar to other inter-demand intervals. Nonetheless, Nguyen (1994) noted that the difference is not significant in the sense that the two systems closely approximate each other under heavy traffic conditions (see also Iglehart and Whitt 1970). We thus proceed with our setup, in which case it will be convenient to define $m_k \equiv \lambda_k^{-1}$, with the interpretation that m_k is the mean "service" time at station k for $1 \leq k \leq d$.

We will assume throughout this paper that for all product types, $\lambda_k < m_{0k}^{-1}$ (i.e., $m_{0k} < m_k$ for $1 \leq k \leq d$), which requires production rates to exceed those of customer demands. Define the "relative" traffic intensities at station j to be

$$\rho_j \equiv \begin{cases} \sum_{k=1}^{c} \lambda_k m_{0k} & j = 0 \\ 1 & 1 \leq j \leq d. \end{cases} \tag{1}$$

Next, define

$$n = \sum_{k=1}^{d} n_k \quad \text{and} \quad \beta_k = \frac{n_k}{n}. \tag{2}$$

Given finite base-stock levels $(n_1, \ldots, n_d)$, the fill rate of type k jobs, $1 \leq k \leq d$, will be some positive number strictly less than one. (For make-to-stock products, the *fill rate* is defined to be the fraction of orders that are filled from inventory.) Denoting type k fill rate by α_k, the actual throughput rates are calculated from $\lambda_k \alpha_k$ and the actual traffic intensities are given by

$$\rho_j^* = \begin{cases} \sum_{k=1}^{d} (\lambda_k \alpha_k) m_{0k} + \sum_{k=d+1}^{c} \lambda_k m_{0k} & j = 0 \\ \alpha_k & 1 \leq j \leq d. \end{cases} \tag{3}$$

One expects the fill rate of type k products to approach 1 as the type k base-stock level increases. In other words, let us fix the proportions of closed jobs $(\beta_1, \ldots, \beta_d) > 0$ while letting $n \to \infty$ (so that $n_k = \beta_k n \to \infty$ as well). Then $\alpha_k \to 1$ for each $k = 1, \ldots, d$ as $n \to \infty$. We may therefore view equation (1) as an initial approximation of (3) when n is large.

Let $Q_{0k}(t)$, $1 \leq k \leq c$, be the number of type k jobs at station 0 at time t. If in addition we define $Q_k(t)$ to be the number of jobs at station $1 \leq k \leq d$ at time t, then $Q_k(t) = n_k - Q_{0k}(t)$.

One interprets Q_{0k} as type k work-in-process inventory and Q_k as the finished-goods-inventory of type k products. We will show herein that when the workstation is roughly balanced, namely, if

$$\rho_0 = \sum_{k=1}^{c} \lambda_k m_{0k} \approx 1,$$

the following approximation holds when the base-stock level n is large:

$$Q_{0k}^n(\cdot) \equiv \frac{1}{n} Q_{0k}(n^2 \cdot) \approx \lambda_k W_0^\circ(t), \tag{4}$$

where $W_0^\circ(\cdot)$ is a reflected Brownian motion (RBM) on the interval $[0, b]$ having drift θ° and variance σ^2, with

$$b = \min\{m_1\beta_1, \ldots, m_d\beta_d\} \tag{5}$$

$$\theta^\circ = n(\rho_0 - 1) \tag{6}$$

$$\sigma^2 = \sum_{k=1}^{c} \lambda_k m_{0k}^2 \left(c_k^2 + c_{0k}^2\right). \tag{7}$$

The remainder of the paper is organized as follows. Section 1 presents the processes that are of interest in the heavy traffic limit result and the approximation procedure. Section 2 continues to set up for the statement of the limit theorem. It discusses centering and scaling conventions as well as the assumptions of heavy traffic. The formal limit theorem is stated in Section 3. The approximation procedure is then presented in Section 4 and its performance is investigated in Section 5. Lastly, Section 6 contains proofs of the main theorems in the paper.

1 The Network Equations

We will take as primitive the demand and service time processes for each customer type. Denote by $D_k(t)$ the number of demands for type k products by time t; one may think of D_k as a renewal process although we will not require this restriction. Let $V_{0k}(m)$ be the sum of the first m processing times of class k products, and let S_{0k} be the renewal process associated with the cumulative sums process V_{0k}; that is, $S_{0k}(t) = \max\{m : V_{0k}(m) \leq t\}$. For concreteness, one may think of $V_{0k}(m)$ as the sum of m independent and identically distributed random variables, in which case S_{0k} would be a renewal process, but again, we need not restrict ourselves to this situation. As explained in the introductory remarks, we will model demands of make-to-stock products by services completed at stations 1 to d. It is thus consistent with our notation to denote by V_k the cumulative sums process corresponding to the counting process D_k for $1 \leq k \leq d$.

Let $T_k(t)$ be the amount of time the workcenter (station 0) has spent processing class k jobs in the the interval $[0,t]$. Denoting by $B_j(t)$, the cumulative "busy" time at station j, we find that

$$B_0(t) = \sum_{k=1}^{c} T_k(t). \tag{8}$$

Define

$$A_{0k}(t) \equiv \begin{cases} D_k(B_k(t)) & 1 \le k \le d, \\ D_k(t) & d+1 \le k \le c, \end{cases} \tag{9}$$

and

$$A_k(t) \equiv S_{0k}(T_k(t)), \qquad 1 \le k \le d. \tag{10}$$

For $d+1 \le k \le c$, $A_{0k}(t)$ is simply the number of demands for make-to-order products of type k to enter the workcenter by time t. For $1 \le k \le d$, one interprets $A_{0k}(t)$ as the number type k make-to-stock demands that have been filled from inventory; equivalently, it is the number of production requests to enter the workcenter by time t. Similarly, $A_k(t)$ is the number of items to enter the type k FGI by time t. Next, define

$$M_{0k}(t) \equiv V_{0k}(A_{0k}(t)), \qquad M_k(t) \equiv V_k(A_k(t)), \tag{11}$$

and set

$$L_j(t) \equiv \begin{cases} \sum_{k=1}^{c} M_{0k}(t) & j = 0 \\ M_j(t) & 1 \le j \le d, \end{cases} \tag{12}$$

with the interpretation of $L_j(t)$ as the amount of work (or load) that has arrived to station j by time t.

Let us assume that initially the FGI of each product type is full and the workcenter contains no production requests:

$$Q_k(0) = n_k, \quad 1 \le k \le d, \quad \text{and} \quad Q_{0k}(0) = 0, \quad 1 \le k \le c. \tag{13}$$

Denote by $W_j(t)$ the sum of all remaining processing times associated with those jobs found at station j at time t (this quantity will be referred to as the "work" at station j). Define the netflow process

$$X_j(t) \equiv W_j(0) + L_j(t) - t. \tag{14}$$

Letting $I_j(t)$ be the cumulative idle time of station j, it follows from the previous definitions that

$$I_j(t) = t - B_j(t) \tag{15}$$

and

$$W_j(t) \equiv W_j(0) + L_j(t) - B_j(t) = X_j(t) + I_j(t). \tag{16}$$

The first-in-first-out service discipline is work-conserving, hence we may make the following additional statements regarding the idleness processes:

$$I_j \text{ is continuous and nondecreasing with } I_j(0) = 0, \tag{17}$$

$$I_j \text{ increases only at times } t \text{ with } W_j(t) = 0. \tag{18}$$

Next, let $\eta(t)$ be the arrival time of the customer in service at station 0 at time t if there is a customer in service, and set it to be t otherwise. It is a consequence of the first-in-first-out policy that the number of type k departures from the workstation at time t is given by $A_k(\eta(t)) - \delta_k(t)$, where $\delta_k(t)$ is 1 if the job currently in service at the workstation belongs to type k and is 0 otherwise. The job count processes at the workstation are thus given by

$$Q_{0k}(t) = A_k(t) - A_k(\eta(t)) + \delta_k(t), \quad 1 \le k \le c, \tag{19}$$

from which we obtain the number of items in the finished-goods-inventory:

$$Q_k(t) = n_k - Q_{0k}(t), \quad 1 \le k \le d. \tag{20}$$

Moreover, the allocation processes $T_k(t)$ obey a similar property:

$$T_k(t) = M_{0k}(\eta(t)) + \epsilon_{1k}(t), \quad 1 \le k \le c, \tag{21}$$

where $\epsilon_{1k}(t)$ is the amount of service the current jobs has received if that job is of class k and $\epsilon_{1k}(t) = 0$ otherwise. Finally, observe that

$$\eta(t) = t - W_0(\eta(t)) + \epsilon_2(t), \tag{22}$$

with $\epsilon_2(t)$ being 0 if $W_0(t) = 0$ and otherwise being equal to the remaining service time of the customer currently occupying station 0.

2 Centering and Scaling

To state the heavy traffic theorem we need to express the processes defined in section 1 in terms of processes that have been "centered" and "scaled." Denoting by $\lfloor x \rfloor$ the integer part of x, let us define the following centered processes for $k = 1, \ldots, c$:

$$\begin{aligned}
\hat{V}_{0k}(t) &\equiv V_{0k}(\lfloor t \rfloor) - m_{0k}\lfloor t \rfloor & \hat{S}_{0k}(t) &\equiv S_{0k}(t) - m_{0k}^{-1}t \\
\hat{A}_{0k}(t) &\equiv A_{0k}(t) - \lambda_k t & \hat{M}_{0k}(t) &\equiv M_{0k}(t) - \lambda_k m_{0k} t \\
\hat{D}_k(t) &\equiv D_k(t) - \lambda_k t & \hat{T}_k(t) &\equiv T_k(t) - \lambda_k m_{0k} t,
\end{aligned} \tag{23}$$

and for $k = 1, \ldots, d$:

$$\hat{V}_k(t) \equiv V_k(\lfloor t \rfloor) - m_k \lfloor t \rfloor. \tag{24}$$

From equations (9)–(11), (21) and (23)–(24), we can write the centered allocation processes as

$$\begin{aligned} \hat{T}_k(t) &= \hat{V}_{0k}(D_k(B_k(\eta(t)))) + m_{0k}\hat{D}_k(B_k(\eta(t))) - \lambda_k m_{0k} I_k(\eta(t)) - \\ &\quad \lambda_k m_{0k} W_0(\eta(t)) + \lambda_k m_{0k}\epsilon_2(t) + \epsilon_{1k}(t) \\ &= \hat{V}_{0k}(D_k(B_k(\eta(t)))) + m_{0k}\hat{D}_k(B_k(\eta(t))) - \\ &\quad \lambda_k m_{0k}\left[X_0(\eta(t)) + I_k(\eta(t)) + I_0(\eta(t))\right] + \lambda_k m_{0k}\epsilon_2(t) + \epsilon_{1k}(t). \end{aligned} \tag{25}$$

Next, set

$$\xi_j(t) \equiv \begin{cases} \sum_{k=1}^{d}\left[\hat{V}_{0k}(D_k(B_k(t))) + m_{0k}\hat{D}_k(B_k(t))\right] + \\ \quad \sum_{k=d+1}^{c}\left[\hat{V}_{0k}(D_k(t)) + m_{0k}\hat{D}_k(t)\right] + (\rho_0 - 1)\,t & j = 0, \\ W_j(0) + \hat{V}_j(S_{0j}(T_j(t))) + m_j\hat{S}_{0j}(T_j(t)) + \frac{m_j}{m_{0j}}\hat{V}_{0j}(D_j(B_j(\eta(t))))+ \\ \quad m_j\hat{D}_j(B_j(\eta(t))) - \xi_0(\eta(t)) + \epsilon_2(t) + \frac{m_j}{m_{0j}}\epsilon_{1j}(t) & 1 \le j \le d. \end{cases} \tag{26}$$

With equations (25)–(26) and the observation $I_0(t) = I_0(\eta(t))$, we can now write the netflow and workload processes in the following form:

$$X_j(t) = \begin{cases} \xi_0(t) - \sum_{k=1}^{d}\lambda_k m_{0k} I_k(t) & j = 0 \\ \xi_j(t) - I_0(t) - I_j(\eta(t)) + \sum_{k=1}^{d}\lambda_k m_{0k} I_k(\eta(t)), & 1 \le j \le d, \end{cases} \tag{27}$$

and

$$W_j(t) = \begin{cases} \xi_0(t) + I_0(t) - \sum_{k=1}^{d}\lambda_k m_{0k} I_k(t) & j = 0, \\ \xi_j(t) - I_0(t) + \sum_{k=1}^{d}\lambda_k m_{0k} I_k(\eta(t)) + [I_j(t) - I_j(\eta(t))] & 1 \le j \le d. \end{cases} \tag{28}$$

We are interested in characterizing these processes when the base stock levels n_k are large and the workcenter is approximately balanced. To do so, we will prove a limit theorem for a *sequence* of networks whose parameters obey these conditions. Let us then consider a sequence of networks that are indexed by n. We fix the relative proportions of base stock inventories $\beta_1, \ldots, \beta_d$, but the level of the base-stock as well as mean processing times and demand rates vary along the sequence. The n^{th} system has a base stock level of $n_k = n\beta_k$ for type k products, and the sum of all base-stock levels is given by n. Of course it makes sense only to consider integer values of n_k, and we can do so by defining n_k to be the integer part plus one of $n\beta_k$ (so we always have positive base-stock levels) with the corresponding total base-stock level being the sum of all n_k's. In the interest of keeping the exposition simple, however, let us

proceed with the n_k's are originally defined while keeping in mind that this is without any loss of representation power.

We will denote parameters and processes associated with the n^{th} system by a superscript "(n)". For example, the demand rates and mean processing times of type k products in the n^{th} system are $\lambda_k^{(n)}$ and $m_{0k}^{(n)}$, respectively. The traffic intensities $\rho_j^{(n)}$ are then defined exactly as in (1) with $\lambda_k^{(n)}$ and $m_{0k}^{(n)}$ in place of λ_k and m_{0k}. We require that $\lambda_k^{(n)} \to \lambda_k$, $m_{0k}^{(n)} \to m_{0k}$ as $n \to \infty$, and the following condition holds for some finite θ:

$$n\left(\rho_0^{(n)} - 1\right) \to \theta \qquad \text{as } n \to \infty. \tag{29}$$

That is, we are interested in the regime where base stock levels are high and the traffic intensity at the workcenter is approximately one.

The heavy traffic limit theorem described in the next section applies to processes whose space and time dimensions have been scaled. Let us describe here the three scaling conventions that will be used. For a "generic" process $X^{(n)}(t)$, set

$$X^n(t) \equiv \frac{1}{n} X^{(n)}(n^2 t), \quad \bar{X}^n(t) \equiv \frac{1}{n} X^{(n)}(nt), \quad \bar{\bar{X}}^n(t) \equiv \frac{1}{n^2} X^{(n)}(n^2 t). \tag{30}$$

We will refer to the first two scaling conventions as the "diffusion scale" and the "fluid scale," respectively. It is understood that, for example, $X^n(t)$ is the process associated with n^{th} system whose state space has been aggregated by a factor of n and whose time dimension has been accelerated by a factor of n^2. One can verify that the *diffusion scaled* workload processes now take the following form:

$$W_j^n(t) = \begin{cases} \xi_0^n(t) + I_0^n(t) - \sum_{k=1}^d \lambda_k^{(n)} m_{0k}^{(n)} I_k^n(t) & j = 0, \\ \xi_j^n(t) - I_0^n(t) + \sum_{k=1}^d \lambda_k^{(n)} m_{0k}^{(n)} I_k^n(\bar{\bar{\eta}}^n(t)) + \left[I_j^n(t) - I_j^n(\bar{\bar{\eta}}^n(t)) \right] & 1 \le j \le d, \end{cases} \tag{31}$$

where

$$I_j^n \text{ is continuous and nondecreasing with } I_j^n(0) = 0, \tag{32}$$

$$I_j^n \text{ increases only at times } t \text{ with } W_j^n(t) = 0, \tag{33}$$

and

$$\xi_j^n(t) \equiv \begin{cases} \sum_{k=1}^d \left[\hat{V}_{0k}^n(\bar{\bar{D}}_k^n(\bar{\bar{B}}_k^n(t))) + m_{0k}^{(n)} \hat{D}_k^n(\bar{\bar{B}}_k^n(t)) \right] + \\ \quad \sum_{k=d+1}^c \left[\hat{V}_{0k}^n(\bar{\bar{D}}_k^n(t)) + m_{0k}^{(n)} \hat{D}_k^n(t) \right] + n\left(\rho_0^{(n)} - 1\right) t & j = 0, \\ W_j^n(0) + \hat{V}_j^n(\bar{\bar{S}}_{0j}^n(\bar{\bar{T}}_j^n(t))) + m_j^{(n)} \hat{S}_{0j}^n(\bar{\bar{T}}_j^n(t)) + \\ \quad \frac{m_j^{(n)}}{m_{0j}^{(n)}} \hat{V}_{0j}^n(\bar{\bar{D}}_j^n(\bar{\bar{B}}_j^n(\bar{\bar{\eta}}^n(t)))) + m_j^{(n)} \hat{D}_j^n(\bar{\bar{B}}_j^n(\bar{\bar{\eta}}^n(t))) - \\ \quad \xi_0^n(\bar{\bar{\eta}}^n(t)) + \epsilon_2^n(t) + \frac{m_j^{(n)}}{m_{0j}^{(n)}} \epsilon_{1j}^n(t) & 1 \le j \le d. \end{cases} \tag{34}$$

3 The Heavy Traffic Limit Theorem

The setting here is the space $\mathbf{D}^{d+1}$, the $d+1$-dimensional product space of functions $f : [0,\infty) \to \Re^{d+1}$ that are right continuous and have left limits. The space $\mathbf{D}^{d+1}$ is endowed with the Skorohod J_1 topology. For a sequence of processes X^n in $\mathbf{D}^{d+1}$ and $X \in \mathbf{D}^{d+1}$, the symbol "$X^n \Longrightarrow X$" means "X^n converges to X in distribution." Moreover, we write "$X^n \to X$ u.o.c." if almost surely, X^n converges to X uniformly on compact sets (see Billingsley 1968).

In addition to the heavy traffic condition (29), we make the following assumptions regarding the primitive demand and service processes.

Assumption 1 *As $n \to \infty$, $\bar{W}_j^n(0) \to m_j\beta_j$ for $1 \le j \le d$ (recall that $W_0^n(0) \equiv 0$), and $(\bar{V}_0^n, \bar{D}^n) \to (\bar{V}_0^*, \bar{D}^*)$ u.o.c., where $\bar{V}_{0k}^*(t) = m_{0k}t$ and $\bar{D}_k^*(t) = \lambda_k t$.*

Here, $\bar{V}_0^n$, $\bar{D}^n$, $\bar{V}_0^*$, and $\bar{D}^*$ are vector processes defined in the obvious way. Henceforth, we will similarly write processes in vector form and assume that their meaning will be clear without further comment or definition.

Assumption 2 *As $n \to \infty$, $(\hat{V}_0^n, \hat{D}^n) \Longrightarrow (\hat{V}_0^*, \hat{D}^*)$, where $\hat{V}_{0k}^*(t)$ is $(0, m_{0k}^2 c_{0k}^2)$ Brownian motion, $\hat{D}_k^*(t)$ is $(0, \lambda_k c_k^2)$ Brownian motion, and all component processes are independent.*

The assumptions above hold, for example, if inter-demand times and service times of all product types are independent sequences of independent and identically distributed random variables.

Recall the the definitions of b, σ^2 and θ from (5), (7), and (29), respectively. We are now ready to state the main result:

Theorem 1 *Under the heavy traffic condition (29) and Assumptions 1 and 2,*

$$(W^n, I^n, Q_0^n) \Longrightarrow (W^*, I^*, Q_0^*),$$

where

$$W_0^*(t) = \xi_0^*(t) + I_0^*(t) - \sum_{j=1}^{d} \lambda_j m_{0j} I_j^*(t),$$

$$W_j^*(t) = m_j\beta_j - W_0^*(t), \quad 1 \le j \le d,$$

ξ_0^* is Brownian motion with drift θ and variance σ^2,

I_j^* is continuous and nondecreasing, $\quad 0 \le j \le d$,

I_j^* increases only at times t when $W_j^*(t) = 0$, $\quad 0 \le j \le d$,

$$Q_{0k}^*(t) = \lambda_k W_0^*(t), \quad 1 \le k \le c.$$

Remark: One can equivalently write

$$W_0^*(t) = \xi_0^*(t) + I_0^*(t) - Y^*(t),$$

where ξ_0^* and I_0^* are as defined in the statement of the theorem and

$$Y^*(t) \equiv \sum_{j=1}^{d} \lambda_j m_{0j} I_j^*(t).$$

Thus defined, Y^* is continuous and nondecreasing with $Y^*(0) = 0$, and Y^* increases only at times t when $W_0^*(t) = m_j\beta_j$ for some $j = 1, \ldots, d$. Because Brownian motion is continuous, we see that the process Y will increase whenever W_0^* reaches the *smallest* of the parameters $m_j\beta_j$, $j = 1, \ldots, d$. In this form, one recognizes W_0^* as one-dimensional reflected Brownian motion on the interval $[0, b]$ with drift θ and variance σ^2, where $b = \min\{m_1\beta_1, \ldots, m_d\beta_d\}$.

4 The Approximation Procedure

Let us now interpret the heavy traffic limit theorem (Theorem 1) in terms of the production inventory model, which will help us to develop an approximation procedure. Recall our notation for describing the production process: demands for type k products occur with rate λ_k and SCV c_k^2, and production times for type k products have mean m_k and SCV c_{0k}^2 ($k = 1, \ldots, c$). Let n_k be the base-stock level of make-to-stock type k products ($k = 1, \ldots, d$); define n to be the sum of all base-stock levels n_k; and set β_k to be the ratio of n_k to the sum n. *Theorem 1 suggests that when the total base-stock level n becomes large, the behavior of the scaled workload process at the production center is approximately that of a one-dimensional reflected Brownian motion (RBM) on an interval;* that is,

$$\frac{1}{n} W_0(n^2 \cdot) \approx W_0^\circ(\cdot),$$

where W_0° is an RBM on the interval $[0, b]$ with drift θ° and variance σ^2, defined as in the introductory remarks:

$$\begin{aligned} b &= \min\{m_1\beta_1, \ldots, m_d\beta_d\}, \\ \theta^\circ &= n(\sum_{k=1}^{c} \lambda_k m_{0k} - 1), \\ \sigma^2 &= \sum_{k=1}^{c} \lambda_k m_{0k}^2 (c_k^2 + c_{0k}^2). \end{aligned}$$

As explained in Harrison and Nguyen (1993), we simply "reverse" the scaling to obtain an approximation for the original workload process W_0. It is straightforward to verify that this procedure results in the approximation of W_0 by $\tilde{W}_0$:

$$W_0(\cdot) \approx \tilde{W}_0(\cdot),$$

where $\tilde{W}_0$ is an RBM on the interval $[0, nb]$ with

$$nb = \min\{m_1 n_1, \ldots, m_d n_d\},$$

whose drift is

$$\mu \equiv \sum_{k=1}^{c} \lambda_k m_{0k} - 1, \tag{35}$$

and whose variance is again σ^2 (as defined in (7)).

We arrive at an approximation that treats the workload at the production center as a *bounded* process, even though the *actual* workload may become arbitrarily large. On an intuitive level, one can justify the approximation with the following argument. For the sake of simplicity, and without loss of generality, let us for now consider the case with one make-to-stock product ($d = 1$). When all n make-to-stock products are queued at the production center, the finished-goods-inventory must be empty. All make-to-stock demands that occur during this time are thus lost, and no new production requests can be initiated. Therefore, during this period of time, the total rate at which work enters the workstation falls below the critical level, and the production center no longer operates in the heavy traffic regime. The make-to-stock products thus have an effect of regulating the production center's workload and preventing it from becoming "too large" relative to its base-stock level. (For more discussion on this issue, readers may refer to Nguyen 1994.)

Continuing with Theorem 1, we find that the job count process is proportional to the workload where the proportionality constant is given by the demand rate; that is we have the approximation

$$Q_{0k}(\cdot) \approx \lambda_k \tilde{W}_0(\cdot).$$

Moreover, we find that Q_k, the finished-goods-inventory of type k products, $1 \leq k \leq d$, is approximated by a reflected Brownian motion on the interval

$$[\frac{1}{m_k}(m_k n_k - nb), n_k]$$

(recall that $\lambda_k^{-1} = m_k$). If k corresponds to the index such that $m_k n_k = nb$, then this states that the type k inventory process approximately follows that of a Brownian motion and may

assume values between 0 and n_k. For a type k product such that $m_k n_k > nb$, however, the approximation of the inventory process is a Brownian motion that is bounded strictly away from zero. That is, in our approximation, such a product type *never* stocks out!

One can explain the "intuition" of this heavy traffic result in the same way that we explained the boundedness of the limiting workload process: Whenever one of the products stocks out, the total arrival rate of work falls below the critical level; the production center essentially has temporary "excess" capacity; it therefore is able to keep all other inventories adequately replenished. Let us interpret $m_k n_k$ as the *expected* duration of time that a full inventory of type k items can satisfy demands without replenishment, and let us refer to this duration as the "buffer time." In the heavy traffic scaling and the eventual limit, it turns out that the product type that *first* stocks out is the one having the smallest buffer time, and consequently, it is the *only* product type to ever stock out. We will call this the "bottleneck product type."

For performance analysis purposes, such a result appears woefully inadequate. As we will see in the next section, where we present some numerical results, it turns out that the heavy traffic approximations for *queue lengths* are quite good. In order to offer a sharper result for *throughput rates*, we will devote some attention to the discussion of bounds. Before doing so, however, let us review the approximation procedure that is suggested by the heavy traffic limit theorem (Theorem 1). First, the workload process at the workcenter is approximated by $\tilde{W}_0$, an RBM whose parameters we have specified. Second, the inventory level of each product type is approximated via the linear relationship $Q_{0k}(\cdot) \approx \lambda_k \tilde{W}_0(\cdot)$. Third, a make-to-stock product of type k will risk stocking out if and only if $m_k n_k = \min\{m_1 n_1, \ldots, m_d n_d\}$.

For illustration, let us suppose that $m_1 n_1 < m_k n_k$ for all $k = 2, \ldots, d$. It follows from Theorem 1 and our interpretations that the RBM $\tilde{W}_0$ is given by

$$\tilde{W}_0(t) = \tilde{\xi}_0(t) + \tilde{I}_0(t) - \lambda_1 m_{01} \tilde{I}_1(t),$$

where $\tilde{\xi}_0$ is Brownian motion with drift μ (defined in equation (35)) and variance σ^2; $\tilde{I}_0$ is the approximating idleness process at the workcenter; and $\tilde{I}_1$ is the approximating idleness process at type 1 finished-goods-inventory. Alternatively, $\tilde{I}_0$ and $\tilde{I}_1$ are the lower and upper regulators, respectively, of the RBM $\tilde{W}_0$ on the interval $[0, m_1 n_1]$. The approximating idleness process at any other finished-goods-inventory is simply zero for all times t. If there are more than one product type achieving the minimum "buffer time," we propose that the idleness process for each of these product types be approximated in the same manner. For example, if $m_1 n_1 = m_2 n_2 = \min\{m_1 n_1, \ldots, m_d n_d\}$, then the approximating idleness process for product type j, $j = 1, 2$ obeys

$$\tilde{W}_0(t) = \tilde{\xi}_0(t) + \tilde{I}_0(t) - \lambda_j m_{0j} \tilde{I}_j(t).$$

Because characterization of all c product types can be collapsed into a one-dimensional RBM, namely, $\tilde{W}_0$, one can explicitly calculate many performance measures of interest, including the fill rates and average inventory levels of each product type. The formulas involved in these calculations are those obtained from steady-state analysis of reflected Brownian motion on an interval, and an excellent reference of this material is Harrison (1985). The details for translating those formulas into performance measures of the production system are presented in Nguyen (1994). In addition, Nguyen (1994) discusses some enhancements to the approximation method, which we will apply here in our calculation. Because these calculations and modifications are now standard in the literature of heavy traffic approximation of networks, we will not repeat them here, and simply point readers to references such as Dai and Harrison (1992), Harrison and Nguyen (1990 and 1993), and Nguyen (1994) for details. We end this section with a discussion of some possible bounds, beginning with the lower bound.

Lower Bound: For illustration, let us suppose that there are two make-to-stock product types and one make-to-order product type ($d = 2$ and $c = 3$), numbered so that $m_1 n_1 < m_2 n_2$. A straightforward application of the heavy traffic limit result would indicate that type 2 products experience 100% fill rate (recall that the fill rate is defined to be the ratio of achieved throughput to demand rate). Let us now consider a variant of the system, called the "alternate system," in which type 1 products are modeled as make-to-order, or *open*, jobs, and the remaining product types are as before. That is, we now have a mixed network with one type of make-to-stock products and two types of make-to-order products. The results of Theorem 1 can be applied to obtain an approximation for the fill rate of type 2 products in the alternate system, and we propose that this estimate be used as a *lower* bound for its actual throughput rate.

To see why such an estimate might serve as a lower bound, note that the average rate at which work enters the workcenter will be higher in the alternate system. In particular, type 1 throughput rate is now simply its demand rate. The utilization of the workcenter will be higher in the alternate system, and so we expect that type 2 products will achieve *lower* throughput than in the original system.

We described the above procedure in the context of a production system with two make-to-stock products and one make-to-order product, but the extension to multiple make-to-order products requires no modification. To extend the procedure to a system that serves several ($d > 2$) types of make-to-stock jobs, one simply applies $d - 1$ iterations of the procedure; at each iteration, the bottleneck product type from the previous iteration is moved to the make-to-order category; this reveals a new bottleneck product type whose throughput rate can then be calculated.

Upper Bound: Again, we will describe the procedure by considering the system with two make-to-stock product types, one make-to-order product type, and $m_1 n_1 < m_2 n_2$. (The extension to the general case is then done exactly as described before.) The procedure is in essence similar to that described for the lower bound; that is, we estimate the throughput rate of type 2 products by considering an alternate system in which type 1 products are "open." However, rather than using the demand rate (λ_1) for type 1 products, we substitute it with the *estimated throughput* rate. The rate at which work arrives to the workcenter in this alternate system is approximately the same as that of the original system, but the arrival process has been made less variable. The original system experiences "bursts" of relatively quick arrivals of type 1 jobs when the type 1 finished-goods-inventory is not empty, alternated with periods of no arrivals when type 1 finished-goods-inventory is depleted. In the alternate system, on the other hand, type 1 jobs arrive "regularly" according to a renewal process, but its arrival rate has been slowed down to achieve the same throughput rate. Because the traffic intensity remains the same while the total variability in the system has been decreased, we propose that this system be used as an *upper* bound for the throughput rate of type 2 jobs.

5 Numerical Results

This section investigates the performance of the approximation procedure described in the previous section. We will consider product-form networks with one make-to-order product type and two make-to-stock product types. It is known that such a mixed network is product-form if, for example, all demand processes are Poisson, all processing times are exponentially distributed, and processing times for all product types at the workstation have the same mean (i.e., they all have the same distribution); see Kelly (1979). For such a network, one can derive the steady-state distribution of queue lengths, from which one can calculate performance measures such as average queue length and throughput rate.

The parameters of the systems we study are shown in Table 1. For each system shown in Table 1 we will consider 12 different base-stock levels (we thus have 24 cases in total). As convention, we will always take $m_1 n_1 \leq m_2 n_2$. To ensure that the network is product-form, we will assume that the demand process of open jobs is Poisson and that all service times are exponentially distributed. Moreover, all job types have the same processing time distribution at the workcenter. It can be verified that the relative traffic intensity of the workcenter in System 1 is 0.975 and equals 0.99 for System 2. If the relative traffic intensity of the workcenter is far from one, the fill rates of closed jobs are essentially 100% even for small base-stock levels. For purposes of testing the approximation, we therefore consider the more challenging cases where

System	System Parameters			
	λ	$m_{01} = m_{02} = m_{03}$	m_1	m_2
1	2.0	0.06	0.125	0.16
2	10.0	0.033	0.10	0.10

Table 1: System Parameters

the relative traffic intensity of the workcenter is close to one.

Table 2 contains the fill rate approximations of each of the closed job types in System 1. The first two columns contain the base-stock levels of the two job types. The next three columns show the heavy traffic approximation of the fill rate of type 1 jobs, the exact fill rate (computed from the product-form solutions), and the percent error of the approximation (computed by dividing the difference between the approximated and exact fill rates by the exact fill rate). The next two columns display the lower and upper bounds of type 2 fill rates. The gap between the two bounds in shown in the next to last column (this is computed by subtracting the lower bound from the upper bound). The exact fill rate of type 2 closed jobs, again computed using the product-form formulas, is shown in the last column.

The ratio $m_1 n_1 / m_2 n_2$ of the first three systems (that is, the first three rows) equals 0.39; the corresponding ratios in the next three sets of systems are 0.42, 0.625, and 0.78, respectively. If $m_1 n_1$ is much smaller than $m_2 n_2$, then the heavy traffic characterization that type 2 jobs "never" stocks out is more likely to hold true, and we expect the heavy traffic approximations to have better performance. We also expect to have better approximations as the base-stock levels n_1 and n_2 increase. For System 1, Table 1 shows that the heavy traffic approximation of type 1 fill rates performs well under all cases. The bounds for type 2 fill rates always contains the exact fill rates. Moreover, the gap between the upper and lower bound is typically less than 3% for moderately large base-stock levels (e.g., $n_1 \geq 20$ and $n_2 \geq 20$).

Tables 3 and 4 display queue length approximations for the same system with the same groupings of parameters. Table 3 contains the average work-in-process (WIP) of make-to-order jobs at the workcenter and the average finished-goods-inventory of type 1 jobs. Again, the performance of the approximations is uniformly good and becomes better as the base-stock levels increase. Table 4 contains the average finished-goods inventory of type 2 jobs. The two heavy traffic approximations correspond to the lower bound and upper bound approximations of the fill rate. (Recall that the inventory of a given job type is related to the underlying reflected Brownian motion via its throughput rate.) The bounds for fill rates do not carry over to become bounds for queue lengths, as can be seen in Table 4. Neither of the approximations

		Fill Rate: Type 1 Jobs			Fill Rate: Type 2 Jobs			
n_1	n_2	HT	Exact	% Err	Lower Bound	Upper Bound	Gap	Exact
5	10	85.7%	85.7%	0.0%	93.2%	99.8%	6.6%	99.3%
10	20	93.3%	93.1%	0.2%	97.7%	100.0%	2.2%	100.0%
15	30	96.0%	95.9%	0.2%	99.0%	100.0%	1.0%	100.0%
16	30	96.4%	96.2%	0.2%	99.0%	100.0%	1.0%	100.0%
32	60	98.9%	98.8%	0.1%	99.9%	100.0%	0.1%	100.0%
48	90	99.6%	99.5%	0.0%	100.0%	100.0%	0.0%	100.0%
16	20	96.4%	96.4%	0.0%	97.7%	99.9%	2.2%	99.7%
32	40	98.9%	98.8%	0.1%	99.5%	100.0%	0.4%	100.0%
48	60	99.6%	99.5%	0.0%	99.9%	100.0%	0.1%	100.0%
10	10	93.3%	94.4%	-1.2%	93.2%	99.2%	6.0%	97.3%
20	20	97.4%	97.7%	-0.3%	97.7%	99.8%	2.1%	99.2%
30	30	98.7%	98.8%	-0.1%	99.0%	99.9%	0.9%	99.7%

Table 2: Fill Rates in System 1

does uniformly better than the other, but both do perform well under all of the cases considered for System 1.

Performance measures for System 2 are contained in Tables 5 – 7, displayed in the same format as that of System 1. The first grouping corresponds to systems where $m_1 n_1 = m_2 n_2$. In this case, the procedure described in Section 4 allows us to obtain performance estimates for both type 1 and type 2 jobs. Moreover, because the two job types are symmetrical, they have the same performance measures and we therefore need to report only one set of numbers for both job types.

We expect that the heavy traffic approximation would have the most difficulty in this range of parameters because there is no single "bottleneck" job type. The figures in Table 5 indicate that the approximation of fill rates are still quite good and becomes much better for larger base-stock levels. Tables 6 and 7 show that the approximations of queue lengths can have large errors for small values of n_1 and n_2, but that their performance moves to the more reasonable range as n_1 and n_2 increase.

In the next grouping, we fix the base-stock level of type 1 jobs while we vary the base-stock level of type 2 jobs. (We still require $m_1 n_1 \leq m_2 n_2$.) As n_2 increases, the heavy traffic characterization that type 2 jobs "never" stock out becomes more representative of the actual dynamics, and we see that the performance of the approximation does increase in this direction.

		Make-to-Order WIP			FGI at Station 1		
n_1	n_2	HT	Exact	% Err	HT	Exact	% Err
5	10	0.71	0.78	-9.1%	2.57	2.67	-3.8%
10	20	1.24	1.30	-4.3%	5.36	5.51	-2.7%
15	30	1.73	1.77	-2.4%	8.37	8.57	-2.4%
16	20	1.82	1.82	-0.1%	8.99	9.26	-2.9%
32	40	3.01	2.99	0.6%	20.10	20.41	-1.5%
48	60	3.78	3.74	1.1%	32.94	33.24	-0.9%
16	30	1.82	1.85	-1.7%	8.99	9.20	-2.3%
32	60	3.01	2.99	0.6%	20.10	20.40	-1.5%
48	90	3.78	3.74	1.1%	32.94	33.24	-0.9%
10	10	1.24	1.20	3.7%	5.36	5.73	-6.5%
20	20	2.16	2.10	2.9%	11.58	12.02	-3.6%
30	30	2.89	2.83	2.0%	18.61	19.04	-2.3%

Table 3: Queue Lengths in System 1: Types 1 and 3

		FGI at Station 2				
n_1	n_2	HT 1	HT 2	Exact	% Err 1	% Err 2
5	10	7.88	7.79	7.60	3.7%	2.5%
10	20	16.17	16.11	15.93	1.5%	1.1%
15	30	24.63	24.60	24.48	0.6%	0.5%
16	20	14.40	14.32	14.34	0.4%	-0.1%
32	40	30.62	30.60	30.67	-0.2%	-0.2%
48	60	48.19	48.18	48.30	-0.2%	-0.2%
16	30	24.35	24.32	24.21	0.6%	0.4%
32	60	50.60	50.60	50.64	-0.1%	-0.1%
48	90	78.18	78.18	78.30	-0.1%	-0.1%
10	10	6.28	6.13	6.47	-3.0%	-5.2%
20	20	13.34	13.26	13.54	-1.5%	-2.1%
30	30	21.04	20.99	21.23	-0.9%	-1.1%

Table 4: Queue Lengths in System 1: Type 2

		Fill Rate: Type 1 Jobs			Fill Rate: Type 2 Jobs			
n_1	n_2	HT	Exact	% Err	Lower Bound	Upper Bound	Gap	Exact
10	10	92.3%	95.2%	-3.1%				
20	20	96.6%	97.9%	-1.4%				
30	30	98.0%	98.8%	-0.8%		Refer to Type 1		
40	40	98.8%	99.3%	-0.5%				
50	50	99.2%	99.5%	-0.3%				
10	10	92.3%	95.2%	-3.1%	91.5%	97.3%	5.8%	95.2%
10	20	92.3%	92.4%	-0.1%	96.4%	99.6%	3.1%	99.8%
10	30	92.3%	92.2%	0.1%	98.0%	99.9%	1.9%	100.0%
10	40	92.3%	92.2%	0.1%	98.7%	100.0%	1.2%	100.0%
10	50	92.3%	92.2%	0.1%	99.2%	100.0%	0.8%	100.0%
10	50	92.3%	92.2%	0.1%	99.2%	100.0%	0.8%	100.0%
20	50	96.6%	96.5%	0.1%	99.2%	99.9%	0.8%	100.0%
30	50	98.0%	98.0%	0.0%	99.2%	99.9%	0.7%	100.0%
40	50	98.8%	98.9%	-0.1%	99.2%	99.8%	0.7%	99.8%
50	50	99.2%	99.5%	-0.3%	99.2%	99.8%	0.6%	99.5%

Table 5: Fill Rates in System 2

(As the parameters enter the region where $m_1 n_1$ is strictly less than $m_2 n_2$, the approximation markedly improves.)

This set of numbers highlights one limitation of the heavy traffic approach: Approximations of all type 1 performance measures remain constant for differing values of n_2. Recall from the discussion of Section 4 that whenever the condition $m_1 n_1 < m_2 n_2$ is satisfied, the approximation essentially treats n_2 as being infinitely large, so the specific value of n_2 has no effect on the approximation. Although we have not attempted to introduce a refinement in this paper, it appears from the numbers in Tables 5 – 7 that the approach performs well under all tested circumstances.

The last grouping fixes n_2 and varies n_1 while still maintaining the condition $m_1 n_1 \leq m_2 n_2$. Consistent with other scenarios discussed above, the approximations are good in all cases.

		Make-to-Order WIP			FGI at Station 1		
n_1	n_2	HT	Exact	% Err	HT	Exact	% Err
10	10	5.17	4.59	12.5%	5.23	5.84	-10.4%
20	20	9.36	8.53	9.8%	10.96	11.84	-7.5%
30	30	13.08	12.10	8.0%	17.18	18.21	-5.7%
40	40	16.33	15.30	6.8%	23.87	24.97	-4.4%
50	50	19.16	18.12	5.8%	31.00	32.11	-3.5%
10	10	5.17	4.59	12.5%	5.23	5.84	-10.4%
10	20	5.17	5.48	-5.6%	5.23	5.32	-1.7%
10	30	5.17	5.53	-6.6%	5.23	5.30	-1.3%
10	40	5.17	5.53	-6.6%	5.23	5.30	-1.3%
10	50	5.17	5.53	-6.6%	5.23	5.30	-1.3%
10	50	5.17	5.53	-6.6%	5.23	5.30	-1.3%
20	50	9.36	9.62	-2.7%	10.96	11.09	-1.2%
30	50	13.08	13.23	-1.1%	17.18	17.36	-1.1%
40	50	16.33	16.20	0.8%	23.87	24.24	-1.5%
50	50	19.16	18.12	5.8%	31.00	32.11	-3.5%

Table 6: Queue Lengths in System 2: Types 1 and 3

		FGI at Station 2				
n_1	n_2	HT 1	HT 2	Exact	% Err 1	% Err 2
10	10					
20	20					
30	30			Refer to Type 1		
40	40					
50	50					
10	10	5.13	4.92	5.84	-12.1%	-15.7%
10	20	14.95	14.84	14.56	2.7%	2.0%
10	30	24.90	24.83	24.47	1.8%	1.5%
10	40	34.87	34.83	34.47	1.2%	1.1%
10	50	44.86	44.83	44.47	0.9%	0.8%
10	50	44.86	44.83	44.47	0.9%	0.8%
20	50	40.69	40.64	40.38	0.8%	0.6%
30	50	37.00	36.93	36.78	0.6%	0.4%
40	50	33.76	33.69	33.88	-0.4%	-0.6%
50	50	30.95	30.87	32.11	-3.6%	-3.9%

Table 7: Queue Lengths in System 2: Type 2

6 Proof of The Heavy Traffic Limit Theorem

Theorem 1 is proved using the same methodology as in Sections 5 and 6 of Nguyen (1994), provided we can establish the existence, uniqueness, and continuity of a certain mapping. To state this result, let us first introduce the following notation. Fix $\epsilon > 0$, and let $\mathbf{D}_\epsilon^{d+1}$ be the set of functions $x \in \mathbf{D}^{d+1}$ such that (i) $x(0) \geq 0$, (ii) $x_0(t) + x_j(t) \geq \epsilon$ for each $j = 1, \ldots, d$ and $t \geq 0$, and (iii) x has a finite number of discontinuities over every finite time interval. We will denote by $\mathbf{C}^{d+1}$ and $\mathbf{C}_\epsilon^{d+1}$ the subset of those functions x in $\mathbf{D}^{d+1}$ and $\mathbf{D}_\epsilon^{d+1}$, respectively, that are continuous. Next, let Λ be the set of functions $a \in \mathbf{D}^1$ that have the following properties: (i) a is nondecreasing, (ii) $0 \leq a(t) \leq t$ for all $t \geq 0$, (iii) for each finite t, there is a finite number of subintervals $0 = s_0 < s_1 < \cdots < s_N = t$ and constants $0 = a_0 < a_1 < \cdots < a_N$ such that either $a(t) = t$ or $a(t) = a_i$ on the interval $[s_i, s_{i+1})$. In particular, observe that $e(t) \equiv t$ is an element of Λ.

Let $x \in \mathbf{D}_\epsilon^{d+1}$, $a \in \Lambda$, and $c_1, \ldots, c_d$ be positive numbers with $\sum_1^d c_k < 1$. We are interested in the mapping $(\Phi, \Psi) : (x, a) \to (w, z)$, where (w, z) are defined by the following:

$$\left[\begin{array}{l} w_0(t) = x_0(t) + y_0(t) - z(t) \\ w_j(t) = x_j(t) - y_0(t) + z(a(t)) + (y_j(t) - y_j(a(t))), \quad 1 \leq j \leq d \\ z(t) = \sum_{k=1}^d c_k y_k(t) \\ y_j \text{ is nondecreasing with } y_j(0) = 0, \quad 0 \leq j \leq d \\ y_j \text{ increases only at times } t \text{ where } w_j(t) = 0, \quad 0 \leq j \leq d. \end{array}\right] \tag{36}$$

Theorem 2 *The mapping (Φ, Ψ) is well defined on $\mathbf{D}_\epsilon^{d+1} \times \Lambda$. That is, for each $x \in \mathbf{D}_\epsilon^{d+1}$ and $a \in \Lambda$, there exists a unique pair of processes (w, z) that satisfies the set of equations (36). If $x \in \mathbf{C}_\epsilon^{d+1}$ and $a(t) = t$, then (Φ, Ψ) is continuous at (x, a). Moreover, y_j, and hence z, are continuous if $x_j - (y_j \circ a)$ has no jumps downwards for each $j = 1, \ldots, d$.*

Remark: Observe that we claim uniqueness for the process z only, and not for the individual processes y_j. To see that y_j are not necessarily unique, consider $d = 2$, $a(t) = t$, $x_0(t) = t + 1$, and $x_j(t) = -t$ for $j = 1, 2$.

We defer discussion of the proof of Theorem 2 to the end of this section. For now, with the aid of Theorem 2, one can follow exactly the same methods as in Sections 5 and 6 of Nguyen (1994) to prove Theorem 1. We will only outline the ideas of the proof here and refer interested readers to Nguyen (1994) for details. By Skorohod's representation theorem, let us first assume that the convergence of Assumption 2 holds u.o.c. The same arguments as in Nguyen (1994)

will show that $\overline{\overline{\eta}}^n \to e$ u.o.c. with $e(t) \equiv t$. From (34) and Assumptions 1 and 2, it follows that $\xi^n \to \xi^*$ u.o.c. where ξ_0^* is (θ, σ^2) Brownian motion and $\xi_j^*(t) + \xi_0^*(t) = m_j\beta_j$ for each $t \geq 0$ and $1 \leq j \leq d$. Observing that the workload and idleness processes can be expressed in terms ξ^n and $\overline{\overline{\eta}}^n$ via the mapping of Theorem 2: $(W^n, I^n) = (\Phi, \Psi)(\xi^n, \overline{\overline{\eta}}^n)$. Because Brownian motion is continuous, one then uses Theorem 2 together with the continuous mapping theorem to argue that convergence indeed takes place as claimed. That is, we have

$$(W^n, I^n) = (\Phi, \Psi)(\xi^n, \overline{\overline{\eta}}^n) \to (W^*, I^*) = (\Phi, \Psi)(\xi^*, \overline{\overline{\eta}}^*),$$

which is equivalent to the statements of Theorem 1.

Turning to the proof of Theorem 2, note that in the case $d = 1$, the statement is simply Theorem 9.1 of Nguyen (1994). We thus propose to prove Theorem 2 by induction on d in the following manner. To begin, observe that without loss of generality, we may assume $\epsilon = 1$. Next, suppose as usual that the theorem has been established on $\mathbf{D}_\epsilon^d$. The *existence* of a mapping (Φ, Ψ) on $\mathbf{D}_\epsilon^{d+1}$ is shown via construction in the following manner. The first two equations of (36) imply that for each $j = 1, \ldots, d$, and fixed t,

$$w_0(t) + w_j(t) = x_0(t) + x_j(t) + y_j(t) - y_j(a(t)) + z(a(t)) - z(t).$$

Set

$$j_t \equiv \arg\max_{1 \leq j \leq d} \{y_j(t) - y_j(a(t))\}; \tag{37}$$

then

$$\begin{aligned}
w_0(t) + w_{j_t}(t) &= x_0(t) + x_{j_t}(t) + y_{j_t}(t) - y_{j_t}(a(t)) - \sum_{k=1}^{d} c_k[y_k(t) - y_k(a(t))] \\
&\geq x_0(t) + x_{j_t}(t) + y_{j_t}(t) - y_{j_t}(a(t)) - \sum_{k=1}^{d} c_k[y_{j_t}(t) - y_{j_t}(a(t))] \\
&= x_0(t) + x_{j_t}(t) + \left(1 - \sum_{k=1}^{d} c_k\right)[y_{j_t}(t) - y_{j_t}(a(t))] \\
&\geq x_0(t) + x_{j_t}(t) \\
&\geq 1,
\end{aligned} \tag{38}$$

where the second to last inequality is due to the monotonicity of y_j and the condition $\sum c_k < 1$, and the last inequality is because $x \in \mathbf{D}_1^{d+1}$. To summarize, for each fixed t, there is an index $1 \leq j_t \leq d$ such that $w_0(t) + w_{j_t}(t) \geq 1$, and consequently, one can find $0 \leq j \leq d$ such that $w_j(t) \geq 1$. We start with $w_{j_0}(0) \geq 1/2$ for some $0 \leq j \leq d$. We set $y_{j_0}(t) = 0$ and solve a problem involving a d-dimensional mapping associated with the remaining components $j \neq j_0$.

If we can guarantee the existence and uniqueness of such a solution, then we have constructed the unique solution of (36) until the first time $t_1 > 0$ such that $w_{j_0}(t_1) = 0$. If $t_1 = \infty$ then our procedure is finished. Otherwise, we are guaranteed by (38) that there is some component j_1 such that $w_{j_1}(t_1) \geq 1/2$. We now freeze the process y_{j_1} from t_1 onwards, and solve the corresponding d-dimensional problem. Iterating in this way we can thus construct (w, z), piece by piece, over any time interval $[0, T]$, and concatenate the pieces in the obvious way.

The details necessary to make such a procedure rigorous are similar to those found in, for example, Chen and Mandelbaum (1991) and Nguyen (1994). Therefore, rather than presenting a complete proof, we will mention only the main ideas that are needed. First, one needs to show that the procedure described above indeed gives the *unique solution* to (36); this will be done in Proposition 1. Second, the pasting procedure involves constructing solutions for certain d-dimensional mappings. If the mapping takes the form of (36) (this happens when the "frozen" component does not correspond to w_0), then the existence and uniqueness of such a mapping can be guaranteed by induction. However, if the "frozen" component is y_0, then the mapping takes a different form and Proposition 2 is needed. The proof for continuity of the mapping under the stated conditions is considerably involved and uses essentially the same methods and ideas as in Nguyen (1994), so we will omit it here.

Proposition 1 *If there exists a pair of processes (w, z) satisfying the system of equations (36), then the solution is unique.*

Proof. The proof proceeds similarly to that of Proposition 2.4 of Chen and Mandelbaum (1991). Let (w, z) be the process constructed as described above, and let (w', z') be another pair satisfying the conditions of this proposition. First, suppose we can show that (a) z and z' coincide on $[0, \delta]$ for some $\delta > 0$; (b) if $z(\tau) = z'(\tau)$ at some $\tau \geq 0$, then the two also coincide on $[\tau, \tau + \delta]$ for $\delta > 0$; (c) if $z(t) = z'(t)$ on $t \in [0, \tau)$ then $z(\tau) = z'(\tau)$. Then defining $\tau \equiv \sup\{t \geq 0 : z(s) = z'(s) \text{ for all } 0 \leq s \leq t\}$, it follows from (a) that $\tau \geq \delta$. If $\tau < \infty$, then (c) holds, hence z and z' coincide beyond τ. But this is in contradiction with the definition of τ, so we must conclude that $\tau = \infty$. It remains only to establish conditions (a) – (c).

To prove (a) let us consider the first interval $[0, s_1)$ associated with the function a. If on

this interval $a(t)$ is a constant then it must be 0, and (36) becomes

$$\left[\begin{array}{l} w_0(t) = x_0(t) + y_0(t) - z(t) \\ w_j(t) = x_j(t) - y_0(t) + y_j(t), \quad 1 \le j \le d \\ z(t) = \sum_{k=1}^{d} c_k y_k(t) \\ y_j \text{ is nondecreasing with } y_j(0) = 0, \quad 0 \le j \le d \\ y_j \text{ increases only at times } t \text{ where } w_j(t) = 0, \quad 0 \le j \le d. \end{array}\right]$$

The existence and uniqueness of (w, y), hence (w, z), is guaranteed by Harrison and Reiman (1981). If, on the other hand, $a(t) = t$ on the interval $[0, s_1)$, then we have

$$\left[\begin{array}{l} w_0(t) = x_0(t) + y_0(t) - z(t) \\ w_j(t) = x_j(t) - y_0(t) + z(t), \quad 1 \le j \le d \\ z(t) = \sum_{k=1}^{d} c_k y_k(t) \\ y_j \text{ is nondecreasing with } y_j(0) = 0, \quad 0 \le j \le d \\ y_j \text{ increases only at times } t \text{ where } w_j(t) = 0, \quad 0 \le j \le d. \end{array}\right]$$

Here, the existence and uniqueness of (w, z) is a direct extension of Theorem 2.5 of Chen and Mandelbaum (1991). In either case, we have shown that the solution is unique over an interval $[0, \delta]$ (say $\delta = s_1/2$).

The proof of (b) follows from the arguments for part (a) by shifting the initial time to τ. For part (c), observe that

$$\begin{aligned} w_0(t) &= [w_0(t^-) + x_0(t) - x_0(t^-)] + [y_0(t) - y_0(t^-)] - \sum_{k=1}^{d} c_k [y_k(t) - y_k(t^-)] \\ w_j(t) &= [w_j(t^-) + x_j(t) - x_j(t^-)] - [y_0(t) - y_0(t^-)] + (1 - c_j)[y_j(t) - y_j(a(t))] - \\ &\quad (1 - c_j)[y_j(t^-) - y_j(a(t^-))] + \sum_{k=1}^{d} c_k [y_k(a(t)) - y_k(a(t^-))]. \end{aligned}$$

If $a(t) = t$, the above system of equations reduces to

$$\begin{aligned} w_0(t) &= [w_0(t^-) + x_0(t) - x_0(t^-)] + [y_0(t) - y_0(t^-)] - \sum_{k=1}^{d} c_k [y_k(t) - y_k(t^-)] \\ w_j(t) &= \{[w_j(t^-) + x_j(t) - x_j(t^-)] - (1 - c_j)[y_j(t^-) - y_j(a(t^-))] + \\ &\quad \sum_{k=1}^{d} c_k [y_k(t^-) - y_k(a(t^-))]\} - [y_0(t) - y_0(t^-)] + \sum_{k=1}^{d} c_k [y_k(t) - y_k(t^-)]; \end{aligned}$$

if otherwise $a(t) < t$, we have

$$w_0(t) = [w_0(t^-) + x_0(t) - x_0(t^-)] + [y_0(t) - y_0(t^-)] - \sum_{k=1}^{d} c_k [y_k(t) - y_k(t^-)]$$

$$w_j(t) = \{[w_j(t^-) + x_j(t) - x_j(t^-)] - (1 - c_j)[y_j(a(t)) - y_j(a(t^-))] + \sum_{k=1}^{d} c_k[y_k(a(t)) - y_k(a(t^-))]\} - [y_0(t) - y_0(t^-)] + (1 - c_j)[y_j(t) - y_j(t^-)].$$

In both cases we consider the first bracketed term to be a known quantity and require that

$$[y_j(t) - y_j(t^-)]w_j(t) = 0, \quad \text{for all } j = 1, \ldots, d.$$

What we have in essence is a linear complementarity problem (LCP) that involves finding $y(t) - y(t^-)$ to satisfy $w = q + M[y(t) - y(t^-)] \geq 0$, $y(t) - y(t^-) \geq 0$, and $w'[y(t) - y(t^-)] = 0$. In both of the cases above, the matrix M can be shown to be positive definite, hence a P-matrix. From Theorem 1.14 of Harker (1993), we conclude that the LCP has a unique solution, hence y has a unique extension from t^- to t. □

Proposition 2 *Let $0 < c_j < 1$, $j = 1, \ldots, d$. For each $x = (x_1, \ldots, x_d) \in \mathbf{D}^d$, $a \in \Lambda$, there exists a unique pair of processes (w, z) satisfying, for $j = 1, \ldots, d$,*

$$\left[\begin{array}{l} w_j(t) = x_j(t) + (1 - c_j)[y_j(t) - y_j(a(t))] + z(a(t)) \geq 0 \\ z(t) = \sum_{k=1}^{d} c_k y_k(t) \\ y_j \text{ is nondecreasing with } y_j(0) = 0 \\ y_j \text{ increases only at times } t \text{ with } w_j(t) = 0. \end{array}\right]$$

Moreover, z is a continuous process if $x - (y \circ a)$ has no jumps downwards.

Proof. The proof of existence is by construction. To begin, consider $[s_0, s_1)$, $s_0 \equiv 0$, $s_1 > 0$, the first interval associated with the function a. On this interval, either $a(t)$ takes the value of a constant, in which case it must be 0, or $a(t) = t$. In the first case where $a(t) = 0$, then $w_j(t) = x_j(t) + y_j(t)$, which is the one-dimensional regulator in Proposition 2.2.3 of Harrison (1985). If in the latter case $a(t) = t$, then $w_j(t) = x_j(t) + z(t)$, and the same result from Harrison (1985) can be used. In either case, (w, z) is uniquely defined on the interval $[s_0, s_1)$.

One can use the same argument as in the proof of Proposition 1 to show that (w, z) can be extended uniquely to s_1. Shifting the initial time to s_1, one can similarly construct (w, z) on the interval $[s_1, s_2)$, and so on. "Pasting" these pieces together, we thus construct (w, z) on any time interval $[0, T]$, and uniqueness of this procedure is shown similarly to Proposition 1.

It remains to show that y is continuous under the stated condition, which is equivalent to show that $y(t) - y(t^-) = 0$ for all $t > 0$. But $y(t) - y(t^-)$ is the unique solution to an LCP

$w = q + M[y(t) - y(t^-)] \geq 0$, $w'[y(t) - y(t^-)] = 0$, either

$$q = w_j(t^-) + x_j(t) - x_j(t^-) - (1 - c_j)[y_j(t^-) - y_j(a(t^-))] + z(t^-) - z(a(t^-))$$

or

$$q = w_j(t^-) + x_j(t) - x_j(t^-) - (1 - c_j)[y_j(a(t)) - y_j(a(t^-))] + z(a(t)) - z(a(t^-))$$

depending on whether $a(t) = t$ or $a(t) < t$, respectively. But $w_j(t^-) \geq 0$, $x - (y \circ a)$ has no negative jumps, and y, z, and a are nondecreasing, so $q \geq 0$ in either case. Hence $y(t) - y(t^-) = 0$ is the unique solution. □

References

BASKETT, F., CHANDY, K. M., MUNTZ, R. R., AND PALACIOS, F. G. 1975. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM* **22**, 248–260.

BILLINGSLEY, P. 1968. *Convergence of Probability Measures.* Wiley, New York.

CARR, S. A., GÜLLÜ, A. R., JACKSON, P. R. AND MUCKSTADT, J. 1993. An Exact Analysis of the No B/C Stock Policy, preprint.

CHEN, H. AND MANDELBAUM, A. 1991. Leontief systems, RBV's and RBM's. *Proceedings of the Imperial College Workshop on Aplied Stochastic Processes,* M. H. A. Davis and R. J. Elliott (eds.), Gordon and Breach Science Publishers (forthcoming).

DAI, J. G. AND HARRISON J. M. 1992. The QNET method for two-moment analysis of closed manufacturing systems. Submitted for publication.

HARKER, P. T. 1993. Lectures on Computation of Equilibria with Equation-Based Methods. *The Core Lecture Series,* Core Foundation, Louvain-La-Neuve, Universite Catholique de Louvain.

HARRISON, J. M. 1985. *Brownian motion and stochastic flow systems.* Wiley, New York.

HARRISON, J. M. 1988. Brownian models of queueing networks with heterogeneous customer populations. *Proceedings of the IMA Workshop on Stochastic Differential Systems,* W. Fleming and P. L. Lions (eds.), IMA Volume **10**, Springer-Verlag.

HARRISON, J. M. AND NGUYEN, V. 1990. The QNET method for two-moment analysis of open queueing networks. *Queueing Systems: Theory and Applications* **6**, 1–32.

HARRISON, J. M. AND NGUYEN, V. 1993. Brownian models of multiclass queueing networks: Current status and open problems. *Queueing Systems: Theory and Applications* **13**, 5–40.

HARRISON, J. M. AND REIMAN, M. I. 1981. Reflected Brownian motion on an orthant. *Annals of Probability* **9**, 302–308.

IGLEHART, D. L. AND WHITT W. 1970. Multiple channel queues in heavy traffic I. *Advances in Applied Probability* **2**, 150–177.

KELLY, F. P. 1979. *Reversibility and Stochastic Networks.* Wiley, New York.

NGUYEN, V. 1994. Fluid and Diffusion Approximations of A Two-Station Mixed Queueing Network. *Mathematics of Operations Research,*.