Sadhana, Vol. 19, Part 6, December 1994, pp 891-939. © Printed in India

Dynamic scheduling in manufacturing systems using Brownian approximations

K RAVIKUMAR and Y NARAHARI

Department of Computer Science and Automation
Indian Institute of Science, Bangalore

Abstract. Recently, Brownian networks have emerged as an effective stochastic model to approximate multiclass queueing networks with dynamic scheduling capability, under conditions of balanced heavy loading. This paper is a tutorial introduction to dynamic scheduling in manufacturing systems using Brownian networks.. The article starts with motivational examples. It then provides a review of relevant weak convergence concepts, followed by a description of the limiting behaviour of queueing systems under heavy traffic. The Brownian approximation procedure is discussed in detail and generic case studies are provided to illustrate the procedure and demonstrate its effectiveness. This paper places emphasis only on the results and aspires to provide the reader with an up-to-date understanding of dynamic scheduling based on Brownian approximations.

Keywords. Brownian networks; dynamic scheduling; manufacturing systems; multiclass queueing networks; heavy traffic approximations; weak convergence; functional cental limit theorem.

1. Introduction

Scheduling as a research area is motivated by important resource allocation questions that arise in manufacturing systems, computer systems, computer communication networks, and in general, in all situations where scarce resources have to be allocated to activities over time to appropriate servers (processors, machines, communication channels, material handling devices, etc.) so as to optimize a performance criterion, while satisfying a set of given constraints. Scheduling problems can be classified as *static* scheduling problems when the jobs to be scheduled comprise a fixed set and *dynamic* when jobs can arrive into the facility in an ongoing and usually, in a random fashion. Another usual way of classifying scheduling problems is to consider them as *deterministic* or *stochastic*. In *deterministic* scheduling, job characterstics such as processing times, due dates, and release dates are known with certainty to the scheduler before the actual processing occurs. In *stochastic* scheduling, the scheduler cannot observe the processing times in advance, but only has knowledge

of a probability distribution for the various processing times. In this paper, the emphasis is on *dynamic* and *stochastic scheduling* of multi-class queueing network models of discrete event systems, using a class of *heavy traffic* approximations, called **Brownian approximations**. Also all our motivating scheduling problems come from the area of manufacturing systems, though the methodology that we discuss is applicable, in general, to any discrete activity scheduling problem with dynamic and stochastic characterstics.

1.1 *Deterministic scheduling*

Much of the research in the area of scheduling has focussed on deterministic scheduling problems. Most of the scheduling problems in this area have been shown to be NP- hard and researchers have explored several different approaches to confront NP-hardness.

- Determine a *lower bound* on the cost of the schedule and then use a *branch and bound* method to determine the optimal solution (Bagchi & Ahmadi 1987; Beloudah *et al* 1988). However, this technique needs exponential amount of computation time in the worst case.
- Use *dynamic programming* (Abdul-Razaq & Potts 1988; Baker & Ahmadi 1978). This technique works very well for many scheduling problems, but like branch and bound technique, needs exponential amount of computation time in the worst case.
- Obtain *sub-optimal* solutions in polynomial time (Hochbaum & Shmoys 1988). Such approximation algorithms are, however, applicable only in specific problem instances and do not yield general methods.
- Use simple *heuristics* (Gere 1987) such as EDD (Ealiest Due Date), SPT (Shortest Processing Time), etc. Heuristics are very efficient and have the ability to react to dynamic changes and have widespread applicability. In general, however, heuristics do not offer the guarantee that the solution is within an acceptable margin of error when compared with the optimal solution.
- *Lagrangian relaxation* based methods (Fisher 1973, 1981) which yield efficient near-optimal solutions with measurable performance as well as important job interaction information to accommodate dynamic changes and to handle new jobs.
- More recently, randomized local search algorithms such as *simulated annealing* (Van Laarhoven *et al* 1992) and *genetic algorithms* (Goldberg 1986) have been applied to deterministic scheduling problems. Another paradigm that has also been used in this context is *neural networks* (Levy & Adams 1987).

1.2 *Stochastic scheduling*

In the area of stochastic scheduling, the results are scattered and technically complicated (Lawler *et al* 1990); they rely on semi-Markovian decision theory and stochastic dynamic optimization. Important results in this dynamic optimization are sur-

veyed by Lawler *et al* (1990), Weiss (1982), Pinedo and coworkers (Pinedo & Weiss 1980, 1987; Pinedo 1981-1983, Pinedo & Scrage 1982) and Forst (1984).

Single class and multiclass queueing networks constitute an important class of stochasic models of discrete event systems (see Walrand 1988). The optimal scheduling of such networks has been attempted by several researchers, but only with limited success. Some of the notable efforts in this area include:

- Priority sequencing in single station queueing systems (Klimov 1974)
- Optimal dynamic scheduling in Jackson networks (Ross & Yao 1989)
- Optimal scheduling control in a flexible machine (Yao & Shantikumar 1990)
- Optimal control of interacting service stations (Hajek 1974)
- Optimal control of service rates in networks of queues (Weber & Stidham 1987)
- Optimal control of admission to a queueing system (Stidham 1985)

However, according to Harrison & Wein (1989), a satisfactory theory for sequencing and scheduling in a queueing network setting has yet to be formulated. Discrete event simulation continues to be the primary tool of analysis and the best hope for further progress appears to be in the analysis of cruder and more tractable models.

Recently, *Brownian networks* (Harrison 1988) have emerged as an effective stochastic model to approximate multiclass queueing networks with dynamic scheduling capability, under conditions of *balanced heavy loading* (see §3). A Brownian network is a crude model but highly tractable and successful in the context of dynamic and stochastic scheduling of queueing networks. This paper attempts to survey the important results in this area. In particular, we present:

- foundational aspects of Brownian networks as applied to the modeling of multiclass queueing networks,
- methodological details of how sequencing and scheduling problems can be approached via the Brownian approximation,
- several illustrative case studies to gain insight into specific methodological details.

Since this paper is intended as a tutorial review, we have used extensively the results from many important papers in this area. These papers include: Harrison (1988), Harrison & Wein (1990), and Wein (1990, 1992). Wherever highly relevant, we shall again explicitly provide a reference to these papers.

1.3 *Motivational examples*

In this section we describe some scheduling problems that occur in dynamic and stochastic environments through some simple and illustrative manufacturing system examples.

1.3.1 *A multiclass make-to-stock queue*

A make-to stock production facility produces a wide variety of products according to a forecast of customer demands and completed jobs enter a finished good inventory, which in turn services the actual customer demand. Figure 1 shows a single machine make-to-stock system with K classes of products.

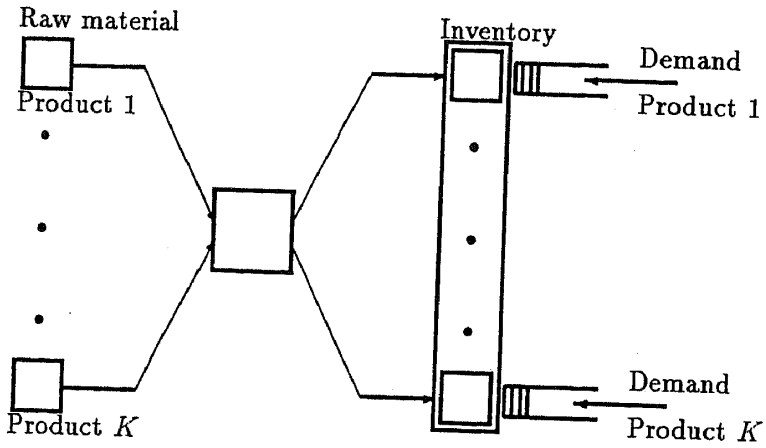

Figure 1. A single machine make-to-stock system.

We shall consider a system with $K = 5$. The class designations summarize all relevant information with regard to processing times and demand patterns of respective products as given below:

- *Class 1*: Product's processing time is low but demands arrive frequently for it.
- *Class 2*: Products have long processing times but demands arrive occasionally for them.
- *Class 3*: These are high priority products with medium processing times and nominal demands. Waiting times of customers arriving for these products shold be low.
- *Class 4*: These are products with medium processing times but very high demands.
- *Class 5*: For these products processing times are medium and demands are occasional.

An arriving class k customer takes a product of the same class, if available; otherwise, backorders for one. Linear costs per unit time are incurred for holding inventory and for backordering.

In a realistic scenario, actual processing times are not known with certainty and one can only have a knowledge of probability distributions of various processing times. It is common that every such manufacturing system experiences some amount

of variability in estimated processing times. Also, sometimes these variations turn out to be unprediclable. For example, variations due to rework, machine failure *etc.* fall under this category. These variations will have impact on costs incurred by the system. For instance, long processing times may reduce inventory holding costs but at the same time they incur high backordering costs. Also, variations in interarrival times of demands produce similar effects. Hence, given the stochastic nature of the problem, deterministic scheduling is less realistic than dynamic scheduling. Further, scheduling policies which perform well in the determinstic setting may not perform well in stochastic setting.

A typical dynamic scheduling decision for the foregoing problem consists of choosing among the following options at each point in time:

- *either* work on a class k job, $k = 1,\ldots,5$
- *or* allow the machine to idle.

Using Brownian analysis methodology, Wein (1992) derived a dynamic scheduling policy for a make-to-stock system under general service distributions and renewal demand patterns. The decision as to whether a machine is to be kept busy or idle at any time point is dictated by the weighted inventory level process (which is a weighted sum of inventory levels of each class, the weights being mean processing times.)at that time. The priority decision derived is reminiscient of the well known $c\mu$ - rule, which awards priority to the class with the largest value of the index $c_k\mu_k$ where c_k is the holding cost and μ_k is the service rate. This policy is discussed in detail in §4.1. Simulation results showing comparison of this policy with other conventional policies are also presented.

1.3.2 *A re-entrant line*

Re-entrant lines are queueing network models for wafer fabrication in a semi-conductor manufacturing system. Wafer fabrication involves a large and complex sequence of processing steps. A charecterstic feature of wafer fabrication is *re-entrancy*, that is each wafer visits the same machine centre multiple number of times. Figure 2 depicts a three-station re-entrant line with a single job type.

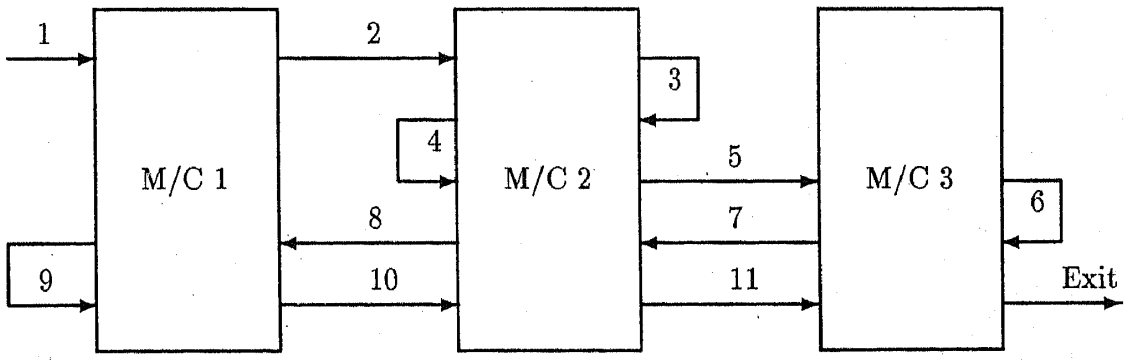

Figure 2. A three station re-entrant line.

Each job has its own deterministic route through the network. We can define a different customer class for each operation of each job. For example, in figure 2, a job has to undergo 11 stages of operation and hence, has eleven classes associated with it. Each job class has its own processing time distribution and different classes contend for service at the same machine center. If the population size of jobs circulating in the system is held constant,*i.e.,* a new job is admitted whenever a job leaves the network, then the above system can be modeled as a three station multiclass closed queueing network. A scheduling problem of relevance in this context is to choose a policy which, at each point in time indicates which class to be serviced at each station. Some of the conventional policies employed in scheduling a re-entrant line are FCFS, FBFS (First Buffer First Served), LBFS (Last Buffer First Served), SEPT (Shortest Expected Processing Time) *etc.*

In re-entrant lines, processing times of various operations are susceptible to unpredictable variability mainly because of complexity and precision requirements involved in performing the operations. Furthermore, due to multiple visits of jobs, many machine centres will be heavily loaded and thus become bottlenecks. In the example shown in figure 2, stations 1 and two are bottleneck stations. These bottlenecks are precisely where large queues form, where most of the waiting time is incurred and where scheduling will have biggest impact. Hence, under such scenario, utilizations of the bottleneck machines should be as high as possible to reduce cycle times and this can be affected through scheduling decisions which take into account the state of the system at any point in time.

A similar scheduling problem for two station closed queueing network is considered in Harrison & Wein (1990) and the scheduling decision considered there is referred to as *workload balancing rule.* It is a static priority sequencing policy, which assigns priority at any station according to an index rule which aims at minimizing workload imbalance between the two stations and there by enhancing the utilization levels of the machines.

This policy works well when the network has more than one bottleneck stations and not too many non-bottleneck stations. If the network has only one bottleneck it is difficult to affect utilizations of bottleneck machines because there are no other bottleneck stations to feed it. Similarly, presence of too many non-bottleneck stations prevent bottleneck stations to feed one another in an effective manner. Details of this policy are given in §4.3. A simulation study is conducted using the workload balancing policy, on the above re-entrant line examples. Section 4.3 also contains results of these experiments.

In the example of figure 2, one can easily see that by admitting a new job into the network, the number of customers of each class goes up by one. Thus, if an open loop release policy which pushes jobs into the system without observing the state of the system is followed then the WIP (Work-In-Process) inventory levels shoot up drastically. From Little's law, it is known that for a given mean arrival rate, mean cycle times are directly proportional to mean WIP. However, the relationship between mean throughput rate and mean WIP is highly non-linear and dependent on the scheduling policy. Using an effective job release policy and a priority sequencing policy combination one can achieve high throughput rates while maintaining low levels of WIP.

An interesting job release policy, known as workload regulating policy, is consid-

ered in Wein (1990b). This policy injects a job into the system whenever the amount of work in the system for the bottleneck stations satisfies certain conditions. The priority sequencing policy uses dynamic reduced costs from a linear program. These policies along with some results of simulation experiment performed on the above reentrant line are presented in §4.4.

The remainder of this paper is organized as follows. In §2.1 we describe weak convergence concepts of relevance and discuss in §2.2, how the heavy traffic limit theorems are proved invoking these concepts. The description of the Brownian network, followed by the approximation procedure, is given in §3.1. Section 3.2 provides workload formulation for the Brownian network of §3.1, which describes the system dynamics of the queueing network in terms of workloads at service centres. Modifications needed to adopt the approximation procedure to the case of closed queueing networks are presented in §3.3. Section 4 illustrates the procedure in the context of various manufacturing systems of practical importance. Section 4.1 deals with a scheduling problem in a single machine make-to-stock queue. Section 4.2 discusses the case of a two-station closed queueing network, with an objective to maximize the throuhput. Section 4.3 gives an interesting scheduling problem in a two-station network with controllable inputs. Here we mention that no attempt is made to provide rigorous proofs for the theorems presented and interested readers are referred to appropriate contributions for a detailed study of the problem concerned.

2. Foundations

2.1 *Weak convergence concepts*

In this section we describe some relevant notions of weak convergence which will be used in subsequent portions of this paper.

Let $\{X(t), t \in T\}$ be a stochastic process on a probability space $(\Omega,\mathcal{E},\mathcal{P})$. Suppose that the index set T is an interval of the real line R. For a fixed $\omega \in \Omega$, the function $X(\omega,.)$ of t gives a sample path of the process. If all such sample paths lie in some fixed collection $\mathcal{X}$ of real valued functions on T, then the process X can be thought of as a map from Ω into $\mathcal{X}$, a random element of $\mathcal{X}$. For example, if a process indexed by [0,1] has continuous sample paths it will be a random element of the space $C[0,1]$ of all real-valued continuous functions on [0,1]. However,the notion of random element needs to be formalized adding measurability requirement as follows:

DEFINITION 2.1.1 An $\mathcal{E}/\mathcal{A}$-measurable map X from a probability space $(\Omega,\mathcal{E},\mathcal{P})$ into a set $\mathcal{X}$ with a σ-field $\mathcal{A}$ is called **random element** of $\mathcal{X}$.

If $\mathcal{X}$ is a metric space, the set of all bounded, continuous $\mathcal{A}/\mathcal{B}(\mathrm{R})$ measurable, real-valued functions on $\mathcal{X}$ is denoted by $\mathcal{C}(\mathcal{X},\mathcal{A})$. Note that if $\mathcal{A}$ is the Borel field generated by closed sets of $\mathcal{X}$, then every continuous function on $\mathcal{X}$ is measurable. A sequence $\{X_n\}$ of random elements of $\mathcal{X}$ converges in distribution to a random element X, written as $X_n \Rightarrow X$, if

$$\int f(X_n)\, d\mathcal{P} \longrightarrow \int f(X)\, d\mathcal{P} \text{ for each } f \in \mathcal{C}(\mathcal{X},\mathcal{A}) \tag{1}$$

A sequence $\{P_n\}$ of probability measures on $\mathcal{A}$ converge weakly to P, written as $P_n \Rightarrow P$ if

$$\int f\, dP_n \longrightarrow \int f\, dP \text{ for each } f \in \mathcal{C}(\mathcal{X},\mathcal{A}) \tag{2}$$

As every random element X of $\mathcal{X}$ induces a probability measure P on $(\mathcal{X}, \mathcal{A})$ defined by

$$P(A) = \mathcal{P}(X^{-1}(A)), \text{ for all } A \in \mathcal{A} \tag{3}$$

convergence in distribution of a sequence of random elements is synonymous to weak-convergence of the corresponding sequence of induced probability measures. However, note that in the latter case X_n and X need not be defined on a same probability space but must induce probability measures P_n and P on the same metric space $(\mathcal{X}, \mathcal{A})$.

Now on, unless otherwise stated, assume that $\mathcal{X}$ is a separable metric space with metric ρ and the Borel σ-field $\mathcal{A}$. If X and Y are defined on a common domain, then $\rho(X, Y)$ is a random variable (see Billingsley 1968). Thus the following definition makes sense.

DEFINITION 2.1.2 A sequence of random elements $\{X_n, n \geq 1\}$ converges in probability to X, written as $X_n \xrightarrow{\mathcal{P}} X$, if X_n and X are defined on a common probability space $(\Omega, \mathcal{E}, \mathcal{P})$ and

$$\rho(X_n, X) \xrightarrow{\mathcal{P}} 0.$$

Here $\xrightarrow{\mathcal{P}}$ denotes convergence in probability of random variables.

Now we state a useful theorem whose application is found frequently in weak convergence results for queueing theory.

Theorem 2.1.1 *Assume that $\{X_n\}$ and $\{Y_n\}$ are sequences of random elements of $\mathcal{X}$ and are defined on a common probability space $(\Omega, \mathcal{E}, \mathcal{P})$. If $X_n \Rightarrow X$ and $\rho(X_n, Y_n) \xrightarrow{\mathcal{P}} 0$, then $Y_n \Rightarrow X$.*

Stochastic processes of interest in queueing theory such as queue length process can often be represented as functions of more basic stochastic processes such as random walks and renewal processes. Consequently limit theorems for stochastic proceses in queueing theory are often obtained from existing limit theorems for these basic processes by showing that the connecting functions preserve convergence. The functions that appear in such proofs are composition, addition, multiplication, supremum, etc. Hence, a natural question that arises in such contexts is: *If $X_n \Rightarrow X$ and f is a measurable mapping from $(\mathcal{X}, \mathcal{A})$ to another separable metric space $(\mathcal{X}', \mathcal{A}')$, does it follow that $f(X_n) \Rightarrow f(X)$?* Observe that the result is trivially true if f is continuous. Interestingly this holds even under slightly weaker assumption as shown by the following theorem:

Theorem 2.1.2 *(*Continuous mapping theorem*) If $X_n \Rightarrow X$ and f is continuous almost surely with respect to the distribution of X, then $f(X_n) \Rightarrow f(X)$.*

The above theorem can be further generalized as given below.

Theorem 2.1.3 *Let $f_n, n \geq 1$ and f be Borel measurable functions mapping the separable metric space $(\mathcal{X}, \mathcal{A})$ into another separable metric space$(\mathcal{X}', \mathcal{A}')$. If $X_n \Rightarrow X$ and $f_n(x_n) \to f(x)$ for all $x \in A$ and $\{x_n \to x\}$, then $f_n(X_n) \Rightarrow f(X)$.*

Most of the weak convergence results in queueing theory rest on the continuous mapping theorem. For an elegant proof of this theorem see Pollard (1984) or Whitt (1980). In queueing theory, the metric spaces of particular interest are $C[0,1]$, the space of all real-valued continuous functions on [0,1] and the space $D[0,1]$ of all real-valued functions that are right continuous at each point of [0,1) with left limits existing at each point of (0,1]. The functions of $D[0,1]$ are called *cadlag* functions.

Thus, the space $D[0,1]$ contains the sample paths of all queue-related processes. Obviously, $C[0,1] \subset D[0,1]$.

The metric on C [0,1] is the uniform metric defined by

$$\rho(x,y) = \sup_{0 \le t \le 1} \mid x(t) - y(t) \mid \text{ for all } x, y \in C[0,1].$$

Under this metric ρ, $C[0,1]$ is complete and separable. But under the same uniform metric $D[0,1]$ is complete but not separable and hence the uniform metric poses some minor measurability difficulties. For instance, under this metric the Borel σ-field turns out to be too large and many interesting stochastic processes fail to be random elements of $D[0,1]$. However, if we consider a strictly smaller σ-field $\mathcal{B}$ generated by closed balls, an interesting weak convergence theory results. $\mathcal{B}$ also coincides with the σ-field generated by co-ordinate projection maps. All interesting functionals on $D[0,1]$ are $\mathcal{B}$ measurable. The lack of a countable dense subset of functions in $D[0,1]$ is surmounted when the limit distributions concentrate on a separable subset of $D[0,1]$ such as $C[0,1]$. For an interesting theory under the uniform metric, see Pollard (1984).

If in the space $D[0,1]$, we define $f_u(t) = I\{t \ge u\}$; $t \in [0,1]$ and $u \in [0,1]$, the collection $\{f_u(.)\}$ is uncountable and under uniform metric two distinct functions are at unit distance. Clearly we want $f_u \to f_v$ whenever $u \to v$. However, this is quite impossible under any topology that leads to a convergence concept which implies pointwise convergence. Skorohod surmounted this difficulty by defining a metric, which leads to pointwise convergence after a suitable rescaling of the time axis that becomes asymptotically negligible. It is difficult to define this metric. It suffices for our purpose to know that such a metric exists and under this metric $D[0,1]$ is separable. Unfortunately, under this metric the space is not complete. Billingsley (1968) defines an equivalent metric under which $D[0,1]$ is both separable and complete. Here onwards, we concentrate on $D[0,1]$ (hereafter to be denoted as D) equipped with the metric under which it is both separable and complete.

Most of the diffusion approximations and heavy traffic limit theorems rely on the so called *Functional central limit theorem*. The functional central limit theorem is an extenstion of classical Lindberg-Levy's central limit theorem (see Breiman 1968) for a sequence of random variables, to the function space D. It is directed at showing that a normalized sequence of random functions converges to a diffusion process. The advantage of functional limit theorems lies in the fact that weak convergence results can be immediately obtained for various functionals of the processes. The standard method to prove weak convergence of a normalized sequence of processes is to first show the convergence of finite dimensional distributions. However, this is not sufficient. A certain *tightness* property of the induced measures needs to be demonstrated and this part poses some technical difficulties.

To get a feel for weak convergence in function spaces, in what follows we state various functional limit theorems of relevance in queueing theory.

Theorem 2.1.4 *(***Donsker's theorem***) Let* $\{\xi_i, i \ge 1\}$ *be a sequence of i.i.d. random variables with mean 0 and variance* $\sigma^2 < \infty$, *defined on* $(\Omega, \mathcal{E}, \mathcal{P})$. *Let*

$$S_k = \xi_1 + \cdots + \xi_k \text{ for all } k \ge 1 \text{ and } S_0 = 0$$

From the $\{S_n\}$, form random elements $\{X_n\}$ of D as:

$$X_n(t) = \frac{S_{[nt]}}{\sigma\sqrt{n}} \text{ for all } t \in [0,1]$$

where $[x]$ is the greatest integer less than or equal to x. Let $W = \{W_t,\ 0 \leq t \leq 1\}$ be the standard Brownian motion. Then

$$X_n \Rightarrow W \text{ in } D$$

For a proof of this see Billingsley (1968).

A generalization of the Donsker's theorem is due to Prohorov and is given by:

Theorem 2.1.5 *(**Prohorov's theorem**) Suppose for each $n \geq 1$ there exists a sequence of i.i.d. random variables $\{\eta_i^n,\ i \geq 1\}$ with mean 0 and variance σ^2. Define partial sums*

$$S_k^n \equiv \eta_1^n + \cdots + \eta_k^n \text{and } S_0^n \equiv 0.$$

Assume that

$$\sigma_n^2 \to \sigma^2 \text{ as } n \to \infty,\ 0 < \sigma^2 < \infty$$

and

$$\sup_{n \geq 1} E\{|\ \eta_1^n\ |^{2+\epsilon}\} < \infty \text{ for some } \epsilon > 0$$

Let

$$X_n(t) \equiv \frac{S_{[nt]}^n}{\sigma\sqrt{n}}.$$

Then,

$$X_n \Rightarrow W \text{ in } D.$$

In the above theorems the partial sums S_n are defined for fixed indices n. Sometimes we encounter cases where the index is random and such a phenomenon is common in renewal processes. Suppose ν_n is a random integer such that ν_n is large with high probability. Define random elements X_n of D as given in Donsker's theorem. Further, define another random element Y_n of D by,

$$Y_n(t) = \frac{S_{\nu_{[nt]}(\omega)}(\omega)}{\sigma\sqrt{n}}$$

Now we seek conditions under which $\{Y_n\}$ weakly converges to some limit. Observe that $Y_n(\omega)$ results from $X_n(\omega)$ by subjecting X_n to a random time scale. If we define $\phi_n(t,\omega)$ by

$$\phi_n(t,\omega) = \nu_{[nt]}(\omega)/n$$

then it follows that

$$Y_n(t,\omega) = X_n(\phi_n(t,\omega),\omega).$$

Thus Y_n is X_n with the time scale subjected to a change represented by random function ϕ_n. Such cases can be dealt with using the following theorem, known as the *Random time change theorem.*

Let D_0 denote the set of elements ϕ of D that are non-decreasing and satisfy $0 \leq \phi(t) \leq 1$ for all $t \in [0,1]$. For $X \in D$, $\phi \in D_0$, let $(X \circ \phi)(t) = X(\phi(t))$. Suppose that in addition we have random elements X_n and ϕ_n of D and D_0 respectively, where X_n and ϕ_n have the same domain (which can vary with n). Note that $X \circ \phi$ and $X_n \circ \phi_n$ for each n lie in D. If D_0 is topologized by relativizing the Skorohod topology of D, then it is easy to see that (X, ϕ) and (X_n, ϕ_n) are random elements of $D \times D_0$ with product topology. The following result is given by Billingsley (1968).

Theorem 2.1.6 *(Random time change theorem) If* $(X_n, \phi_n) \Rightarrow (X, \phi)$ *and* $\mathcal{P}(\mathcal{X} \in C) = \mathcal{P}(\phi \in C) = 1$, *then*

$$X_n \circ \phi_n \Rightarrow X \circ \phi$$

where $C \equiv C[0,1]$.

The proof of the above theorem is based on continuous mapping theorem and also on the fact that Skorohod topology relativized to C coincides with the topology generated by uniform metric on C. The theorem is useful in deriving functional central limit theorem for renewal processes.

Theorem 2.1.7 *Let* $\eta_1, \eta_2 \cdots$ *be an i.i.d. sequence of random variables with mean* μ *and variance* $\sigma^2 < \infty$. *Define,*

$$\nu_t = \max\{k : \sum_{i=1}^{k} \eta_i \leq t\}, \text{ with } \nu_t = 0 \text{ if } \eta_1 \geq t$$

Thus ν_t *gives number of renewals upto time* t. *Define*

$$Z_n(t,\omega) = \frac{\nu_{nt}(\omega) - nt/\mu}{\sigma \mu^{-3/2} \sqrt{n}}.$$

Then, $Z_n \Rightarrow W$ *in* D.

For a proof of this refer Billingsley (1968).

These notions of weak convergence are used in proving heavy traffic limit theorems for queue related processes as we shall see in §2.3.

2.2 *Brownian motion*

As discussed in the previous section the limit process in the functional central limit theorem is the standard Brownian motion. In this section we define the standard Brownian motion and the reflected Brownian motion which is a functional of Brownian motion. Many of the interesting queue related processes converge to the latter.

DEFINITION 2.2.1 A standard Brownian motion or Wiener process is a stochastic process $\{X(t), 0 \leq t \leq 1\}$ on $(\Omega, \mathcal{E}, \mathcal{P})$ having continuous sample paths, and stationary independent increments such that for any fixed $t \in [0,1]$, $X(t)$ is normally distributed with mean 0 and variance t.

Thus, a standard Brownian motion starts at level zero almost surely.

DEFINITION 2.2.2 A process $\{Y(t),\ 0 \leq t \leq 1\}$ is called a (μ, σ) Brownian motion if it has the form:

$$Y(t) = Y(0) + \mu t + \sigma X(t). \tag{4}$$

where $X(t)$ is the standard Brownian motion and $Y(0)$ is independent of X.

It follows that $Y(t+s)-Y(t) \sim N(\mu s, \sigma^2 s)$. μ is called the drift and σ^2 the variance of $Y(t)$.

The normality requirement in the above definition is superfluous because if Y is a continuous path process and has independent increments, then Y is a Brownian motion and normality follows as a consequence of these assumptions. Refer Breiman (1968) and Cox & Miller (1965) for further properties of Brownian motion.

DEFINITION 2.2.3 Let $f : D \rightarrow D$ be defined for all $Y \in D$ as $f(Y) = Z$ where

$$Z(t) = Y(t) - \inf_{0 \leq s \leq t} \{Y(s)\}, 0 \leq t \leq 1,$$

where $Y(t)$ is (μ, σ)Brownian motion with $Y(0) = 0$. Then {Z(t),$0 \leq t \leq 1$} is called *reflected Brownian motion*, denoted by RBM(μ, σ).
Whitt (1980) proves that f above is continuous in Skorohod topology.

For processes in R^K, the limit processes of the functional central limit theorem is a K-dimensional Bownian motion, $\{\overline{Y}(t), 0 \leq t \leq 1\}$, specified by a K-dimensional drift vector, $\overline{c}$ and a $K \times K$ covariance matrix A, denoted by BM(c, A), *i.e.*, $\overline{Y}(t)$ is a K-dimensional vector stochastic process with continuous sample paths in R^K with $\overline{Y}(0)$=0 and stationary independent increments.

Similarly, a reflected Brownian motion on the non-negative orthant R_+^K was defined and characterized by Harrison & Reiman (1981) and is discussed in detail by Harrison & Williams (1987). It behaves like Brownian motion on the interior of its state space R_+^K and reflects instantaneously in a fixed direction at each boundary hyperplane. The reflection directions are given in $K \times K$ reflection matrix R, where the k-th row of R gives the reflection direction for the boundary corresponding to $X_k(t) = 0$. This process is thus completely specified by (c, A, R) where c and A correspond to the drift vector and covariance matrix of the underlying Brownian motion.

2.3 *Heavy traffic limit theorems*

A queueing system is stable if the input rate is less than the output rate, *i.e.*, if the traffic intensity, ρ is strictly less than unity. If $\rho \geq 1$, the system is unstable and the queueing processes tend to blow up. For example, in a GI/G/1 queue, if $\rho \geq 1$, then for any $\mathcal{K} < \infty$,

$$\lim_{n \rightarrow \infty} \mathcal{P}\{W_n \geq \mathcal{K}\} = 1,$$

where W_n is the waiting time of the n-th customer. Thus if $\rho \geq 1$ the queue is said to be under heavy traffic. However, even under heavy traffic conditions, properly normalized sequences of queueing processes converge weakly to diffusion processes. Heavy traffic limit theorems formalize this fact. The diffusion approximation procedures stem from these limit theorems.

In this section, we discuss a simple case of GI/G/1 queue under heavy traffic and give an intuitive feel for how the heavy traffic limit theorems are proved invoking the weak convergence concepts discussed in §2.1.

Consider a standard GI/G/1 queue determined by two independent sequences of i.i.d. random variables $\{u_n, n \geq 1\}$ and $\{v_n, n \geq 0\}$. Assume that the 0-th customer arrives at time $t = 0$ to find a free server. Let v_n represent the service time of the n-th customer and u_n represent the inter-arrival time between the $(n-1)$-st customer

and n-th customer. We define,

$$\rho = E(v_1)/E(u_1)$$

$$Y_n = v_{n-1} - u_n$$

$$W_{n+1} = [W_n + Y_{n+1}]^+, \text{ for all } n \geq 0 \quad \text{and} \quad W_0 = 0.$$

W_n gives the waiting time of the n-th customer. It is well known that if $\rho < 1$, then there exists a non-degenerate random variable W such that :

$$W_n \Rightarrow W \text{ as } n \to \infty.$$

Under appropriate moment conditions one can show that

$$(\sigma^2 n)^{-1/2}[\sum_{k=1}^{n} W_k - n\,E(W)] \Rightarrow N(0,1).$$

In this case, the events $\{W_k{=}0\}$ are regenerative points for $\{W_n, n \geq 1\}$ and $\{W_k{=}0\}$ occurs infinitely often, w.p.1. Thus, $\{W_n, n \geq 1\}$ is a regenerative process and $\{\sum_{k=1}^{n} W_k,\ n \geq 1\}$ is a cumulative process. Thus, $\{W_n\}$ can be broken up into i.i.d. blocks and consequently, eventhough $\{W_n\}$ is itself not i.i.d., the theory of sequence of i.i.d. random variables can be applied for a proof of the above convergence.

But, in the case when $\rho = 1$, the situation is more delicate and in the context of Markov chains this case corresponds to null recurrence. If $\rho = 1$, $W_n \leq \mathcal{K}$ for $\mathcal{K}$ finite, infinitely often w.p.1. But the expected time between epochs when customers arrive to find a free server is infinite. However, observe that,

$$W_n = S_n - \min\{S_k, 0 \leq k \leq n\},\ n \geq 0 \tag{5}$$

where $S_n = \sum_{k=1}^{n} Y_k$ and $S_0 = 0$.

It is apparent that the limit behaviour of $\{W_n\}$ is closely related to the limit behaviour of $\{S_n\}$ and not the same, because W_n is a function of the initial segment $\{S_k, 0 \leq k \leq n\}$ and not just the single S_n. This relation between initial segments $\{W_k, 0 \leq k \leq n\}$ and $\{S_k, 0 \leq k \leq n\}$ can be established by inducing, for each n, an appropriate stochastic process in D. Let:

$$\widetilde{W}_n \equiv \widetilde{W}_n(t) = \frac{W_{[nt]}}{a_n}, \quad 0 \leq t \leq 1$$

$$\widetilde{S}_n \equiv \widetilde{S}_n(t) = \frac{S_{[nt]}}{a_n}, \quad 0 \leq t \leq 1.$$

where a_n is a normalizing constant such that $a_n \to \infty$ as $n \to \infty$.

$\widetilde{W}_n$ and $\widetilde{S}_n$ are continuous time processes with sample paths in D. It is apparent from (5) that:

$$\widetilde{W}_n = f(\widetilde{S}_n)$$

where, $f : D \to D$ is defined by,

$$f(X)(t) = X(t) - \inf_{0 \leq s \leq t} X(s)\ , \quad 0 \leq t \leq 1.$$

Hence the desired limit theorems follow from the Donsker's theorem and continuous mapping theorem. Thus, $\widetilde{S}_n$ converges weakly to a Brownian motion and hence $\widetilde{W}_n$ converges to a reflected Brownian motion.

For the case when $\rho > 1$, $\min\{S_k, 0 \leq k \leq n\}$ in equation 5 converges weakly to a non-degenerate random variable. The limiting behaviour of W_n is obtained from the known results for random walks because from the convergence together theorem it follows that with normalization $\{W_n\}$ and $\{S_n\}$ have the same limiting behaviour.

Thus, heavy traffic limit theorems for queueing processes are proved in general by expressing them as functions of some basic processes for which limit theorems exist and then invoking theorems such as the continuous mapping theorems. See Whitt (1974) for an interesting exposition of heavy traffic limit theorems. In the above case, the basic process turned out to be a random walk.

Instead of concentrating on a single stochastic process, Reiman (1984) considered a sequence of GI/G/1 queues indexed by $n = 1,2,\ldots$ such that the traffic intensity ρ_n approaches to 1 in the limit. As a consequence, he obtained heavy traffic limits for various queue related processes. Using this approach we discuss in some detail about heavy traffic limit of *unfinished work process* in a GI/G/1 queue and only state the results for other queue related processes.

Consider a sequence of GI/G/1 queues on probability spaces $\{(\Omega_n, \mathcal{E}_n, \mathcal{P}_n)\}$ with FIFO service discipline. For each $n \geq 1$, let $\{u_i(n), i \geq 1\}$ and $\{v_i(n), i \geq 1\}$ be i.i.d. sequences of positive inter-arrival times and service times respectively, with

$$\lambda_n^{-1} = E[u_1(n)] \quad a_n = var[u_1(n)]$$
$$\mu_n^{-1} = E[v_1(n)] \quad s_n = var[v_1(n)].$$

Let

$$T_n(k) \equiv \sum_{i=1}^{n} u_i(n)\ , k \geq 1\ ,\ n \geq 1 \text{ and } T_n(0) = 0$$

be the arrival time of the k-th customer in the n-th system. With the inter-arrival time sequence and service time sequence, we can associate the following renewal processes respectively:

$$A_n(t) = max\{k \geq 0 : T_n(k) \leq t\}, \tag{6}$$

$$S_n(t) = \begin{cases} 0 & \text{if } v_1(n) \geq t \\ max\{k \geq 1 : \sum_{i=1}^{k} v_i(n) \leq t\} & \text{if } v_1(n) < t. \end{cases}$$

Further, let

$$L_n(t) = \sum_{i=1}^{A_n(t)} v_i(n), \tag{7}$$

$$V_n(t) = L_n(t) - t. \tag{8}$$

The unfinished work process $U_n(t)$ is the sum of the service times of the customers in the queue and the remaining service time of the customer in service, if any. It is easy to see that

$$U_n(t) = V_n(t) - \inf_{0 \leq s \leq t} V_n(s). \tag{9}$$

Consider the following normalized processes: For $0 \leq t \leq 1$,

$$\widetilde{A}_n(t) = n^{-1/2}[A_n(nt) - \lambda_n nt] \quad (10)$$
$$\widetilde{S}_n(t) = n^{-1/2}[S_n(nt) - \mu_n nt] \quad (11)$$
$$\widetilde{V}_n(t) = n^{-1/2}[V_n(nt)] \quad (12)$$
$$\widetilde{U}_n(t) = n^{-1/2}[U_n(nt)]. \quad (13)$$

In addition we need the following: Let

$$c_n = \sqrt{n}(\lambda_n - \mu_n) \quad (14)$$
$$X_n(t) = n^{-1/2}\sum_{i=1}^{[nt]}(v_i(n) - \mu_n^{-1}) \quad (15)$$
$$\alpha_n(t) = n^{-1}A_n(nt) \quad (16)$$
$$\eta_n(t) = n^{-1}S_n(nt), \quad \text{for } 0 \leq t \leq 1 \text{ and } n \geq 1. \quad (17)$$

Assume that

$$c_n \to c,\ \lambda_n \to \lambda,\ \mu_n \to \mu,\ a_n \to a,\ s_n \to s \quad \text{as } n \to \infty. \quad (18)$$

Further, assume that

$$\sup_{n\geq 1} E[(u_1(n))^{2+\epsilon}] < \infty \text{ for some } \epsilon > 0 \quad (19)$$

$$\sup_{n\geq 1} E[(v_1(n))^{2+\epsilon}] < \infty \text{ for some } \epsilon > 0 \quad (20)$$

Theorem 2.3.1 *If (18), (19), and (20) hold, then*

$$\widetilde{U}_n \Rightarrow \widehat{U} \equiv RBM[c/\mu, \lambda(a+s)] \text{ in } D.$$

Proof: Combining the normalized processes $\widetilde{A}_n(t), X_n(t)$, and $\alpha_n(t)$, we can write

$$\widetilde{V}_n(t) = X_n \, o \, \alpha_n(t) + \mu^{-1}[\widetilde{A}_n(t) + c_n t] \text{ for } 0 \leq t \leq 1 \text{ and } n \geq 1. \quad (21)$$

From the *functional central limit theorem for renewal processes*, it follows that $\widetilde{A}_n(t) \Rightarrow \overline{A}(t) = \text{BM}(0, \lambda^3 a)$. Hence, $\mu_n^{-1}[\widetilde{A}_n(t) + c_n t] \Rightarrow \text{BM}[c/\mu, \lambda^3\mu^{-2}a]$. Now, $\alpha_n(t) = n^{-1/2}\widetilde{A}_n(t) + \lambda_n t$. Hence, $\alpha_n(t) \Rightarrow \lambda t$ because the first term on the RHS converges to zero functional.
Thus, from the *random time change theorem*, it follows that

$$X_n \, o \, \alpha_n(t) \Rightarrow \text{BM}(0, \lambda s)$$

From the asymptotic independence of the two terms on the RHS of (21), it is easy to see that

$$\widetilde{V}_n(t) \Rightarrow \overline{V} = \text{BM}[c/\mu, \lambda(a+s)] \text{ in } D$$

as λ/μ=1 is a necessary condition for c to be finite.

If $f : D \to D$ is defined as:

$$f(X) = X(t) - \inf_{0 \le s \le t}\{X(s)\}, \quad 0 \le t \le 1, \text{ for all } X \in D,$$

then, from the continuity of f on D in Skorohod topology, we get

$$\widetilde{U}_n \Rightarrow \overline{U} = f(\overline{V}) = \text{RBM}[c/\mu, \lambda(a+s)] \text{ in } D.$$

using the *continuous mapping theorem* of §2.1. □.

Under the assumptions (18), (19), and (20), the normalized waiting time process $\widetilde{W}_n$ and the normalized queue length process $\widetilde{Q}_n$ converge weakly to the following limits respectively:

$$\widetilde{W}_n \Rightarrow \overline{U} \equiv \text{RBM}[c/\mu, \lambda(a+s)] \tag{22}$$

$$\widetilde{Q}_n \Rightarrow \overline{Q} \equiv \text{RBM}[\, c\,, \lambda^3(a+s)]. \tag{23}$$

For a proof of this see Flores (1985). Reiman (1984) extended these results to queueing networks in which K GI/G/1 queues are inter-connected to form a network and the servers serve customers in FIFO order. In this case, the vector queue length process converges weakly to a K-dimensional reflected Brownian motion on the non-negative orthant R_+^K. Also results are available for sojourn time process, which is more important than queue length process in communication networks. Also results are available for the case where different types of routing and where dependencies between the arrival and service processes exist. See Reiman (1982, 1984) for details. Flores (1985) gives a survey of the results available in heavy traffic theory.

Now, suppose that we want to approximate the behaviour of the queue length process $Q_n(.)$ when the n-th system is stable but only just so. If we set $\gamma_n = \frac{\lambda_n - \mu_n}{c}$, then $n^{1/2} \sim 1/\gamma_n$ for large n so that expansion of time scale by a factor n and normalization of the process $Q_n(.)$ that appeared in heavy traffic limit theorem is equivalent to expansion of time scale by a factor of $1/\gamma_n^2$ and normalization by a factor of $1/\gamma_n$. Thus, we can interpret equation (23) as $\gamma_n Q_n(./\gamma_n^2)$ converges weakly to RBM[c, $\lambda(a+s)$]. For each fixed t, the distribution of $\gamma_n Q_n(t/\gamma_n^2)$ converges in distribution to $\overline{Q}(t)$. Hence, for large n, we might consider approximating the behaviour of $\gamma_n Q_n(./\gamma_n^2)$. Moreover, the limiting distribution of $\overline{Q}(t)$ is given by,

$$lim_{t\to\infty} \mathcal{P}\{\overline{Q}(t) \le x\} = 1 - e^{-2|\,c\,|\,.\,x/\sigma^2} \text{ for each } x \ge 0.$$

Thus, if $Q_n(t) \Rightarrow Q_n{}'$ as $t \to \infty$, then for sufficiently large n we might approximate the distribution of $\gamma_n Q_n{}'$ by the exponential distribution with parameter $2 \mid c \mid /\sigma^2$. Diffusion approximations are based on this idea and when queueing systems are under heavy traffic such approximations yield good results. Lemoine (1978) gives a tutorial introduction to diffusion approximations.

However, when a queue is stable and if we approximate its behaviour by its limiting behaviour under heavy traffic, the effectiveness of such approximations depends on the parameters of the approximating diffusion process. Different approximating diffusions may lead to limiting diffusions with identical parameters because, in the limit, traffic intensity is equal to 1 and hence, several parameters are equal. Further, the diffusion approximations of different processes are related and thus, may lead to different approximations for the same quantity. Then these approximations need to be evaluated by their performance relative to various consistency checks. Three such approximations are given for mean in Flores (1985). These are are evaluated

according to the existing upper bounds for mean delay. Whitt (1982) discusses several possible refinements to these approximations.

In the case of queueing networks, the situation is more complicated. Here also the parameters of a limiting diffusion can be written in several ways because in the heavy traffic, the arrival and service rates are equal and can be interchanged. This gives different approximations for stable systems. However, by a careful selection of the parameters of the limiting diffusion, the exact behaviour of stable queueing systems are obtained in simple cases. For example, diffusion approximation gives exact value of mean queue length for Jackson networks. See Flores (1985) for further details.

3. Brownian networks

As seen in the previous section, when a network of queues is under heavy traffic, *i.e.*, when each queue is loaded to its capacity, various queueing processes converge weakly to multi-dimensional reflected Brownian motion. This is the underlying idea in the Brownian network model to be discussed in this section. The approximation involved is a system approximation; not just the approximation of one stochastic process by another. This feature gives dynamic control capability to any problem under consideration as we shall see in later sections.

In §3.1, we discuss development of Brownian network model for a multi-class open queueing network. In §3.2, a useful model in terms of workloads at stations is derived. In §3.3, we describe how the Brownian model of §3.1 can be modified to address the case of closed queueing networks. For this we need the following probabilistic setting which will be used throughout this paper.

A stochastic process will be described RCLL if its sample paths are right continuous and have left limits w.p.1. When we say X is a K-dimensional (μ, Σ) Brownian motion, it is assumed that there is given a *filtered probability space* $(\Omega, \mathcal{F}, \mathcal{F}_t, X, \mathcal{P}_x)$, where $(\Omega, \mathcal{F})$ is a measurable space, and $X = X(\omega)$ is a measurable mapping of Ω into $C(R^K)$ which is the space of continuous functions on R^K. $\mathcal{F}_t \equiv \sigma(X(s), s \leq t)$ is the filtration generated by X and $\mathcal{P}_x$ is a family of probability measures on Ω such that the process $\{ X(t), t \geq 0\}$ is a Brownian motion with drift vector μ and covariance matrix Σ and initial state x under $\mathcal{P}_x$. Let E_x be the expectation operator associated with $\mathcal{P}_x$. If $Y = \{Y(t), t \geq 0\}$ is a process that is $\mathcal{F}_t$ measurable for all $t \geq 0$, then we say that the process Y is non-anticipating w.r.t X when Y is adapted to the coarsest filtration w.r.t which X is adapted. (See Harrison 1985).

Three basic notions in a *Brownian network* model are:

- *resources* indexed by $i = 1, \ldots, I$
- *activities* indexed by $j = 1, \ldots, J$
- *stocks* indexed by $k = 1, \ldots, K$

The system dynamics of a *Brownian network* can be compactly expressed by:

P.3:

$$\begin{aligned} Z(t) &= X(t) + RY(t) \in S \quad \forall t \geq 0 \\ U(t) &= AY(t) \text{ is a nondecreasing process with } U(0) = 0 \end{aligned}$$

where $X(t)$ is a K-dimensional Brownian motion, R and A are $K \times J$ *input-output* matrix and $I \times J$ *resource consumption* matrix respectively. In the ensuing sections we shall see how the dynamics of a queueing network are related to that of the corresponding Brownian network.

3.1 *Brownian approximations for scheduling multiclass open queueing networks*

Consider an open queueing network with I single server stations (index i=1,...,I) and with K customer classes, indexed by k=1,...,K. It is assumed that the class designation of a customer, summarizes all relevant and observable properties of the customer, including possibly its past processing history that may be used in dynamically scheduling the network. Customers of class k arrive according to a renewal process at an average rate of λ_k. It is assumed that customers of class k visit station $s(k)$ for service and service times are i.i.d. with mean m_k and finite variance. The arrival processes and service time sequences for various classes are assumed to be mutually independent.

A customer of class k after completion of service at station $s(k)$ will turn into a class j customer with probability P_{kj} independent of previous history. The Markovian switching matrix $(K \times K)$, $P = (P_{kj})$ is assumed to be transient and hence, a customer of class k leaves the system with a positive probability $1 - \sum_j P_{kj}$. Let $C(i)$ denote the constituency of server i, *i.e.*,

$$C(i) = \{k : s(k) = i\}, \quad i = 1 \ldots I$$

From the description above, it follows that $C(i) \bigcap C(j) = \phi$, $i \neq j$. As the number of classes is allowed to be arbitrary, the above routing structure is extremely general. The case where a system is populated by various customer types, each of which has an arbitrary deterministic route through the network can also be handled by assigning different class for each combination of customer type and its stage of completion. Further, Markovian switching enables to incorporate probabilistic route structure arising out of rework, spoilage, etc.

In view of the aforementioned Brownian network model (P.3), it is easy to see that *queue lengths* correspond to stocks, *servers* at I stations play the role of resources and *servicing* of class j customer corresponds to activity j. One unit of activity j is interpreted as one time unit allocated to class j customer by server $s(j)$. Activity j consumes resource i at rate,

$$A_{ij} = \begin{cases} 1 & \text{if } i = s(j) \\ 0 & \text{otherwise.} \end{cases}$$

and total amount of resource available is, $b_i = 1$, $i = 1 \ldots I$. Queue length of class k decreases by an activity j at a rate of,

$$R_{kj} = \mu_j(\delta_{jk} - P_{jk}), \text{ where } \mu_j = 1/m_j. \tag{24}$$

where δ_{jk} is the Dirac delta function, given by

$$\delta_{jk} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise.} \end{cases}$$

In the matrix form, (24) can be rewritten as:

$$R = (I - P)^T D^{-1} \tag{25}$$

where D is the diagonal matrix with elements $m_1 \ldots m_k$ and I is the $K \times K$ identity matrix.

Since P is transient, R is non-singular with:

$$R^{-1} = D[I + P + P^2 + \cdots] \tag{26}$$

Thus, there exists a unique solution $\beta = (\beta_k)$ to:

$$\lambda - R\beta = 0 \tag{27}$$

where $\lambda = (\lambda_k)$ is the K-vector of arrival rates. Here β_k can be interpreted as the average amount of time the server $s(k)$ must assign to class k customer in order to maintain material balance over the long run.

The traffic intensity at station i, ρ_i, is defined as

$$\rho_i = \sum_{k \in C(i)} \beta_k \tag{28}$$

Define $\alpha = (\alpha_k)$, the K-vector of work load proportion by:

$$\alpha_k = \frac{\beta_k}{\rho_i}. \tag{29}$$

α_k represents the long run fraction of server's active time at station $s(k)$ that is to be devoted to class k customer in order to maintain material balance.

Basic flow processes involved in the queueing network can be given in terms of number of customers as a K-dimensional vector process, $F^j = \{F^j(t),\ t \geq 0\}$ indexed by $j = 0,1,\ldots K$. $F_k^0(t)$ is interpreted as exogenous arrival process for class k customer and $F_k^{(j)}(t)$, $j = 1,\ldots K$ is interpreted as the flow *out* of class k resulting from the t time units that server $s(i)$ devotes to class j.

Denote by R^j, the jth column of the $K \times K$ matrix R. Using the results of *renewal theory* (see Wolff 1989) or Karlin & Taylor (1981), it can be shown that

$$E[F^0(t)] \sim \lambda t \text{ and } E[F^j(t)] \sim R^j(t),\ j = 1,\ldots K \tag{30}$$

Scheduling policy is expressed as a family of allocation processes,

$$T_k = \{T_k(t),\ t \geq 0\},\ k = 1,\ldots K$$

where $T_k(t)$ gives the cumulative amount of time that server $s(k)$ allocates to class k customers during the interval $[0,t]$. Then the K-dimensional queue length process $\{Q(t),\ t \geq 0\}$ can be written in terms of the flow processes $\{F^j(t),\ t \geq 0\}$ as follows:

$$Q(t) = F^0(t) - \sum_{j=1}^{K} F^j(T_j(t)). \tag{31}$$

Similarly, the I-dimensional cumulative idle time process $\{I(t),\ t \geq 0\}$ can be defined by:

$$I_i(t) = t - \sum_{k \in C(i)} T_k(t),\ i = 1,\ldots,I. \tag{32}$$

The allocation process $T = (T_k)$ reflects a scheduling policy for the queueing network and thus, we can say T is a feasible policy if it satisfies:

$$T \text{ is continuous with } T(0) = 0. \tag{33}$$
$$T \text{ is is nondecreasing.} \tag{34}$$
$$T \text{ is nonanticipating with respect to } Q. \tag{35}$$
$$I \text{ is nondecreasing with } I(0) = 0. \tag{36}$$
$$Q(t) \geq 0 \ \forall t \geq 0. \tag{37}$$

Only (35), (36),and (37) need explanation. (35) demands that scheduling policy is to be based on observable quantities. (36) expresses that a server has only $t - s$ units of time for allocation in any interval $[s, t]$. (37) enforces that the server at station $s(k)$ must stop allocating time to class k when $Q_k(t)$ hits zero.

Thus, having expressed the basic queue processes in terms of the flow processes and the allocation process, we set out to define centred versions of these processes so as to establish connection between the system dynamics of the original queueing network and that of the approximating Brownian network, that appeared at the beginning of this section.

If we set $T_k(t) = \alpha_k t$, it is easy to see that such an allocation process T fully utilizes all available resources. In fact, observe that

$$A\alpha = b \tag{38}$$

Such an allocation is referred to as *nominal activity plan.* Now, for each $k = 1,\ldots K$, define a centred allocation process by:

$$V_k(t) \equiv \alpha_k t - T_k(t). \tag{39}$$

expressing the actual allocation to class k, $(T_k(t))$ as a decrement from the nominal alocation $(\alpha_k t)$ or in vector form (39) can be written as,

$$V(t) = \alpha t - T(t). \tag{40}$$

Similarly, we can define centred flow processes as,

$$\eta^0(t) = F^0(t) - \lambda t \text{ and } \eta^j(t) = F^j(t) - R^j t,, \ j = 1,\ldots K. \tag{41}$$

Using these centred processes the queue length process can be re-expressed as

$$Q(t) = (\eta^0(t) + \lambda t) - \sum_{j=1}^{K}[\eta^j(T_j(t)) + R^j T(t)] \tag{42}$$

$$= (\eta^0(t) + \lambda t) - \sum_{j=1}^{K}[\eta^j(T_j(t)) - RT(t)]. \tag{43}$$

(43) can be compactly written as,

$$Q(t) = \varsigma(t) + RV(t) \tag{44}$$

where,

$$\zeta(t) \;=\; \eta^0(t) \;-\; \sum_{j=1}^{K} \eta^j(T_j(t)) \;+\; (\lambda \;-\; R\,\alpha)\,t. \tag{45}$$

A similar representation for the cumulative idleness process is given as follows: observe that $I(t) \;=\; b\,t \;-\; A\,T(t)$. Hence it follows that

$$I(t) \;=\; A\,V(t). \tag{46}$$

(44) and (46) describe the system dynamics of the original queueing network and processes involved resemble the corresponding processes that appeared in the approximating Brownian network described in (P.3), with an exception that a K-dimensional Brownian motion is present instead of $\zeta(t)$. Thus, the essence of Brownian approximation lies in the approximation for ζ.
Suppose that in (45), the allocation process $T_j(t)$, is replaced by $\alpha_j\,t$ for all $j = 1, \ldots K$. Then, it is easy to verify that (45) reduces to,

$$\eta(t) \;\equiv\; F^0(t) \;-\; \sum_{j=1}^{K} F^j(\alpha_j(t)). \tag{47}$$

As the allocation process involved in (47) is the nominal one, the process $\eta(t)$ is called the nominal queue length process.

The approximation is carried out in two steps; at the first level ζ is approximated by $\eta(t)$, which in turn at the secondary level is approximated by a Brownian motion whose drift vector and covariance matrix coincide with the asymtotic drift and covariance of $\eta(t)$. If the original queueing network is under balanced heavy loading conditions and if the relevant processes are normalized in a manner consistent with the state of affairs, the above procedure provides good approximation.

The asymptotic drift vector Υ and covariance matrix Γ of the process $\eta(t)$ can be calculated using standard results of *renewal theory*. Interested reader is referred to Reiman (1984) or Harrison (1988). Thus,

$$E\{\eta(t)\} \;\sim\; \Upsilon\,t \text{ and } cov\{\eta(t)\} \;\sim\; \Gamma\,t, \text{ as } t \;\to\; \infty. \tag{48}$$

Using the *central limit theorem for random vectors* (Breiman 1968) and the *central limit theorem for renewal processes* (Wolff (1989), Karlin & Taylor (1981), it can be shown that

$$n^{-1/2}\,[\eta(n) \;-\; n\,\Upsilon] \;\xrightarrow{\mathrm{D}}\; N(0,\Gamma) \text{ as } n \;\to\; \infty.$$

Thus, the asymptotic distribution of η is the multi-variate normal distribution with mean 0 and covariance matrix Γ, or more generally , for each $t > 0$ fixed,

$$\xi^*(t) \;\equiv\; n^{-1/2}\,[\xi nt \;-\; n\,\Upsilon\,t] \;\xrightarrow{\mathrm{D}}\; N(0,\Gamma\,t) \text{ as } n \;\to\; \infty. \tag{49}$$

Thus, if $B(t)$ is a K-dimensional Brownian motion with drift 0 and covariance matrix Γ, then $\xi^*(t)$ and $B(t)$ have approximately the same distribution for each fixed t and for large n.

Now, we will discuss the scaling operation that appeared in (49). Assume that the total work load at each station is approximately equal to its capacity in the

following sense:

there exists a large integer n such that

$$n^{1/2} \mid 1 - \rho_i \mid \text{ is of moderate size for } i = 1, \ldots I. \tag{50}$$

In this case the system has balanced flow and this condition is referred to as *balanced heavy loading condition.* This n serves as an essential parameter in scaling various queueing processes. In most cases scaling expresses time as multiples of n and queue lengths as multiples of $n^{1/2}$. For example, K-dimensional scaled queue length process is defined by,

$$Z(t) \equiv n^{-1/2} Q(nt), \quad t \geq 0. \tag{51}$$

Similarly, the scaled versions of the processes ζ, V, and I are defined by,

$$X(t) \equiv n^{-1/2} \zeta(nt); \; Y(t) \equiv n^{-1/2} V(nt) \text{ and } U(t) \equiv n^{-1/2} I(nt) \tag{52}$$

(44) and (46) can be re-expressed in terms of the above scaled processes as

$$Z(t) = X(t) + RY(t) \tag{53}$$
$$U(t) = AY(t). \tag{54}$$

and the scaled and centred allocation process $Y(t)$ is feasible iff

$$Y \text{ is continuous with } Y(0) = 0 \tag{55}$$
$$Y(t) - Y(s) \leq n^{1/2} \alpha (t - s) \text{ if } t > s \tag{56}$$
$$Y \text{ is nonanticipating w.r.t } Z \tag{57}$$
$$U \text{ is nondecreasing with } U(0) = 0 \tag{58}$$
$$Z(t) \geq 0 \;\; \forall \; t \geq 0 \tag{59}$$

If we define $\theta = n^{1/2} \Upsilon$ and $T_j^*(t) = n^{-1}(T_j(nt))$, $X(t)$ can be rewritten in terms of these quantities as,

$$X(t) = n^{-1/2} [\eta^0(nt) - \sum_{j=1}^{K} n^{-1/2} \eta^j (n T_j^*(t)) + \theta t]. \tag{60}$$

Using the nominal activity plan for $T_j^*(t)$, $X(t)$ can be expressed in terms of centred and scaled nominal queue length process $\xi^*(t)$ as

$$X(t) = \xi^*(t) + \theta t. \tag{61}$$

As mentioned earlier, for sufficiently large n, $\xi^*(t)$ can be well approximated by a (0,Γ) Brownian motion. Hence, from (61), it follows that $X(t)$ can be well approximated by a (θ, Γ) Brownian motion process.

Replacement of $T_j^*(t)$ by $\alpha_j t$ can be articulated as follows: if n is large, $t > 0$ is moderate, from the balanced heavy loading condition(50), it follows that total server idleness at each station over the long interval [0,nt] is small compared to n, under any policy which calls for all servers to be busy whenever there is work for them to do. Hence, under such full allocation policies, the relative amounts of time that servers allocate to customers of their constitunecies must coincide with workload proportions (α_k) over the long run.

Thus, in the approximating Brownian network model we can take $X(t)$ to be the (θ, Γ) Brownian motion and define Z and U in terms of X and Y as given in (53) and (54). The feasibility conditions (57)–(59) can be further simplified. With X as a Brownian motion and $Z = X + RY$ by definition, condition (59) is equivalent to a conceptually simpler requirement that Y be non-anticipating w.r.t X. Constraints (57) and (58) are too stringent to impose. We can replace them by a weaker requirement that Y be RCLL. For a defence of this proposal, consider the constraint (58). For sufficiently large n, this constraint, which imposes a limit on the rate of increase of $Y_k(t)$, is loose in the sense that we can enforce rapid upward movements that closely approximate even positive jumps.

Thus, in view of the above suggested changes the approximating Brownian network takes the form:

P.3.1.1:

choose a K-dimensional RCLL process Y such that,

$$Z(t) = X(t) + RY(t) \tag{62}$$

$$U(t) = AY(t) \tag{63}$$

$$U \text{ is nondecreasing with } U(0) = 0 \tag{64}$$

$$Y \text{ is nonanticipating w.r.t } X \tag{65}$$

$$Y(0) = 0 \tag{66}$$

where $X(t)$ is a K-dimensional Brownian motion.

Once the scaling parameter n, satisfying (50) is chosen, the calculation of the drift vector θ and the covariance matrix Γ for $X(t)$ entails knowledge of only the first and second moments of the arrival and service patterns. Hence, the approximating Brownian network is insensitive to the specifc form of the arrival and service distributions.

The decision problem in the Brownian network can be transformed into a more intuitively appealing workload problem which, besides being amenable to analytical tractability, has the advantage that given a performance measure the optimal solution is easier to interpret than that obtained by solving the original problem.

3.2 *Workload formulation for a Brownian network problem*

Define an $I \times K$ matrix $M = (M_{ik})$ by

$$M \equiv AR^{-1} = AD[I + P + P^2 + \cdots] \tag{67}$$

M_{ik} represents the expected total time that server i must allocate to a class k customer before it eventually leaves the system. Define an I-dimensional *workload* process $W = (W_i)$ as

$$W \equiv MZ(t). \tag{68}$$

$W_i(t)$ gives the expected total amount of work embodied in those customers present anywhere in the network at time t. (Recall that all the processes in (68) are expressed in scaled units). The state space S of W is,

$$S = \{w \in R^I : w = MZ,\ Z \in R_+^K\}$$

Define an I-dimensional Brownian motion B as,

$$B(t) = M X(t). \tag{69}$$

The drift vector and covariance matrix of $B(t)$ are $M\,\theta$ and $M\,\Gamma\,M^T$ respectively. With the above modifications, the decision maker's problem can be redefined as:

P.3.2.1:

Choose a pair of RCLL processes $(Z\ ,\ U)$ such that

$$U \text{ is nonanticipating w.r.t } B \tag{70}$$

$$U \text{ is nondecreasing with } U(0) = 0 \tag{71}$$

$$Z(t) \geq 0 \quad \forall t \geq 0 \tag{72}$$

$$M\,Z(t) = B(t) + U(t) \quad \forall t \geq 0. \tag{73}$$

The allocation process $Y(t)$ can be expressed in terms of $(Z\ ,\ U)$ as,

$$Y(t) = R^{-1}\,[Z(t) - X(t)]$$

Equivalence of the two formulations (P.3.1.1) and (P.3.2.1) follows from the fact that $Y(t)$ given above satisfies all the coditions given in (P.3.1.1).

3.3 *The case of closed queueing networks*

In this subsection we discuss how the approximating Brownian network described in §3.1 can be modified to address the case of closed queueing networks. In a closed queueing network, a constant population of customers circulates indefinitely through the network, with no exogenous arrivals and departures. An initial queue length vector $Q(0)$ is specified apriori. Further, the switching matrix P in this case is irreducible and hence, is of rank $K-1$, which further implies that the input-output matrix is of rank $K-1$.

Analogous to the case discussed in §3.1, we seek a K-vector (β) of average activity rates satisfying,

$$R\,\beta = 0 \tag{74}$$

Equation (74) has strictly positive solution unique only up to a scale constant. Thus the traffic intensities,

$$\rho_i = \sum_{k\,\in\,C(i)} \beta_k, \quad i = 1, \ldots I. \tag{75}$$

are determined up to a scale constant. To resolve this ambiguity, the average activity rates $(\beta_1, \ldots, \beta_K)$ are scaled so that $max_i\ \rho_i = 1$. In this case the traffic intensities ρ_i express the relative amounts of work that servers at the various stations must do to maintain material balance. If we call the station k with $\rho_k = 1$ as bottleneck station, then ρ_i represents the fraction of time that server i would be kept busy if the bottleneck station is never idle.

The analog of heavy traffic condition (50) in this case is that:

There exists a large integer n such that

$$n^{1/2}\ |\ 1 - \rho_i\ |\ \text{is of moderate size} \tag{76}$$

$$n^{-1/2}\,[Q_1(0) + \cdots + Q_K(0)] \;=\; 1 \tag{77}$$

In other words, the total population size N in the network should be such that $\mid \rho_i - \rho_j \mid$ for $i \neq j$, is of order N^{-1} or smaller for each pair of i and j and that we choose $n = N^2$ as scaling parameter for the approximating Brownian network.

Following the notation of the §3.1, the vector queue length process is given by,

$$Q(t) \;=\; Q(0) \;-\; \sum_{j=1}^{K} F^j(T_j(t)). \tag{78}$$

The components of $F^j(t)$ sum to zero so that the queue length remains constant over time.

Nominal allocation for class k over $[0, t]$ can be taken to be $\alpha_k\, t$, as in the earlier case, where,

$$\alpha_k \;=\; \frac{\beta_k}{\rho_i} \quad \forall\, k \in C(i) \tag{79}$$

Embedding the initial queue length vector $Q(0)$ in the definition of $\zeta(t)$, we get

$$\zeta(t) \;=\; Q(0) \;-\; \sum_{j=1}^{K} \eta^j\,(T_j(t)) \;-\; R\,\alpha\,t. \tag{80}$$

Then the identity (44) remains valid in the closed network case. Similarly, incorporattion of $Q(0)$ in the definition of $\xi(t)$ gives the nominal queue length process as

$$\xi(t) \;=\; Q(0) \;-\; \sum_{j=1}^{K} F^j(\alpha_j\, t). \tag{81}$$

The asymptotic drift vector and covariance matrix of $\xi(t)$ satisfy,

$$e^T\,\Upsilon \;=\; 0 \;\text{ and }\; e^T\,\Gamma\,e \;=\; 0, \text{where } e \text{ is a } K - \text{dimensional sum vector.}$$

Justification for using nominal allocation in the approximating Brownian network, in this case, can be given as follows: in closed queueing networks, the decision maker's problem is to maximize rate of circulation, which boils down to maximizing the fraction of the time that any server is kept busy. Hence, full allocation policy is justified and the approximation is valid under any such policy.

A few more changes need be taken into account in the case of closed networks. The underlying Brownian motion $X(t)$ has now the initial state :

$$X(0) \;=\; Z(0) \;=\; n^{-1/2}\,Q(0). \tag{82}$$

It is easy to see that $e^T\,X(0) \;=\; e^T\,X(t) \;=\; e^T\,Z(t) = 1$, consistent with constant population size. $Z_k(t)$, hence, can be interpreted as *the fraction of the total population* that belongs to class k at time t.

Workload formulation given in §3.2 cannot be extended to the closed network case because here R is singular. But, using a modeling artifice, a similar transformation can be achieved as we shall see in §4.2.

4. Methodology and numerical results for three different queueing systems

4.1 *Scheduling a multiclass make-to-stock queue*

In a make to stock production system, products are made according to a forecast of demand and completed jobs enter a finished good inventory which services actual customer demand. Here, we consider a simple case of *make-to-stock* system with a single machine centre. K classes of products are made and service times of products of class k have a general distribution with mean m_k and finite squared co-efficient of variation, ${v_s}_k{}^2$. Demand for products of class k is a renewal process with rate λ_k and squared coefficient of variation, ${v_d}_k{}^2$. Holding cost of h_k units per unit time is incurred for maintaining inventory of class k products and a back order cost of b_k units per unit time is incurred if inventory of class k is not available.

It is assumed that ample amount of raw material is available for all types of products and also that no set up time/cost is incurred when the machine switches over from one class to another. The scheduing problem is to choose among $K+1$ options, *i.e.*,

- either work on a class k job, $k = 1, \ldots, K$
- or allow the machine to idle.

with a view to minimize the long run expected cost incurred.

Let $\{S_k(t),\ t \geq 0\}$ be the renewal process associated with the service times of class k, giving at any point of time t, number of service completions in the interval $[0,\ t]$. Let $\{D_k(t),\ t \geq 0\}$ be the point process for demands which gives number of class k demands up to time t. The inventory level process $Z_k(t)$ is given by,

$$Z_k(t) \equiv S_k(T_k(t)) - D_k(t), \tag{83}$$

where $T_k(t)$ is the allocation process which at time t gives the cumulative amount of time alloted to class k in the interval [0,t]. Thus, the decision maker's problem is to

(P.4.1.1)

choose a K-dimensional allocation policy $T = (T_k)$ to

$$\textit{minimize} \quad \limsup_{T \to \infty} \frac{1}{T} E\left[\int_0^T \sum_{k=1}^{K} c_k(Z_k(t))\, dt\right]$$

where,

$$c_k(x) = \begin{cases} h_k\, x & \text{if } x \geq 0 \\ -b_k\, x & \text{if } x < 0. \end{cases}$$

subject to

$$T \text{ is nondecreasing and continuous with } T(0) = 0 \tag{84}$$

$$T \text{ is nonanticipating w.r.t } Z \tag{85}$$

$$I \text{ is nondecreasing with } I(0) = 0 \tag{86}$$

To develop the Brownian approximation for the problem (P.4.1.1), we consider centred and scaled versions of all the related processes.

Define the traffic intensity of the system by,

$$\rho = \sum_{k=1}^{K} \rho_k \quad \text{where } \rho_k = \frac{\lambda_k}{\mu_k}. \tag{87}$$

ρ gives average server utilization required to satisfy the average demand.

Define $\alpha_k = \rho_k/\rho$ to be the proportion of the server's busy time that should be devoted to class k to meet average demand. The centred allocation process and the centred renewal process generated by service completions are given respectively as:

$$Y_k(t) = \alpha t - T_k(t) \tag{88}$$

$$\eta_k(t) = S_k(t) - \mu_k t \tag{89}$$

Furthermore define,

$$\Psi_k(t) \equiv (\mu_k \alpha_k - \lambda_k)t + \eta_k(T_k(t) - D_k(t) + \lambda_k t \quad \text{for } k = 1, \ldots K \text{ and } t \geq 0 \tag{90}$$

(Note that $\Psi(t)$ corresponds to the process $\zeta(t)$ of (45)). Then the queue length process and idle time process can be reexpressed in terms of (88) and (89) as follows:

$$Z_k(t) = \Psi_k(t) - \mu_k Y_k(t) \quad \forall\, k = 1, \ldots, K \text{ and } t \geq 0. \tag{91}$$

$$I_k(t) = \sum_{k=1}^{K} Y_k(t) \quad \forall\, t \geq 0 \tag{92}$$

Now, choosing the scaling parameter as $(1-\rho)^2$, the above basic processes will be normalized as given below. (For notational convenience, the same symbols are used for scaled processes).

$$Z_k(t) = \frac{Z_k(nt)}{\sqrt{n}}, \quad k = 1, \ldots, K \quad \forall\, t \geq 0 \tag{93}$$

$$Y_k(t) = \frac{Y_k(nt)}{\sqrt{n}}, \quad k = 1, \ldots, K \quad \forall\, t \geq 0 \tag{94}$$

$$I(t) = \frac{I(nt)}{\sqrt{n}}, \quad \forall\, t \geq 0 \tag{95}$$

We get the nominal inventory level process by replacing $T_k(t)$ in (90) by $\alpha_k t$. Then, using the central limit theorem for renewal processes(see Wolff 1989), the random time change theorem and the continuous mapping theorem discussed in §(2.1), we can show that the nominal inventory level process $\Psi_k(t)$ converges weakly to a Brownian motion $X_k(t)$ with drift $\sqrt{n}\,(\lambda_k - \mu_k \alpha_k)$ and variance $\lambda_k\,({v_{s_k}}^2 + {v_{d_k}}^2)$.

Thus, the approximating Brownian control problem for (P.4.1.1) appears as follows:

P.4.1.2 :

Choose a policy (Y) to

$$\textit{minimize} \quad \limsup_{T \to \infty} \frac{1}{T} E_x \left[\int_0^T \sum_{k=1}^{K} c_k \left(Z_k(t) \right) dt \right]$$

subject to

$$\begin{aligned} Z_k(t) &= X_k(t) - Y_k(t) \quad \text{for } k = 1, \ldots, K, \ \forall\, t \geq 0 && (96) \\ I(t) &= \sum_{k=1}^{K} Y_k(t) && (97) \\ I &\text{ is nondecreasing with } I(0) = 0 && (98) \\ Y &\text{ is nonanticipating w.r.t } X \text{ and } Y(0) = 0 && (99) \end{aligned}$$

Workload formulation:
The workload process $W(t)$, which gives at any time t the expected amount of total work embodied in the system, is given by,

$$W(t) = \sum_{k=1}^{K} m_k \, Z_k(t). \tag{100}$$

Define the one dimensional Brownian motion B by

$$B(t) = \sum_{k=1}^{K} m_k \, X_k(t), \ \forall\, t \geq 0,$$

so that B has drift $\delta = \sqrt{n}(1 - \rho) > 0$ and variance $\sum_{k=1}^{K} \lambda_k \, {m_k}^2 ({v_{s_k}}^2 + {v_{d_k}}^2)$. Then, the workload formulation for the problem (P.4.1.2) is:

P.4.1.3:

choose the pair (Z, I) so as to

$$\textit{minimize} \quad \limsup_{T \to \infty} \frac{1}{T} E_x \left[\int_0^T \sum_{k=1}^{K} c_k \left(Z_k(t) \right) dt \right]$$

subject to

$$\begin{aligned} W(t) &= B(t) - I(t), \quad \forall\, t \geq 0 && (101) \\ I &\text{ is nondecreasing with } I(0) = 0 && (102) \\ Z \text{ and } I &\text{ are nonanticipating w.r.t } B && (103) \end{aligned}$$

The equivalence of (P.4.1.2) and (P.4.1.3) can be established easily (see Wein 1992b).

The problem (P.4.1.3) is easier to solve and its solution easier to interpret in terms of the original queueing system, compared to problem (P.4.1.2). We briefly sketch the solution procedure for (P.4.1.3) and urge the reader to see Wein (1992) for further details.

Observe that given $I(t)$ at each point of time t, which satisfies the constraints (102) and (103), embedded in the problem (P.4.1.3) is a *linear programming problem.* A simple closed form solution can be obtained in terms of $W(t)$ by reformulating the

problem as a linear programming problem with seperate variables for the positive and negative parts of $Z_k(t)$.

Define the indices j and l by

$$min_{1 \le k \le K} \frac{h_k}{m_k} = \frac{h_j}{m_j} \tag{104}$$

$$min_{1 \le k \le K} \frac{b_k}{m_k} = \frac{b_l}{m_l}. \tag{105}$$

Because the problem for a given $I(t)$ has only one constraint, it is easy to see that the optimal solution for the linear programming problem is:

$$Z_k^*(t) = \begin{cases} \frac{W(t)}{m_k} & \text{if } k = j \text{ and } W(t) \ge 0, \\ 0 & \text{if } k \ne j \text{ and } W(t) \ge 0. \end{cases}$$

$$Z_k^*(t) = \begin{cases} \frac{W(t)}{m_k} & \text{if } k = l \text{ and } W(t) < 0 \\ 0 & \text{if } k \ne l \text{ and } W(t) < 0. \end{cases}$$

Hence, the optimal solution to (P.4.1.3) is dependent on $I(t)$ through $W(t)$. Thus, the work load formulation reduces to choosing an optimal policy $I(t)$, which should be an RCLL process and non-anticipating w.r.t B. So the resulting Brownian control problem is to find such an $I(t)$ to
(P.4.1.4):

$$minimize \ \limsup_{T \to \infty} \frac{1}{T} E \left[\int_0^T f(W(t)) \, dt \right]$$

subject to

$$W(t) = B(t) - I(t) \quad \forall \, t \ge 0 \tag{106}$$

where,

$$f(x) = \begin{cases} \frac{h_j x}{m_j} & \text{, if } x \ge 0 \\ \frac{b_l x}{m_l} & \text{, if } x < 0. \end{cases}$$

Observe that from the positive drift of $B(t)$ and the nature of $f(t)$, it is natural to consider a policy $I(t)$ which keeps $W(t)$ in an interval of the form $[-\infty, c]$ while exerting minimum amount of control $I(t)$. The process $W(t)$ under such a policy is called *regulated Brownian motion* on $[-\infty, c]$. A candidate policy $I(t)$, given by

$$I(t) = \sup_{0 \le s \le t} [B(s) - c]^+, \quad \forall \, t \ge 0. \tag{107}$$

satisfies all the requirements specified above [see Chapter 1 of Harrison (1985)]. Thus, if we confine ourselves to the policies of the type (107), the cost function appearing in (P.4.1.4) can be expressed as a function of c. For this, we need the following proposition from Harrison (1985).

Theorem 4.1.1 *Suppose that B is a (μ, σ^2) Brownian motion, I is as defined in (107) and thus, $W = B - I$ is an RBM on $[-\infty, c]$. Then, W has exponential steady state distribution with density,*

$$p(x) = \begin{cases} \nu\, e^{\nu(x-c)} & if\ x \leq c \\ 0 & if\ x > c \end{cases}$$

Furthermore, for each starting state $x < c$, there exists a constant C such that

$$E_x\,[W^2(t)] < C,\ \forall\, t \geq 0$$

Using the above proposition, the cost function in (P.4.1.4) can be written as

$$F(c) = -\int_{-\infty}^{0} b\, x\, \nu\, e^{\nu(x-c)}\, dx \ + \ \int_{0}^{c} h\, x\, \nu\, e^{\nu\,(x-c)}\, dx. \qquad (108)$$

and the value of c that minimizes $F(c)$ is

$$c^* = \frac{\sigma^2}{2\,\mu}\, ln(1 + \frac{b}{h})$$

with $F(c^*) = \frac{h\,\sigma^2}{2\,\mu}\, ln(1 + \frac{b}{h})$. For the proof of optimality of the policy $I^* = \sup_{0 \leq s \leq t}[B(s) - c^*]$, see Wein (1992a).

Eventhough the underlying processes in deriving the optimal policy I^* are scaled versions of the corresponding processes of the original problem, still the solution can provide insights to develop an effective scheduling policy.

Recall that $I(t)$ represents the scaled cumulative idleness process and under the optimal policy $I^*(t)$, the scaled and weighted inventory process $W(t)$ is an RBM. The process $I^*(t)$ increases only when $W(t)$ is equal to c^* or in otherwords the server is idle only at times t when $W(t)$ is c^* and otherwise is busy. Let $w(t)$ be the actual unscaled weighted inventory process. Then $W(t)$ and $w(t)$ are related by,

$$W(t) = \frac{w(nt)}{\sqrt{n}},\ t \geq 0.$$

Thus, the machine should be kept busy whenever $w(t) < \sqrt{n}\, c$ or when

$$w(t) < \frac{\sum_{k=1}^{K} \lambda_k\, m_k{}^2\, (v_{s_k}^2 + v_{d_k}^2)}{2\,(1-\rho} \, ln\,(1 + \frac{b}{h}).$$

In a similar fashion, the priority scheduling decision can be in terms of the optimal inventory level process Z^*. Whenever $w(t) < 0$, only one component of the inventory is seen to be at the positive level. In particular, no inventory is held and back orders are all of the class with the minimum value of the index $b_k\, \mu_k$ and hence the backordered demands of this class should only be satisfied when this class is the only one that is backordered at time t. In heavy traffic, the scaled number of backorders of other classes will be negligible and it does not matter in which order these backorders are satisfied. To resolve this ambiguity in priority assignment for these classes, an intuitively appealing decision would be to give priority to the class with the largest value of the index $b_k\, \mu_k$ among all the classes that are back ordered at time t.

Table 1. Data for the model in figure 1

Class	Int. arr. time distribution (mean, [std.dev])	Service distribution (mean, [std.dev])	Backorder cost	Holding cost
1	Uniform(24.0, 11.54)	Uniform(2.0, 1.0)	2.0	2.0
2	Exp(150)	Exp(150)	10.0	10.0
3	Exp(60.0)	Uniform(10.0, 2.0)	100.0	5.0
4	Uniform(20.8, 8.66)	Exp(5.0)	5.0	10.0
5	Exp(60.0)	Normal(15.0, 4.0)	5.0	5.0

Extending the same arguments to the case when $w(t) > 0$, an effective scheduling policy is to process the class with the minimum value of $h_k \mu_k$ whenever no jobs are backordered.

However, the foregiven scheduling policy has a shortcoming that it does not anticipate backorder job classes and does not respond to the class until its inventory level is negative. To compensate for this, a parametric policy is suggested in terms of parameters, ϵ_k, $k = 1,\ldots K$ which at any time t indicate which of the classes are in *danger of being backordered.* A class k is in danger of being backordered if, $Z_k(t) < \epsilon_k$ at any time t. In terms of the parameters ϵ_k, the above scheduling policy can be modified as follows: the machine is idle whenever the weighted inventory process $w(t) \geq \sqrt{n}\, c$ and no classes are in danger of being backordered; otherwise, is busy. Among the subset of the classes that are in danger of being backordered, priority is based on the value of the index $b_k \mu_k$ and the class with minimum value of this index is served first. When no class is in danger of being backordered, the machine processes the classes based on the index $h_k \mu_k$ serving the class with the minimum value of $h_k \mu_k$ first. For a detailed description of these policies see Wein (1992b) and Veatch & Wein (1992).

4.1.1 *An example: A five class make-to-stock system*

A simulation study is performed on the example described in §1.4. The processing time distributions, the customer inter-arrival distributions and backordering and holding costs for all the five classes are shown in table 1. High priority for class 3 is taken into account by assigninig high backordering cost and low holding cost. The BROWNIAN policy described above is compared against other scheduling policies such as FCFS (First Come First Served), MIN (MINimum inventory level), SEPT (Shortest Expected Processing Time) policies. In all these policies the Busy/Idle decision is according to an dependent (S-1,S) policy for each class. Under this policy, an arriving customer simultaneously takes a class k product from the inventory (if not available, backorders for one)and initiates request for a class k product. The machine centre is busy only when requests are queued. Safety stock levels for FCFS and MIN policies correspond to the optimal stock levels obtained by performing a Brownian analysis similar to that given above. In the case of SEPT policy, the safety stock levels are arbitraily selected using some intuitive arguments.

For all the above policies, safety stock levels and total cost achieved for that safety stock are given in table 2. It can be easily observed that BROWNIAN policy outperforms all the other policies.

Another simulation experiment is conducted for different utilization levels of the

Table 2. Costs for various policies.

Policy	Safety Stock	Avg. Cost
FCFS	(6, 1, 11, 4, 2)	772.2
MIN	(6, 1. 11, 4, 2)	676.8
SEPT	(0, 0, 3, 0, 0)	527.4
BROWNIAN	(0, 0, 2, 0, 0)	432.9

machine center and the costs incurred for all the above policies are presented in table 3. Referring to the results, it is seen that under light load conditions the Brownian policy is not as effective as the other policies. This behaviour is due to the fact that under low utilizations, the BROWNIAN policy tries to keep the machine center busy even though the inventory levels exceed the safety stock levels and the arrival rate of demands is very low. But at higher utilizations, the BROWNIAN policy dominates the other policies.

4.2 *Scheduling a two-station closed queueing network*

Here we consider the problem of optimally scheduling a two station closed queueing network with K customer classes to maximize the long run expected average throughput of the network. We describe a Brownian model for the problem under the setting given in §§3.1 and 3.4 and follow the same notation given there.

An approximating Brownian network is developed along the same lines as described earlier except for a change in the scaling phenomenon. Here scaling re-expresses time as multiples of N^2 and queue length as multiples of N, where N is the total population size, *i.e.*,

$$Z_k(t) = \frac{Q_k(N^2\,t)}{N} \quad \forall\ k = 1, \ldots, K \tag{109}$$

$$U_i(t) = I_i(t) = \frac{I_i(N^2\,t)}{N} \text{ for } i = 1, 2,\ \forall\, t \geq 0 \tag{110}$$

The allocation process $T(t)$ is centred by the vector $\alpha = (\alpha_k)$ of workload proportions and then scaled to give,

$$Y_k(t) = \frac{\alpha_k\, N^2\, t - T_k(N^2\, t)}{N},\ \forall\, k = 1, \ldots, K \tag{111}$$

α_k is as given in (79). Recall that in closed queueing networks $Z_k(t)$ gives the fraction of the total population that belongs to class k at any time t.

In a closed queueing network, maximizing the long run average throughput rate is equivalent to minimizing the long run average amount of idleness at either station. Without loss of generality, here we seek to minimize U_1.

Thus, the Brownian control problem is to,

(P.4.2.1):

Choose a policy (Y) to

$$\textit{minimize} \quad \limsup_{T \to \infty} \frac{1}{T}\, E\,[U_i(T)]$$

Table 3. Costs at different utilizations.

Utilization	FCFS	MIN	SEPT	BROWNIAN
0.03	282.2	282.1	282.3	1505.4
0.20	289.6	288.4	288.8	662.4
0.90	499.4	437.2	348.5	303.9
0.99	5365.5	3162.5	1144.7	921.1

subject to

$$Y \text{ is nonanticipating w.r.t } X \tag{112}$$
$$Z(t) = X(t) + RY(t), \ \forall t \geq 0 \tag{113}$$
$$U(t) = AY(t), \ \forall t \geq 0 \tag{114}$$
$$U \text{ is nondecreasing with } U(0) = 0 \tag{115}$$
$$e^T Z(t) = 1, \ \forall t \geq 0 \tag{116}$$
$$Z(t) \geq 0. \tag{117}$$

The drift δ and covariance matrix Σ of the Brownian motion X are

$$\delta = -N R \alpha \tag{118}$$
$$\Sigma = \sum_{k=1}^{K} [\alpha_k m_k^{-1} P_{kj} (\delta_{jl} - P_{kl}) + \alpha_k m_k^{-1} s_k^2 R_{jk} R_{lk}] \tag{119}$$

where R is the input-output matrix and P is the switching matrix given in §3.1. Assume that P is irreducible. As mentioned in §3.4, the above problem defies reformulation in terms of workloads for in this case, the matrix R is singular. However, the following modeling artifice obviates this difficulty.

As it is assumed that each type has its own deterministic route through the network, each class, $k = 1, \ldots, K$, corresponds to a particular stage along a type's route. Denote all the classes that correspond to the last stage along the route of some customer type as potential exit classes. Arbitrarily select a potential exit class, say K. Let $q = (q_k)$ be the Kth column of P^T. Thus, the elements of q give the probability of transitions from the potential exit class K and q_k is positive only for classes that correspond to the first stage along some customer type's route.

Define $K \times K$ matrix Δ as:

$$\Delta^j = (P^T)^j \text{ for } j = 1, \ldots, K-1 \tag{120}$$
$$\Delta^K = (0) \tag{121}$$

where B^j denotes the j th column of matrix B and (0) is the K-dimensional vector of zeroes.

Since P is irreducible, $(I - \Delta)^{-1}$ exists. Let D be the diagonal matrix with diagonal elements $m_1, \ldots, m_K$. Define the matrix H by,

$$H = D(I - \Delta)^{-1}.$$

Now, we can define workload profile matrix, $\overline{M}$ for closed networks as,

$$\overline{M} = AH. \tag{122}$$

$\overline{M}_{ik}$ represents the expected total time the server i devotes to a class k customer until that customer next completes service as a class K customer, *i.e.*, until that customer *exits*.

Now, define the two dimensional scaled workload process $W(t)$ by

$$W(t) \equiv \overline{M}\, Z(t) \tag{123}$$

$W_i(t)$ at any time t gives the expected total amount of work for server i embodied in all customers in the network at time t until they next complete service as a class K customer. Define the two dimensional Brownian motion (B) by,

$$B(t) = \overline{M}\, X(t), \ \forall t \geq 0 \tag{124}$$

B has drift $\overline{M}\, \delta$ and covariance matrix $\overline{M}\, \Sigma\, \overline{M}^T$.

However, in order to calculate the actual workload at any time t, we have to take into account the expected total time, v_i, that server i must devote to a class k customer until he next exits. It is easy to see that

$$v = \overline{M}\, q. \tag{125}$$

Average number of such *newly exiting* customers, $\theta(t)$, is given by,

$$\theta(t) = m_k{}^{-1}\, Y_k(t) \tag{126}$$

Then, the workload formulation for the problem (P.4.2.1) is given as follows:

(P.4.2.2):

choose RCLL processes $(Z,\ U,\ \theta)$ to

$$\textit{minimize} \quad \limsup_{T \to \infty} \frac{1}{T}\, E\,[U_i(T)]$$

subject to

$$\begin{aligned}
Z,\ U,\ \theta \ \text{ are } & \text{nonanticipating w.r.t } X && (127)\\
M\, Z(t) = & \ B(t) + U(t) - v\, \theta(t), \ \forall t \geq 0 && (128)\\
U \ \text{ is } & \text{nondecreasing with } U(0) = 0 && (129)\\
e^T\, Z(t) = & \ 1 \ \forall t \geq 0 && (130)\\
Z(t) \geq & \ 0 \ \forall t \geq 0 && (131)
\end{aligned}$$

For a proof of equivalence of the formulations (P.4.2.1) and (P.4.2.2) see Harrison & Wein (1990).

Interestingly, it turns out that the vector of traffic intensities ρ is proportional to the vector v, *i.e.*, $\rho_i = c\, v_i$. This observation facilitates further reduction in dimensionality of the problem (P.4.2.2). To see this, define the one-dimensional workload imbalance process $\widehat{W}$ by

$$\widehat{W} = \rho_2\, W_1(t) - \rho_1\, W_2(t), \ \forall t \geq 0 \tag{132}$$

If $\widehat{W} > 0$, then the workload in the system is imbalanced towards station 1. Define the one-dimensional Brownian motion $\widehat{B}$ by,

$$\widehat{W} = \rho_2\, B_1(t) - \rho_1\, B_2(t), \ \forall t \geq 0 \tag{133}$$

ı, $\widehat{B}$ has drift $\mu = \rho^T \overline{M} \rho$ and variance $\sigma_2 = \rho^t \overline{M} \Sigma \overline{M}^T \rho$, where $\rho = \begin{pmatrix} \rho_2 \\ -\rho_1 \end{pmatrix}$.
easy to prove that $\mu = N(\rho_2 - \rho_1)$. Define one dimensional processes R and
',

$$R = \rho_2 U_1(t), \quad \forall\, t \geq 0 \tag{134}$$
$$L = \rho_1 U_1(t), \quad \forall\, t \geq 0. \tag{135}$$

nd L can be interpreted as right and left movements exerted by the controller
).
sing the fact that $\rho = c\,v$, the workload problem (P.4.2.2) can be reformulated
single dimensional problem as follows:
1.2.3):

hoose a pair $(R\,,\ L)$ to

$$\textit{minimize} \quad \limsup_{T \to \infty} \frac{1}{T} E \left[\frac{R(t)}{\rho_2} \right]$$

ject to

$$R \text{ and } L \text{ are nonanticipating w.r.t } \widehat{B} \tag{136}$$
$$\widehat{W}(t) = \widehat{B}(t) + R(t) - L(t), \quad \forall\, t \geq 0 \tag{137}$$
$$R \text{ and } L \text{ are nondecreasing with } R(0) = L(0) = 0 \tag{138}$$
$$\widehat{W}(t) = \sum_{k=1}^{K} (\rho_2 \overline{M}_{1k} - \rho_1 \overline{M}_{2k})\, Z_k(t) \tag{139}$$
$$e^T Z(t) = 1 \quad \forall\, t \geq 0 \tag{140}$$
$$Z(t) \geq 0 \quad \forall\, t \geq 0. \tag{141}$$

athwise solution which minimizes $R(t)$ and $L(t)$ for all times t simultaneously
.1 can be found by initially ignoring the process $Z(t)$ and replacing the con-
aints (136)– (141) of the problem (P.4.2.3) by a surrogate condition that the
cess $\widehat{W}(t)$ be confined to an interval $[a, b]$. In view of the constraints (139), and
0), it is easy to see that the interval end points a and b are given respectively by,

$$a = \rho_2 \overline{M}_{12} - \rho_1 \overline{M}_{22} \equiv min_k\, (\rho_2 \overline{M}_{1k} - \rho_1 \overline{M}_{2k}) \tag{142}$$
$$b = \rho_2 \overline{M}_{11} - \rho_1 \overline{M}_{12} \equiv max_k\, (\rho_2 \overline{M}_{1k} - \rho_1 \overline{M}_{2k}) \tag{143}$$

follows that $a \leq 0 \leq b$, and class 1 customers are served at station 1 and class
at station 2.
The pair of RCLL processes (R, L) are feasible policies only if the associated
ocess $\widehat{W}(t)$ is kept within $[a, b]$. Among all the feasible policies (R, L), the polices
ven by

$$R(t) = \sup_{0 \leq s \leq t} [a - \widehat{B}(s) + L(s)]^+ \tag{144}$$
$$L(t) = \sup_{0 \leq s \leq t} [\widehat{B}(s) + R(s) - b]^+ \tag{145}$$

minimize the valves of $R(t)$ and $L(t)$ for all t simultaneouly w.p.1. (for a proof, see Chap. 2 of Harrison (1985))

For the polices defined by (144) and (145), R and L increase only when $\widehat{W}(t) = a$ and $\widehat{W}(t) = b$ respectively. To find out a control process $Z(t)$, that completes the pathwise solution, define,

$$\gamma(t) = \frac{\widehat{W}(t) - a}{b - a} \quad \forall\, t \geq 0. \tag{146}$$

Let $Z(t)$ be defined by,

$$Z(t) = \begin{cases} \gamma(t) & , \text{if } k = 1 \\ 1 - \gamma(t) & , \text{if } k = 2 \\ 0 & , \text{otherwise.} \end{cases}$$

Thus, (Z, R, L) defined by (144) and (145) respectively give a pathwise solution to the problem (P.4.2.3). Proof of this can be found in Harrison & Wein (1990). The solution (Z, U, θ) for (P.4.2.2) can be found from the existing relations.

From the solution for $Z(t)$, it follows that in heavy traffic limit only two components indexed by 1 and 2 are positive. Class 1 is served at station 1 and class 2 at station 2. This solution can be interpreted to mean that the classes 1 and 2 are to be given lowest priority at the respective stations. Under the heavy traffic condition, it does not matter in whatever order the other $K-2$ classes are served. However, to be specific, a natural ordering based on workload imbalance indices minimizes the idle time of any server. To see this, observe that idleness is incurred only when $\widehat{W}(t) = a$ or b. Order the classes now according to the values

$$\rho_2 \overline{M}_{1k} - \rho_1 \overline{M}_{2k}. \tag{147}$$

Suppose that priority rule assigns highest priority at station 1 (respectively, at station 2) to the classes with smaller (respectively, larger) values of the index. Then, the workload imbalance process $\widehat{W}(t)$ is kept within the interval $[a, b]$. As a result, idleness will be incurred less often than any other sequencing policy, such as SPT, SRPT etc. For a formal justification of this fact, see Harrison & Wein (1990). The foregoing scheduling problem in multi-class case was discussed by Chevalier & Wein (1993).

4.2.1 *An example: A closed re-entrant line*

We now present the report of simulation studies performed on the re-entrant line example of §1.4. The service time distributions for all the classes are given in table 4. In this example, machine centers 1 and 2 act as bottlenecks. The simulation experiment is conducted for different population sizes where WBAL (Workload BALancing) scheduling policy is followed at stations 1 and 2 and FCFS is followed at the machine center 3. Also the experiment is performed with other conventional scheduling policies which include FCFS, SEPT, LBFS (Last Buffer First Served) and FBFS (First Buffer First Served) policies. WBAL policy awards priorty, from high to low, to classes (1,9,8) at station 1 and (7,4,10,3,2) at station 2 whereas the priority order for SEPT policy is easily seen to be (9,8,1), (2,7,10,4,3) and (5,11,6)

Table 4. Data for the model in figure 2.

Class	Service Distribution (mean, [std.dev])
1	Exp(9.0)
2	Uniform(1.0, 0.25)
3	Exp(8.0)
4	Exp(6)
5	Uniform(2.0, 0.7)
6	Normal(6.0, 1.0)
7	Uniform(3.0, 0.7)
8	Exp(8.0)
9	Normal(6.0, 1.0)
10	Exp(5.0)
11	Exp(5.0)

ations one, two, three respectively. Mean cycle times and variances of cycle s for a given throughput rate are then compared in table **5**.
ıe reason for comparing under constant throughput rate rather than under ;ant population size is the fact that many manufacturing systems which use closed loop input will attempt to produce at the rate at which products are ınded and will choose population sizes accordingly.
ıe results are tabulated for three different throughput rates which correspond 3.9%, 63% and 99.4% utilization levels. At low and medium utilization levels ıe policies performed equally well but at the utilization level of 99.4% WBAL :y outperforms all the other policies. Further, table **5** shows WBAL policy ;ves the desired through put at lower population sizes compared with the other :ies under heavy loading conditions.

Scheduling a two-station network with controllable inputs

scheduling problem is relevant for any production system which is obliged to ıtain a specific average throughput rate of a certain product mix but can exert rol on the timing of inputs. Make-to-stock production systems stand as an nple to such situations. Advantage of controlling inputs lies in the fact that it lts in considerable reduction in WIP and in cycle times, there by improving the bility of the system.
ere we consider a simple case of a two station network with an endless queue of omers waiting to get entry into the system. Each customer has an exogenously ified class designations which are assigned such that the long-run proportion of s k customers released into the system is q_k for $k = 1, \ldots, K$, satisfying,

$$\sum_{k=1}^{K} q_k = 1$$

input decision allows full discretion over timing of release of customers into system but no control can be exerted on the choice of which class to inject. thermore, there is a lower bound $\overline{\lambda} = (\lambda_k)$, $k = 1, \ldots, K$ on the long-run

Table 5. Simulation results for the model in figure 2.

	Throughput rate = .0274		
Policy	Mean Cycle Time	Var.of Cycle time	Population
WBAL	72.88	40.82	2
FBFS	72.88	40.82	2
FCFS	72.88	40.82	2
LBFS	72.88	40.82	2
SEPT	72.88	40.82	2
	Throughput rate = 0.0433		
WBAL	138.57	206.99	6
FBFS	415.59	1789.51	18
FCFS	207.6	178.88	8
LBFS	161.58	93.31	7
SEPT	161.67	128.08	7

average throughput rate. Holding cost of c_k is incurred per unit time the class k spends in the system; but no set up costs are incurred during switchovers.

Let $\{N(t),\ t \geq 0\}$ denote the input process which at any time t gives the cumulative number of customers released into the network during the interval $[0, t]$. Scaling and centering of the input process yields $\theta(t)$ as,

$$\theta(t) = n^{-1/2}\,[\overline{\lambda}\, n\, t - N(nt)\,].$$

Following the notation of §3.1, the Brownian network formulation for the problem described is as follows:

(P.4.3.1):

Choose a pair $(Y\ ,\ \theta)$ to

$$\textit{minimize} \quad \limsup_{T \to \infty} \frac{1}{T} E_x \left[\int_0^T \sum_{k=1}^{K} c_k\, Z_k(t)\, dt \right]$$

subject to

$$Y \text{ and } \theta \quad \text{are} \quad \text{nonanticipating w.r.t } X \tag{148}$$

$$Z(t) = X(t) + RY(t) - q\,\theta(t),\ \forall\, t \geq 0 \tag{149}$$

$$U(t) = AY(t),\ \forall\, t \geq 0 \tag{150}$$

$$U \quad \text{is} \quad \text{nondecreasing with } U(0) = 0 \tag{151}$$

$$Z(t) \geq 0,\ \forall\, t \geq 0 \tag{152}$$

$$\limsup_{T \to \infty} \frac{1}{T} E_x\,[U_i(t)] \leq \gamma_i,\ i = 1, 2. \tag{153}$$

where

$$\gamma_i = \sqrt{n}\,(1 - \rho_i). \tag{154}$$

Constraint (153) is a surrogate constraint to stipulate that the longrun average throughput should be greater than or equal to $\overline{\lambda}$. As in the open network case (see §3.1), because of non-singularity of the input-output matrix R, there exists a unique non-negative K-dimensional vector $\beta = (\beta_k)$ satisfying flow balance equations,

$$\lambda = R\beta$$

$\lambda \equiv q\,\overline{\lambda}$.
e vector of traffic intensities $\rho = (\rho_i)$, $i = 1,2$, and the vector of workload ortions $\alpha = (\alpha_k)$, $k = 1,\ldots,K$ are defined as in §3.1. The drift δ and covariance ix Σ of the Brownian motion X can be computed as discussed in §3.1 .
e workload profile matrix M is defined by,

$$M = A\,R^{-1}$$

gives the expected total amount of time that server i must devote to a class k mer before it exits from the system. However, to find out total workload in ystem, variations due to input control need be accounted for.
this end, define the two dimensional vector $v = (v_i)$ by

$$v = M\,q$$

iat v_i can be interpreted as expected total amount of time the server i devotes ach customer.
ow, the two dimensional scaled workload process defined by, $W(t) = M\,Z(t)$ the additional workload due to input control, given by, $v\,\theta(t)$, gives the total load in the system at any time t. Thus, the workload formulation for (P.4.3.1)

.3.2):
ose RCLL processes (Z, U, θ) to

$$\textit{minimize} \quad \limsup_{T\to\infty} \frac{1}{T} E_x \left[\int_0^T \sum_{k=1}^{K} c_k\, Z_k(t)\, dt \right]$$

ject to

$$\begin{aligned}
Z,\ U,\ \text{and}\ \theta \quad &\text{are} \quad \text{nonanticipating w.r.t } X && (155)\\
U \quad &\text{is} \quad \text{nondecreasing with } U(0) = 0 && (156)\\
Z(t) \quad &\geq \quad 0,\ \forall\, t \geq 0 && (157)\\
\limsup_{T\to\infty} \frac{1}{T} E_x\,[U_i(T)] \quad &\leq \quad \gamma_i,\ \text{for } i = 1,2 && (158)\\
M\,Z(t) + v\,\theta(t) \quad &= \quad B(t) + U(t),\ \forall\, t \geq 0 && (159)
\end{aligned}$$

ere $B(t)$ is defined by,

$$B(t) = M\,X(t)$$

l thus, has drift $M\delta$ and covariance matrix $M\Sigma M^T$. For a proof of the equivalence P.4.3.1) and (P.4.3.2), refer Wein (1990b).
Given the policy $U(t)$ at any fixed time t, embedded in (P.4.3.2) is a linear pro- mming problem in Z and θ. As the RHS of the constraint set of (P.4.3.2) varies h t, it would be easier to consider the corresponding dual LP, which has a static nstraint set. Thus, the dual program for (P.4.3.2) is:

$$\textit{maximize}_{\pi_1(t),\pi_2(t)}\ [B_1(t) + U_1(t)]\,\pi_1(t) + [B_2(t) + U_2(t)]\,\pi_2(t)]$$

bject to

$$\begin{aligned}
M_{1k}\,\pi_1(t) + M_{2k}\,\pi_2(t) \quad &\leq \quad c_k,\ \forall\, k = 1,\ldots K && (160)\\
v_1\,\pi_1(t) + v_2\,\pi_2(t) \quad &= \quad 0. && (161)
\end{aligned}$$

It can be shown easily that $\rho_i = v_i \overline{\lambda}$, $i = 1, 2$. This fact renders it possible to simplify the dual LP further. To see this, define the workload imbalance process $\widehat{W}(t)$ by,

$$\widehat{W}(t) = \rho_2 W_1(t) - \rho_1 W_2(t), \ \forall t \geq 0 \tag{162}$$

Then, the dual LP reduces to,

$$maximize_{\pi_1(t)} \left[\frac{\widehat{W}(t)}{\rho_2} \pi_1(t) \right]$$

subject to

$$c_k^{-1} (\rho_2 M_{1k} - \rho_1 M_{2k}) \pi_1(t) \leq \rho_2. \tag{163}$$

Order the classes $k = 1, \ldots, K$ so that

$$arg_k \ max \ c_k^{-1} (\rho_2 M_{1k} - \rho_1 M_{2k}) = 1 \tag{164}$$

$$arg_k \ min \ c_k^{-1} (\rho_2 M_{1k} - \rho_1 M_{2k}) = 2. \tag{165}$$

From the complementary slackness condition, it follows that,

$$Z_k(t) = 0, \ \forall k \neq 1 \ , \text{ if } \widehat{W}(t) > 0 \tag{166}$$

$$Z_k(t) = 0, \ \forall k \neq 2 \ , \text{ if } \widehat{W}(t) < 0. \tag{167}$$

Using this, it is easy to derive that when $\widehat{W}(t) > 0$,

$$Z_k(t) = \begin{cases} \frac{\widehat{W}(t)}{\rho_2 M_{11} - \rho_1 M_{21}} & \text{if } k = 1 \\ 0 & \text{if } k \neq 1, \end{cases}$$

$$Z_k(t) = \begin{cases} \frac{\widehat{W}(t)}{\rho_2 M_{12} - \rho_1 M_{22}} & \text{if } k = 2 \\ 0 & \text{if } k \neq 2. \end{cases}$$

Thus, the optimal queue length process $Z(t)$ does not depend on the control process θ and depends on the control process U only through the workload imbalance process. The cost function corresponding to the optimal queue length process is given as a function of $\widehat{W}(t)$, *i.e.*,

$$\sum_{k=1}^{K} c_k Z_k(t) = h(\widehat{W}(t))$$

where,

$$h(x) = \begin{cases} -h_1 x & \text{if } x < 0 \\ h_2 x & \text{if } x > 0. \end{cases}$$

with $h_1 = c_2/(\rho_1 M_{22} - \rho_2 M_{12})$ and $h_1 = c_1/(\rho_1 M_{11} - \rho_2 M_{21})$.

Hence, the workload problem is reduced to finding out optimal two dimensional cumulative idleness process, U. Further simplification is possible if we define the Brownian motion $\widehat{W}$ and the right and the left control processes R and L by,

$$\widehat{W} \equiv \rho_2 B_1(t) - \rho_1 B_2(t), \ \forall t \geq 0 \tag{168}$$

$$R(t) \equiv \rho_2 U_1(t), \ \forall t \geq 0 \tag{169}$$

$$L(t) \equiv \rho_1 U_2(t), \ \forall t \geq 0 \tag{170}$$

Then, $\widehat{W} = \widehat{B}(t) + R(t) - L(t), \ \forall\, t \geq 0$. Further, notice that $\widehat{W}$ has drift $\mu = \sqrt{n}\,(\rho_1 - \rho_2)$. Using the relation (154), it is easy to see that,

$$\gamma_i = \frac{(1 - \rho_i)\,\mu}{\rho_1 - \rho_2}.$$

Now, the limiting control problem is,
(P.4.3.4):

Choose a pair (R, L) to

$$\textit{minimize} \quad \limsup_{T \to \infty} \frac{1}{T} E_x \left[\int_0^T h(\widehat{W}(t))\, dt \right]$$

subject to

$$\limsup_{T \to \infty} \frac{1}{T} E_x\,[R(t)] \leq \frac{\rho_2\,(1 - \rho_1)\,\mu}{\rho_1 - \rho_2} \tag{171}$$

$$\limsup_{T \to \infty} \frac{1}{T} E_x\,[L(t)] \leq \frac{\rho_1\,(1 - \rho_2)\,\mu}{\rho_1 - \rho_2}. \tag{172}$$

The problem (P.4.3.4) can be solved using *Lagrangian Multipliers* method. For this we need the following *Lagrangian cost function*:

$$K(x) = \limsup_{T \to \infty} \frac{1}{T} E_x \left[\int_0^T h(\widehat{W}(t))\, dt + r\,R(T) + l\,L(t) \right] \tag{173}$$

where r and l are the Lagrangian multipliers corresponding to the constraints (171) and (172). Call this problem as *Lagrangian problem.* With the aid of the following theorem, the constrained problem (P.4.3.4) can be solved by making an appropriate choice of multipliers and then minimizing the cost function $K(x)$.

Theorem 4.3.1 *Suppose r and l are nonnegative real numbers and suppose (R^*, L^*) is a solution to the Lagrangian problem. Furthermore, suppose*

$$\limsup_{T \to \infty} \frac{1}{T} E_x\,[R^*(t)] = \frac{\rho_2\,(1 - \rho_1)\,\mu}{\rho_1 - \rho_2} \tag{174}$$

$$\limsup_{T \to \infty} \frac{1}{T} E_x\,[L^*(t)] = \frac{\rho_1\,(1 - \rho_2)\,\mu}{\rho_1 - \rho_2}. \tag{175}$$

Then, (R^*, L^*) is a solution to the constrained control problem (P.4.3.4).

See Wein (1990a). Taksar (1985) developed sufficient conditions for the optimality of the Lagrangian problem. The optimal policy is one among a special class of policies called *control limit policies.* Such a policy brings the controlled process $\widehat{W}(t)$ within a certain interval $[a, b]$ instantaneously and keeps it within that interval while exerting minimum amount of control. The process $\widehat{W}(t)$ under such a policy is an RBM in the interval $[a, b]$.

The control limit policy on $[a, b]$ is defined by,

$$R(t) = \sup_{0 \leq s \leq t} [a - \widehat{B}(t) + L(s)]^+ \tag{176}$$

$$L(t) = \sup_{0 \leq s \leq t} [\widehat{B}(s) + L(s) - b]^+. \tag{177}$$

Taksar (1985) gives sufficiency conditions for a control limit policy on $[a, b]$ to be a solution to the Lagrangian problem. Wein (1990a) using this result and the theorem (4.3.2) derives sufficiency conditions for a control limit policy to be a solution to problem (P.4.3.4). Thus, the problem is reduced to finding out candidate interval end points a^* and b^* corresponding to R^* and L^* of the theorem (4.3.2).

In order to find these points a^* and b^*, the following lemma from Harrison (1985) is needed.

Lemma 4.3.1 *Let $\widehat{B}$ be a (μ, σ^2) Brownian motion and R and L be as in (176) and (177) and thus, $\widehat{W} = B + R - L$ is an RBM on the interval $[a, b]$. Then, $\widehat{W}$ has truncated exponential steady state distribution with density,*

$$p(x) = \frac{\nu\, e^{\nu\,(x\,-\,a)}}{e^{\nu\,(b\,-\,a)} - 1} \quad \text{for } a \leq x \leq b. \tag{178}$$

where $\nu = \frac{2\mu}{\sigma^2}$. Furthermore,

$$\limsup_{T \to \infty} \frac{1}{T} E_x\,[R(T)] = \frac{\mu}{e^{\nu\,(b\,-\,a)} - 1} \tag{179}$$

$$\limsup_{T \to \infty} \frac{1}{T} E_x\,[L(T)] = \frac{\mu}{1 - e^{\nu\,(b\,-\,a)}} \tag{180}$$

In view of the theorem (4.3.1), the interval end points can be found by solving the following problem:

(P.4.3.5):

Among the class of control limit policies, find a policy (R, L) to

$$\textit{minimize} \quad \limsup_{T \to \infty} \frac{1}{T} \mathrm{E}_x \left[\int_0^T h(\widehat{W}(t))\, dt \right]$$

subject to,

$$\limsup_{T \to \infty} \frac{1}{T} E_x\,[R(T)] = \frac{\rho_2\,(1 - \rho_1)\,\mu}{\rho_1 - \rho_2} \tag{181}$$

$$\limsup_{T \to \infty} \frac{1}{T} E_x\,[L(T)] = \frac{\rho_1\,(1 - \rho_2)\,\mu}{\rho_1 - \rho_2}. \tag{182}$$

The above lemma enables to express the constraints in (P.4.3.5) directly in terms of the end points a and b and thus, establishes a relation between a and b. As a result, the problem reduces to a search over values of a.

For a detailed description of the solution procedure, refer Wein (1990a). Now, in the optimal solution only class 1 customers have positive queue length whenever $\widehat{W}(t) > 0$ and class 2 customers have positive queue length whenever $\widehat{W} < 0$. This can be interpreted to mean that customers of class 1 are given the lowest priority whenever $\widehat{W}(t) > 0$. As for the priority of the other classes, a natural policy would be to award highest priority at each station at any time t to the customer with largest *reduced cost*. The reduced cost for a class k customer at time t gives the increase in the objective function value of the problem (P.4.3.3) per unit increase in RHS of the corresponding constraint. These *dynamic reduced costs*, c_k $(k = 1, \ldots, K)$ can be easily found out from the dual program (P.4.3.4). For further details, see Wein

90b). Since $w = M\,Z$, the optimal solution Z^* implies that the workload process des on the boundary of a cone in $R^2{}_+$. Further, the optimal control policies R^* l L^* are such that the control, U_1(respectively, U_2) is exerted only when $\widehat{W}(t) =$ respectively, b^*). In other words, idleness is incurred when $\widehat{W} = a^*$. This can expressed in terms of the workload process $W(t)$ using the optimal queue length ›cess.

The interval end points a^* and b^* are the reflecting barriers on the boundary the cone beyond which $W(t)$ may not enter. $W(t)$ must reside on a portion of : boundary of the cone as shown in figure 3. The optimal solution tells that ıtrols U_1 and U_2 are exerted only when $W_1(t) = c_1^*$ and $W_2(t) = c_2^*$ re- ɛctively, (See figure 3). Otherwise, only the input process $\theta(t)$ is used to keep : workload process on the boundary of the cone. To be more precise, input increased relative to the nominal input rate whenever the process $W(t)$ lies the shaded regions and input is withheld whenever $W(t)$ is inside the cone.

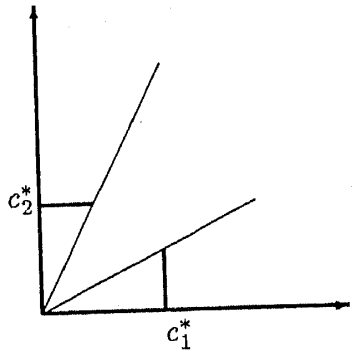

Figure 3. Cone of confinement for the process $W(t)$.

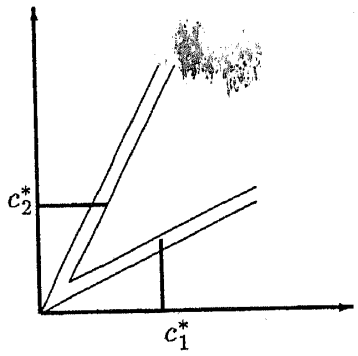

Figure 4. Inner cone to adopt ıput control.

lowever, in the actual queueing system, the process $W(t)$ may reside outside the one in figure 3. This is because the state space of the workload process is the cone $W = M\,Z,\ Z \geq 0\}$ and its extremal rays are generated by the two customer lasses with

$$arg\ max_k \frac{M_{1k}}{M_{2k}}$$

$$arg\ min_k \frac{M_{1k}}{M_{2k}}$$

which may not coincide with the classes 1 and 2 of the priority rule described earlier.

In the idealized Brownian model, when the scaled workload process is on the lower ɛay and $W_2(t) < c_1^*$, then there are zero customers at station 1; but station 1 is not ıdle according to the input rule described above. This apparent paradox is due to the scaling process involved in heavy traffic limit. Eventhough, in the actual system there are enough customers at station 1, these customers vanish in the scaled space of heavy traffic.

Table 6. Data for the model in figure 5.

Class	Service Dist.
1	Uniform(2.0, 1.7)
2	Normal(5.0, 1.0)
3	Exp(4.0)
4	Exp(8.0)
5	Normal(6.0, 1.0)
6	Exp(9.0)

In order to adopt the input rule to the actual system, it is necessary to consider a cone which is generated from the original one by building up a boundary layer of thickness, say ϵ, (see figure 4), inside the original cone. Now, the input rule admits customers as long as the workload process is in the enlarged shaded area. Selection of such a suitable ϵ is dependent on the network topology and also on how balanced the network is. Further, in the whole description given above, we have considered only the *scaled* workload process $W(t)$. In order to adopt the policy, this has to be reexpressed in unscaled terms. The procedure is the same as that has been done in the make-to-stock case.(See §4.1) and for further details, interested readers are referred to Wein (1990b). The case of multi-station closed networks is discussed in detail in Wein (1992a).

4.3.1 *An example: A two-station re-entrant line.*

A two station re-entrant line shown in figure 5 is considered for the performance study of workload regulating release policy and dynamic reduced cost based priority sequencing policy through simulation. The service distributions for the classes shown in figure 5 are given in table 6. All c_i's are assumed to be equal to 1.0. To achieve a throughput rate of 20 jobs per unit time, the values of c_1^* and c_2^* (see figure) should be 87.6 and 56.3 respectively. The boundary layer thicknesses ϵ_1 and ϵ_2 are set at 1.0. Different combinations of input release policies and priority sequencing policies are experimented. The results are presented in table 7.

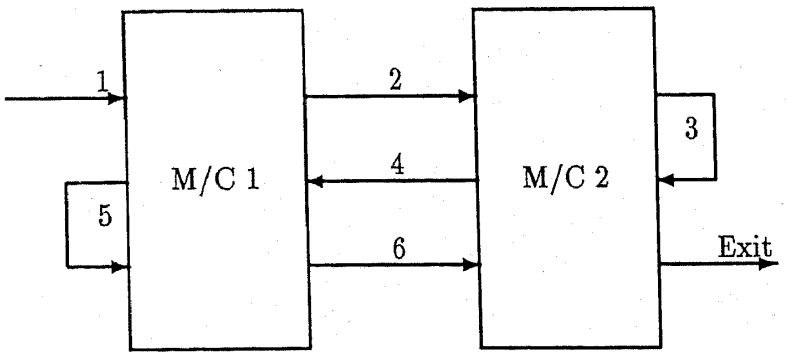

Figure 5. A two station re-entrant line

Table 7. Simulation results for the model in figure 5.

policy	Policy	Time	Cycle Time
DRC	WR	137.3	1499.4
FCFS	Deterministic	184.3	5861.1
SRPT	Deterministic	172.0	4135.2
LBFS	Deterministic	182.7	3842.6
FBFS	Deterministic	181.1	10397.0
FCFS	Poisson	255.0	16544.8
SRPT	Poisson	227.8	12025.9
LBFS	Poisson	228.9	12025.9
FBFS	Poisson	372.9	53683.6
FCFS	Uniform	190.7	7292.3
SRPT	Uniform	181.2	5398.4
LBFS	Uniform	188.7	4560.2
FBFS	Uniform	191.4	13591.3

One can see that DRC in association with WR release policy performed better than other policies in reducing cycle times where as under poisson input release policy, large cycle times are incurred. As poisson release policy can be thought of as representative to open loop release policy that is independent of the state of the system, one can say that by exercising control over input release, and thus regulating the amount of work at bottleneck stations, WIP levels can be reduced considerably.

5. Future work

The use of Brownian motion in the dynamic scheduling of multiclass queueing networks is now well established. In the past five years, there is a large body of literature on this subject. Much of the early work focussed on single station and two station networks (see Harrison & Wein 1989, 1990) and Wein (1990b, 1992b), but results are now available for multi-station networks (see Wein 1992a), Chevalier & Wein (1993). On the theoretical front, a heavy traffic limit theorem for very general networks is still eluding researchers. Also, recently, multiclass queueing networks that do not have a satisfactory Brownian network approximation in heavy traffic have been presented (Dai & Vien Nguyen 1992). This explains the need for characterizing classes of networks having Brownian networks that approximate them satisfactorily under heavy traffic assumptions. It is also an interesting open issue to investigate the range of values of traffic intensitiies for which a given network can be satisfactorily approximated by a Brownian network.

There are several interesting scheduling problems that can be attempted using Brownian approximations. There is a large variety of scheduling problems that one sees in the real world since every factory or manufacturing setup has its own peculiar and unique scheduling problems. The following gives a list of real-world features that are worthwhile to be taken into account while scheduling resources in a manufacturing facility.

- Delayed or stochastic availability of raw material: Since the raw materials are usually procured from sources external to the machine shop, one is never sure unless the raw material is in hand. Brownian models have so far assumed

a perennial supply of raw material and also do not account for raw material holding cost.

- In a real-life factory, machines or tools are prone to breakdowns and these events are non-deterministic. Also, the processed parts need not always conform to the required quality standards. Usually, periodic inspection of processed parts will decide whether the quality of parts is acceptable or not. Parts identified for reworking cause extra load on the system whereas every rejected part entails complete reprocessing and also material waste. Modelling such features is important.

- In a multiclass production system, switchover times or set-up times can have a significant effect on the way parts are scheduled. Existing Brownian models do not address the issue of scheduling in the presence set-up times and set-up costs.

- The objective function chosen for minimization in the existing literature usually takes into account factors such as inventory costs, backorder costs, mean waiting times, and machine utilizations. Since variability of performance measures is also a very important criterion, there is need to include it as part of the objective functions. There are also other measures of performance such as makespan and total tardiness. This also brings out the issue of modeling due dates.

It is also important and useful to evaluate the performance of Brownian policies and derive performance degradation of Brownian policies when the underlying network does not satisfy heavy traffic conditions. Finally, one has to look into the computational effort involved in arriving at Brownian policies for various networks.

This research was supported by the Office of Naval Research and the Department of Science and Technology grant N00014-93-1017. We would also like to acknowledge the encouragement and comments of Professor N. Viswanadham and several critical comments of the reviewers of this paper.

References

Abdul-Razaq T S, Potts C N 1988 Dynamic programming state space relaxation for single machine scheduling. *J. Oper. Res. Soc.* 39: 141–152

Bagchi U, Ahmadi R H 1987 An improved lower bound for minimizing weighted completion times with deadlines. *Oper. Res.* 35: 311–313

Baker K R, Ahmadi R H 1978 Finding an optimal sequence by dynamic programming: An extension to precedence-related tasks. *Oper. Res.* 35: 111-120

Beloudah H, Posner M E, Potts C N 1988 A branch and bound algorithm for scheduling jobs with release dates on a single machine to minimize total weighted completion time. Preprint, Faculty of Mathematical Studies, University of Southampton

Billingsley P 1968 *Convergence of probability measures* (New York: John Wiley and Sons)

Breiman L 1968 *Probability* (Reading, MA: Addison-Wesley)

Chevalier P B, Wein L M 1993 Scheduling network of queues: Heavy traffic analysis of a multistation closed network. *Oper. Res.* 41: 743–758

Cox D R, Miller H D 1965 *The theory of stochastic processes* (London: Methuen)

Dai J G, Vien Nguyen 1992 On the convergence of multiclass queueing networks in heavy traffic. School of Mathematics and Industrial/Systems Engineering, Georgia Institute of Technology

Fisher M L 1973 Optimal solution of scheduling problems using Lagrangian multipliers. *Oper. Res.* 21: 1114 –1127

Fisher M L 1981 Lagrangian relaxation method for integer programmming problems. *Manage. Sci.* 27: 1–18

Flores C 1985 Diffusion approximations for computer communication networks. *Proc. Symp. Appl. Math.* : 83–124

Forst F G 1984 A review of the static, stochastic, job sequencing literature. *Opsearch* 21: 127–144

Gere W S 1987 Heuristics in job shop scheduling. *Manage. Sci.* 13: 167-190

Goldberg D E 1986 *Genetic algorithms in search, optimization, and machine learning.* (Reading, MA: Addison-Wesley)

Hajek B 1974 Optimal control of two interacting service stations. *IEEE Trans. Autom. Contr.* 29: 491–499

Harrison J M 1985 *Brownian motion and stochastic flow systems* (New York: John Wiley and Sons)

Harrison J M 1988 Brownian models of queueing networks with heterogeneous customer populations. *Stochastic differential systems, stochastic control theory and applications, IMA* (eds) W Fleming, P L Lions (New York: Springer-Verlag) 10: 147–186

Harrison J M, Reiman M 1981 Reflected Brownian motion on an orthant. *J. Appl. Probab.* 9: 302–308

Harrison J M, Wein L M 1989 Scheduling network of queues: Heavy traffic analysis of a simple open network. *Queueing Syst.: Theor. Appl.* 5: 265–280

Harrison J M, Wein L M 1990 Scheduling network of queues: Heavy traffic analysis of a two-station closed network. *Oper. Res.* 38: 1052–1064

Harrison J M, Williams R J 1987 Multidimensional reflected Brownian motion having exponential stationary distributions. *Ann. Probab.* 15: 115-137

Hochbaum D S, Shmoys D B 1989 A polynomial approximation scheme for machine scheduling on uniform processors using the dual approximation approach. *SIAM J. Comput.* 17: 539-555

Karlin S, Taylor H M 1981 *A first course in stochastic processes* (New York: Academic Press)

Klimov G P 1974 Time sharing service systems. *Theor. Appl. Probab. Appl.* 19: 532–551

Lawler J K, Lenstra J K, Rinooy Kan A H G, Shmoys D B 1990 Sequencing and scheduling: Algorithms and complexity. *Handbook of operational research and management* (eds) S C Graves, A H G Rinooy Kan, P Zipkin (Amsterdam: North-Holland) 4: 51–61

Lemoine A J 1978 Network of queues: A survey of weak convergence results. *Manage. Sci.* 24: 175–1193

Levy B C, Adams M B 1987 Global optimization with stochastic neural networks. *Proceedings of IEEE International Conference on Neural Networks*, San Diego, pp 681–686

Pinedo M L 1981 A note on the two-machine job shop with exponential processing times. *Naval Res. Logist. Q.* 28: 693–696

Pinedo M L 1982 Minimizing the expected make-span in stochastic flow shops. *Oper. Res.* 30: 148–162

Pinedo M L 1983 Stochastic scheduling with release dates and due dates. *Oper. Res.* 31: 559–572

Pinedo M L, Scrage L 1982 Stochastic shop scheduling: A survey. *Deterministic and stochastic scheduling* (eds) M A H Dempster, J K Lenstra, A H G Rinooy Kan (Dordrecht: Reidel) pp 181–196

Pinedo M L, Weiss G 1980 Scheduling tasks with exponential service times on non-identical processors to minimize various cost functions. *J. Appl. Probab.* 17: 187–202

Pinedo M L, Weiss G 1987 The "Largest variation first" policy in some stochastic scheduling problems. *Oper. Res.* 35: 884–894

Pollard D 1984 *Convergence of stochastic processes* (New York: Springer-Verlag)

Reiman M I 1982 Heavy traffic diffusion approximation for sojourn times in Jackson networks. *Applied probability, computer science, the interface* (eds) R L Disney, T T Ott (Boston: Birkhauser) pp 409-422

Reiman M I 1984 Open queueing networks in heavy traffic. *Math. Oper. Res.* 9: 441–458

Ross K W, Yao D D 1989 Optical dynamic scheduling in Jackson networks. *IEEE Trans. Autom. Contr.* 34: 47–53

Stidham S Jr 1985 Optimal control of admission to a queueing system. *IEEE Trans. Autom. Contr.* 30: 44–52

Taksar W 1985 Average optimal singular control and a related stopping problem. *Math. Oper. Res.* 10: 63–81

Van Laarhoven P J M, Aarts E H L, Lenstra J K 1992 Job shop scheduling by simulated annealing. *Oper. Res.* 40: 1156–1179

Veatch M H, Wein L M 1992 Scheduling a make-to-stock queue: Index policies and hedging points. Working Paper, Massachusetts Institute of Technology, September

Walrand J 1988 *An introduction to queueing networks* (Englewood Cliffs, NJ: Prentice Hall)

Weber R R, Stidham S 1987 Optimal control of service rates in network of queues. *Adv. Appl. Probab.* 19: 202–218

Wein L M 1990a Optimal control of a two-station Brownian network. *Math. Oper. Res.* 38: 215–242

Wein L M 1990b Scheduling network of queues: Heavy traffic analysis of a two-station network with controllable inputs. *Oper. Res.* 38: 1065–1078

Wein L M 1992 Scheduling network of queues: Heavy traffic analysis of a multi-station network with controllable inputs. *Oper. Res.* 40: S312–S334

Weiss G 1982 Multiserver stochastic scheduling. *Deterministic and stochastic scheduling* (eds) M A H Dempster, J K Lawler, A H G Rinooy Kan (Dordrecht: Reidel) pp 157–179

Whitt W 1974 Heavy traffic limit theorems for queues: A survey. *Mathematical methods in queueing theory* (ed.) A B Clarke (New York: Springer-Verlag) pp 307–350

Whitt W 1985 Some useful functions for functional central limit theorem. *Math. Oper. Res.* 5: 67–85

Whitt W 1982 Refining diffusion approximations for queues. *Oper. Res. Lett.* 1: 165–169

Wolff R A 1989 *Stochastic processes and queueing theory* (Englewood Cliffs, NJ: Prentice Hall)

Yao D D, Shantikumar J G 1990 Optimal scheduling control of a flexible machine. *IEEE Trans. Robotics Autom.* 6: 706–712