On Dynamic Scheduling of Stochastic Networks in Heavy Traffic and Some New Results for the Workload Process

M. Bramson[1]
School of Mathematics
University of Minnesota
Minneapolis MN 55455
bramson@math.umn.edu

R. J. Williams[2]
Department of Mathematics
University of California, San Diego
La Jolla CA 92093-0112
williams@math.ucsd.edu

Abstract

Dynamic scheduling of stochastic networks has applications to the control of modern telecommunications, manufacturing and computer systems. Most models of such networks cannot be analyzed exactly, and one is naturally led to consider more viable approximations. As one approach, J. M. Harrison proposed Brownian control problems (BCPs) as formal heavy traffic approximations to dynamic scheduling problems for stochastic networks. Subsequently, various authors combined analysis of BCPs with interpretation of their optimal solutions to suggest original and attractive policies for certain specific stochastic network control problems. Despite these successes for specific problems, there is, as yet, no general rigorous approach to analyzing BCPs, inferring good policies from their solutions, and proving asymptotic optimality of such policies. We are interested in developing such an approach. This paper is a step in that direction. In particular, we (a) provide a detailed stochastic network model, (b) give a fluid model interpretation of the notion of heavy traffic, (c) derive a formula for the dimension of the workload process in terms of basic model parameters, and (d) identify a mild condition under which the components of the workload process are all non-negative.

1 Introduction

Stochastic networks are used as models for complex manufacturing, telecommunications and computer systems. Some of these networks allow for flexible scheduling of jobs through dynamic (state-dependent) alternate routing and sequencing, hereafter collectively referred to as dynamic scheduling. Usually these models cannot be analyzed exactly, and it is a challenging problem to design control policies for such networks that are simple to implement and yet are approximately optimal in an appropriate sense. As one approach to this problem, some authors have followed a scheme first suggested by Harrison [5], where analysis of Brownian control problems (formal heavy traffic approximations) is combined with interpretation of their optimal solutions to suggest "good" policies for stochastic network control problems. These policies have usually performed well when simulated, and there are a few proofs of asymptotic optimality for specific networks [8, 12, 13, 15, 17]. However, despite these successes, there is, as yet, no general rigorous approach to analyzing Brownian control problems, inferring good policies from their solutions, and proving asymptotic optimality of such policies. This paper is a contribution to the development of such an approach.

Before describing the results of this paper, we first summarize the steps (I)–(V) involved in applying the Brownian control problem approach. Some or all of these steps have been used, at least implicitly, by the authors who have followed Harrison's scheme.

(I) Formulate a stochastic network model.
(II) Formulate a notion of heavy traffic. (There is no conventional notion of heavy traffic, since a general network model should allow alternate routing, and the nominal or average load on a server may then depend on the routing of jobs.)
(III) Formulate a formal diffusion approximation, i.e., a Brownian control problem (BCP), for the network control problem. Reduce the dimension of this problem by deriving an equivalent workload formulation (EWF).
(IV) Analyze the BCP (or EWF) and "interpret" its solution by giving a proposed control policy for the original network.
(V) Investigate the performance of the policy pro-

[1]Research supported in part by NSF Grant DMS-9971248.
[2]Research supported in part by NSF Grant DMS-0071408, and a gift from the David and Holly Mendel Fund.

posed in (IV). In particular, determine whether it is asymptotically optimal (in the heavy traffic limit) and whether it achieves the same cost as the solution of the BCP.

For step (I), Harrison [7] proposed a very general model. However, this model is not defined in terms of primitive network processes; for the purposes of applying stochastic network models and also for proving asymptotic optimality of a control policy, a more detailed model is desirable. For an important subcollection of the situations encompassed by [7], such a detailed model is described in Section 2. This model is used throughout the current paper. For step (II), Harrison [7] (see also [10, 16]) proposed a notion of heavy traffic in terms of linear programs. An interpretation of this condition, in terms of fluid models, is described in Section 3. It is anticipated, based on experience gained in proving asymptotic optimality in [1, 2], that, in general, this interpretation and related results will be important for step (V). Step (III) was initiated by Harrison in [5] for networks with dynamic sequencing capabilities, and expanded on by Kelly-Laws [11] for examples of networks that also allow dynamic routing. A Brownian model covering both types of control was described in Harrison [7], and a dimension reduction (or state-space collapse) of this model was developed by Harrison et al. in [7, 10]. In the examples that have been analyzed to date, this reduced model (the equivalent workload formulation), has played a key role in solving the BCP. The Brownian control problem is reviewed in Section 4 and its equivalent workload formulation is described in Section 5.

Except in some special cases, such as when the workload process is one-dimensional (and non-negative), it is currently not known how to solve the BCP or EWF analytically. In Section 6, we present some new results concerning properties of the workload process. These are likely to be of use in solving the EWF. In particular, we characterize the dimension of the workload process in terms of basic model parameters, and identify a mild condition under which the components of the workload process are all non-negative.

With regard to steps (IV) and (V), there is no systematic method (other than by discretization [6, 14]) for interpreting solutions of the BCP when they are available, and there are very few proofs of asymptotic optimality of control policies derived using the formal BCP methodology [8, 12, 13, 15, 17]. When the workload process is one-dimensional (and non-negative), the EWF, and hence the BCP, can be solved explicitly (cf. [9, 19]). However, it is by no means obvious how to interpret this solution in terms of the original stochastic network, nor how to prove asymptotic optimality of a proposed policy. In separate related work of Williams [19] and Bell-Williams [1, 2], a systematic execution of steps (IV) and (V) is illustrated for a class of models known as parallel server systems, under a complete resource pooling condition (i.e., the workload process is one-dimensional). In particular, the solution of the BCP is interpreted in terms of a "continuous review threshold policy" and it is shown that this policy is asymptotically optimal.

2 Stochastic Network Model

Network Structure. A schematic for the model is shown in Figure 1. There are $\mathbf{I}$ infinite capacity buffers for holding jobs awaiting service and $\mathbf{K}$ (non-identical) servers for processing jobs. Arrivals to a buffer may come from outside the system and/or from the internal movement of jobs that have already received at least one service in the system. Several different servers may be capable of processing (or serving) jobs from a particular buffer. Service of a given buffer i by a given server k is called a processing activity. It is assumed that there is at most one activity for each pair (i, k) and so the number of activities $\mathbf{J}$ is at most $\mathbf{I} \cdot \mathbf{K}$. The activities are labelled by $j = 1, \ldots, \mathbf{J}$. The correspondences between activities and buffers, and activities and servers, are described by two deterministic matrices C and A, where C is an $\mathbf{I} \times \mathbf{J}$ matrix with $C_{ij} = 1$ if activity j processes buffer i, and $C_{ij} = 0$ otherwise, and A is a $\mathbf{K} \times \mathbf{J}$ matrix with $A_{kj} = 1$ if server k performs activity j, and $A_{kj} = 0$ otherwise. Each activity j has exactly one buffer and one server associated with it, and so each column of C contains the number one exactly once and similarly for A. It is also assumed that each row of C and each row of A contains the number one at least once (i.e., each buffer is capable of being processed by at least one activity and each server is capable of performing at least one activity).

It is assumed that once a job has commenced service under an activity, it must complete its service with that activity, even if its service is interrupted for some time (e.g., by preemption by a job from another buffer). In addition, an activity must complete service of any job it has started serving before commencing service of another job from its associated buffer. When taking a new job from a buffer, a server always takes the oldest arrival in the buffer that has not yet commenced service. (This last restriction corresponds to an HL (head-of-the-line) discipline in systems without dynamic routing, cf. [3, 18].)

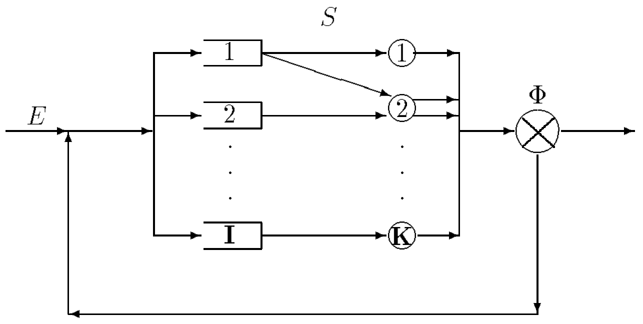

Figure 1: Schematic for a stochastic network

Stochastic Primitives. The primitive stochastic processes for the model are given by a triple (E, S, Φ). Here, E is an $\mathbf{I}$-dimensional exogenous arrival process such that for each i, $E_i(t)$ is a (delayed) renewal process which denotes the number of jobs that have arrived to buffer i from outside the system in $[0,t]$. The long run arrival rate vector associated with E is a non-negative vector α. For each activity j, there is a (delayed) renewal process S_j such that $S_j(t)$ denotes the number of complete jobs that could be processed by activity j in $[0,t]$ if the associated server worked continuously and exclusively on jobs from the associated buffer in $[0,t]$. The long run rate vector associated with S is a strictly positive vector β. The process $\Phi = (\Phi^1, \Phi^2, \ldots, \Phi^{\mathbf{J}})$, where Φ^j is an $\mathbf{I}$-dimensional routing process associated with activity j such that $\Phi^j(n) = \sum_{\ell=1}^{n} \phi^j(\ell)$, and $\{\phi^j(\ell)\}_{\ell=1}^{\infty}$ is a sequence of i.i.d. random routing vectors taking values in $\{e^0, e^1, \ldots, e^{\mathbf{I}}\}$. Here, e^0 is the identically zero vector and $e^1, \ldots, e^{\mathbf{I}}$ are the non-negative unit basis vectors in $\mathbb{R}^{\mathbf{I}}$. If $\phi^j(\ell) = e^i$ for $i \in \{1, \ldots, \mathbf{I}\}$, one interprets this to mean that, upon completion of its service, the ℓ^{th} job processed by activity j is sent to buffer i, and if $\phi^j(\ell) = e^0$, the job exits the system. Let $P_{ji} = \mathbf{P}(\phi^j(\ell) = e^i)$, $i = 1, \ldots, \mathbf{I}$, $j = 1, \ldots, \mathbf{J}$. In addition to (E, S, Φ), a random variable $\chi(0)$ describing the initial state of the system is needed.

Control. Scheduling control is exerted through allocations of server time by each server to its associated processing activities. Formally, this is specified by a $\mathbf{J}$-dimensional stochastic process $T = \{T(t), t \geq 0\}$ where for $j = 1, \ldots, \mathbf{J}$ and $t \geq 0$, $T_j(t)$ is the cumulative amount of service time devoted to activity j by the associated server in the time interval $[0,t]$. The process T must satisfy certain properties that go along with its interpretation. In particular, there will, in general, be constraints on what T can depend on. Here, it is assumed that T is adapted to a filtration $\{\mathcal{F}_t, t \geq 0\}$, where $\mathcal{F}_t$ is a σ-algebra denoting the information known about the history of the system at time t. Typically this filtration is endogenous (i.e., depends on T) and may depend on the situation being modeled.

Performance Processes and Model Equations. Processes used for measuring the performance of the network under a given scheduling control policy T are the queue length process Q and the cumulative idle-time process I. For each buffer i, $Q_i(t)$ is the number of jobs in buffer i at time t, including any jobs from that buffer which have been partially served. For each server k, $I_k(t)$ denotes the total amount of time in $[0,t]$ that server k has been idle. Noting that $S_j(T_j(t))$ is the number of complete jobs that have been processed by activity j in the time interval $[0,t]$, one sees that the descriptive processes (Q, I) satisfy the following vector equations:

$$
\begin{aligned}
Q(t) &= Q(0) + E(t) + \Phi(S(T(t)) - CS(T(t)), \\
I(t) &= et - AT(t),
\end{aligned}
$$

where e is the $\mathbf{K}$-dimensional vector of all ones. (Here, notation has been abused slightly in that the i^{th} component of $\Phi(S(T(t)))$ is $\sum_{j=1}^{\mathbf{J}} \Phi_i^j(S_j(T_j(t)))$, which corresponds to the number of arrivals to buffer i from *inside* the network in $[0,t]$, and the i^{th} component of $CS(T(t))$ is $\sum_{j=1}^{\mathbf{J}} C_{ij} S_j(T_j(t))$, which corresponds to the number of departures from buffer i in $[0,t]$.) Note that for each k and i, I_k is continuous and non-decreasing, $I_k(0) = 0$, and $Q_i(t) \geq 0$ for all $t \geq 0$.

Relationship with other models. The stochastic network model introduced above is more specific than the very general model in Harrison [7]. In particular, it uses more primitive stochastic processes than [7] and thereby has some elements in common with [3, 12, 18, 19]. Also, in our model, each activity serves exactly one buffer and is processed by exactly one server, whereas Harrison [7] allows a more general setup.

3 Fluid Model, Heavy Traffic and a Linear Program

The *fluid model* corresponding to the network model described above can be thought of as a formal deterministic analogue in which the primitive stochastic processes E, S and Φ are replaced by linear flows which move at rates equal to the long run rates of the original processes. In addition, the processes Q, I and T are replaced by flows $\bar{Q}$, $\bar{I}$ and $\bar{T}$ which satisfy the following for all allowed values of i, j and k:

$$\begin{aligned}
&\bar{Q}(t) = \bar{Q}(0) + \alpha t - R\bar{T}(t), \\
&\bar{I}(t) = et - A\bar{T}(t), \\
&\bar{Q}_i(t) \geq 0 \text{ for all } t \geq 0, \\
&\bar{T}_j \text{ is non-decreasing, Lipschitz continuous} \\
&\quad \text{with Lipschitz constant 1, and } \bar{T}_j(0) = 0, \\
&\bar{I}_k \text{ is continuous \& non-decreasing, } \bar{I}_k(0) = 0,
\end{aligned}$$

where $R \equiv (C - P')\mathcal{B}$, $\mathcal{B}$ is the diagonal matrix with the entries of the service rate vector β on its diagonal, and $'$ denotes transpose. The above conditions define the fluid model and a $\bar{T}$ satisfying these conditions is called a *fluid control.* For a given fluid control $\bar{T}$, the fluid system is said to be *balanced* if the associated fluid "queue length" $\bar{Q}$ does not change with time. In addition, the fluid system *incurs no idleness* (or all fluid servers are fully occupied) if $\bar{I} \equiv 0$, i.e., $A\bar{T}(t) = et$ for all t.

Definition. *The fluid model is in heavy traffic if*
(i) there is a unique fluid control $\bar{T}^*$ under which the fluid system is balanced, and
(ii) under $\bar{T}^*$, the fluid system incurs no idleness.

Since any fluid control is differentiable at almost every time, one can readily convert the above notion of heavy traffic into one involving the rates $x^*(t) = \frac{d}{dt}\bar{T}^*(t)$. This leads to the notion of heavy traffic as formulated by Harrison [7] (see also [10, 16]) in terms of the linear program LP specified below.

Lemma 3.1 *(cf. [19]) The fluid model is in heavy traffic if and only if the following two conditions hold:*
(i) there is a unique optimal solution (x^*, ρ^*) of the following linear program (LP):

$$\textit{minimize } \rho \ \ \textit{s.t. } Rx = \alpha, \quad Ax \leq \rho e \quad \textit{and} \quad x \geq 0,$$

(ii) the solution of the linear program is such that $\rho^* = 1$ and $Ax^* = e$.

Assumption. *For the remainder of this paper, we assume that the fluid model is in heavy traffic.*

4 Brownian Control Problem

The Brownian control problem (BCP) is a formal approximation (under diffusive scaling) to a control problem for the stochastic network (cf. [7, 10]). Mathematically this is handled by considering a sequence of stochastic networks indexed by r, where r tends to infinity through a sequence of positive real numbers. The networks in this sequence all have the same basic structure as that described in Section 2, except that the control policy, initial conditions and form of the cost function (which is defined below) may depend on r. Quantities associated with the system indexed by r will have a superscript r attached to them. For concreteness, the following average cumulative discounted holding cost will be used for the r^{th} network:

$$J^r(T^r) \quad = \quad \mathbf{E}\left(\int_0^\infty e^{-\gamma t}\, h'\hat{Q}^r(t)\, dt\right),$$

where $\gamma > 0$ is a fixed constant called the discount factor, $h = (h_1, \ldots, h_{\mathbf{I}})'$ with $h_i > 0$ for $i = 1, \ldots, \mathbf{I}$, is a constant vector of holding costs, $\hat{Q}^r(\cdot) = r^{-1}Q^r(r^2\cdot)$ is the normalized queue length process associated with the control T^r, and $\mathbf{E}$ denotes expectation. Consider the centered and normalized processes $\hat{E}^r(t) = r^{-1}\left(E(r^2t) - \alpha r^2 t\right)$, $\hat{S}^r(t) = r^{-1}\left(S(r^2t) - \beta r^2 t\right)$, and $\hat{\Phi}^{j,r}(t) = r^{-1}\left(\Phi^{j,r}([r^2t]) - P^j[r^2t]\right)$, where P^j denotes the column vector whose entries are given by the j^{th} row of the matrix P, and $[\cdot]$ denotes the integer part. Under suitable second moment and independence conditions on the stochastic primitives (cf. [18]), one has a functional central limit theorem result of the form:

$$(\hat{Q}^r(0), \hat{E}^r, \hat{S}^r, \hat{\Phi}^r) \quad \Longrightarrow \quad (\tilde{Q}(0), \tilde{E}, \tilde{S}, \tilde{\Phi}),$$

as $r \to \infty$, where $\hat{\Phi}^r = (\hat{\Phi}^{1,r}, \ldots, \hat{\Phi}^{\mathbf{J},r})$, $\tilde{\Phi} = (\tilde{\Phi}^1, \ldots, \tilde{\Phi}^{\mathbf{J}})$, and $\tilde{E}$, $\tilde{S}$, $\tilde{\Phi}^j$, $j = 1, \ldots, \mathbf{J}$ are independent Brownian motions that are jointly independent of $\tilde{Q}(0)$, and $\Longrightarrow$ denotes convergence in distribution for stochastic processes.

An important assumption in the formal derivation of the BCP is that, in the fluid (law of large numbers) limit, the allocation processes achieve the nominal levels $\bar{T}^*(t) \equiv x^*t$. The control process $\tilde{Y}$ appearing below is the formal limit of the *deviation processes* $\hat{Y}^r(t) = r^{-1}\left(x^*r^2t - T^r(r^2t)\right)$. A consequence of this is that for j such that $x_j^* > 0$, there is no sign constraint on $\tilde{Y}_j$ (indeed, $\tilde{Y}_j$ need not even be of bounded variation), whereas for j such that $x_j^* = 0$, $\tilde{Y}_j$ must be non-positive and non-increasing. The activities j for which $x_j^* > 0$ are called *basic* activities and those for which $x_j^* = 0$ are called *non-basic* activities. The non-basic activities only come into play at the diffusion level, not at the coarser fluid level. Without loss of generality, it can be assumed that the first $\mathbf{B}$ activities are the basic ones and the last $\mathbf{J} - \mathbf{B}$ are the non-basic ones. For later reference, we partition the matrices R and A: $R = [H \quad M]$ and $A = [B \quad N]$, where H is $\mathbf{I} \times \mathbf{B}$, M is $\mathbf{I} \times (\mathbf{J} - \mathbf{B})$, B is $\mathbf{K} \times \mathbf{B}$ and N is $\mathbf{K} \times (\mathbf{J} - \mathbf{B})$.

Brownian control problem (BCP).

$$\text{minimize } \mathbf{E}\left(\int_0^\infty e^{-\gamma t}\, h'\tilde{Q}(t)\, dt\right)$$

using a $\mathbf{J}$*-dimensional control process* $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_{\mathbf{J}})'$ *such that*

$$\begin{aligned}
&\tilde{Q}(t) = \tilde{Q}(0) + \tilde{X}(t) + R\tilde{Y}(t) \quad \textit{for all } t \geq 0,\\
&\tilde{I}(t) = A\tilde{Y}(t) \quad \textit{for all } t \geq 0,\\
&\tilde{Q}_i(t) \geq 0 \textit{ for all } t \geq 0,\ i = 1, \ldots, \mathbf{I},\\
&\tilde{I}_k \textit{ is non-decreasing, } \tilde{I}_k(0) = 0,\ k = 1, \ldots, \mathbf{K},\\
&\tilde{Y}_j \textit{ is non-increasing, } \tilde{Y}_j(0) = 0,\ j = \mathbf{B}+1, \ldots, \mathbf{J},\\
&\tilde{Q} \textit{ and } \tilde{Y} \textit{ do not anticipate the future of } \tilde{X},
\end{aligned}$$

where $\tilde{X}(\cdot) = \tilde{E}(\cdot) + \sum_{j=1}^{\mathbf{J}} \tilde{\Phi}^j(\beta_j x_j^* \cdot) + (P' - C)\tilde{S}(x^* \cdot)$ *is an* $\mathbf{I}$*-dimensional driftless Brownian motion that starts from the origin. (Here the* j^{th} *component of* $\tilde{S}(x^* \cdot)$ *is* $\tilde{S}_j(x_j^* \cdot)$*.)*

5 Equivalent Workload Formulation and the Dual Program

Harrison et al. [7, 10] have shown that the BCP has an "equivalent workload formulation" in which the dimension of the problem is usually reduced (or at least not increased) by replacing the "queue length" state descriptor $\tilde{Q}$ by a "workload" process $\tilde{W} = \Lambda\tilde{Q}$, where Λ is a matrix whose rows are determined by certain optimal solutions of the dual linear program to LP. More precisely, suppose that $(y^1, z^1), \ldots, (y^{\tilde{\mathbf{L}}}, z^{\tilde{\mathbf{L}}})$ are the extreme points of the feasible set of solutions to the *dual program* (DP):

$$\text{maximize } y'\alpha \ \text{ s.t. } \ y'R \leq z'A, \ \ z'e \leq 1 \ \text{ and } \ z \geq 0,$$

that satisfy $(y^\ell)'\alpha = 1$, $\ell = 1, \ldots, \tilde{\mathbf{L}}$, i.e., the maximum in the dual program DP is achieved at these extreme points. By relabelling if necessary, suppose that $y^1, \ldots, y^{\mathbf{L}}$ is a maximal linearly independent set of vectors from $y^1, \ldots, y^{\tilde{\mathbf{L}}}$. Let Λ be the matrix whose rows are given by $y^1, \ldots, y^{\mathbf{L}}$. (Note that in general Λ is not unique, although its range space is unique.) For future reference, let Π be the $\mathbf{L} \times \mathbf{K}$ matrix whose rows are given by $z^1, \ldots, z^{\mathbf{L}}$. A key to the reduction of Harrison et al. [7, 10] is the fact that if

$$\mathcal{N} = \{\delta \in \mathbb{R}^{\mathbf{I}} : \delta = Rx,\ Ax = 0, x_N = 0\},$$

where x_N denotes the components of x indexed by the non-basic activities $j = \mathbf{B}+1, \ldots, \mathbf{J}$, then

$$\mathcal{N}^\perp = \text{span}\{y^1, \ldots, y^{\mathbf{L}}\},$$

where $\mathcal{N}^\perp$ denotes the orthogonal complement of $\mathcal{N}$ in $\mathbb{R}^{\mathbf{I}}$. Recall the decomposition of R and A given just before the definition of the BCP. It follows from complementary slackness that $\Lambda H = \Pi B$. Combining this with $(y^i)'R \leq (z^i)'A$, we see that $(y^i)'M \leq (z^i)'N$, $i = 1, \ldots, \mathbf{L}$. It then follows that the $\mathbf{L} \times (\mathbf{K}+\mathbf{J}-\mathbf{B})$ matrix $G = [\Pi \quad \Pi N - \Lambda M]$ has entirely non-negative entries and $\Lambda R = GK$, where K is the $(\mathbf{K}+\mathbf{J}-\mathbf{B}) \times \mathbf{J}$ matrix given by

$$K = \begin{bmatrix} B & N \\ 0 & -I \end{bmatrix}.$$

Here, I is the $(\mathbf{J}-\mathbf{B}) \times (\mathbf{J}-\mathbf{B})$ identity matrix (not to be confused with the idletime process).

The BCP has the equivalent workload formulation described below, cf. [7, 10]. Here, the *workload process* $\tilde{W} = \Lambda Q$ is $\mathbf{L}$-dimensional, and the $(\mathbf{K}+\mathbf{J}-\mathbf{B})$-dimensional control process $\tilde{U}$ is related to the processes in the BCP by the relation

$$\tilde{U} = \begin{bmatrix} \tilde{I} \\ -\tilde{Y}_N \end{bmatrix},$$

where $\tilde{Y}_N$ is the vector process formed from $\tilde{Y}_j$, $j = \mathbf{B}+1, \ldots \mathbf{J}$, the $\mathbf{J}-\mathbf{B}$ non-basic components of $\tilde{Y}$.

Equivalent Workload Formulation (EWF).

$$\text{minimize } \mathbf{E}\left(\int_0^\infty e^{-\gamma t}\, h'\tilde{Q}(t)\, dt\right)$$

using a $(\mathbf{K}+\mathbf{J}-\mathbf{B})$*-dimensional control process* $\tilde{U}$ *such that*

$$\begin{aligned}
&\tilde{W}(t) = \tilde{W}(0) + \Lambda\tilde{X}(t) + G\tilde{U}(t) \quad \textit{for all } t \geq 0,\\
&\tilde{W}(t) = \Lambda\tilde{Q}(t) \textit{ for all } t \geq 0,\\
&\tilde{Q}_i(t) \geq 0 \textit{ for all } t \geq 0,\ i = 1, \ldots, \mathbf{I},\\
&\tilde{U}_j \textit{ is non-decreasing and } \tilde{U}_j(0) = 0,\\
&\quad \textit{for } j = 1, \ldots, \mathbf{K}+\mathbf{J}-\mathbf{B},\\
&\tilde{U}(t) \textit{ takes values in the Euclidean space spanned}\\
&\quad \textit{by the columns of } K, \textit{ for all } t \geq 0,\\
&\tilde{Q} \textit{ and } \tilde{U} \textit{ do not anticipate the future of } \tilde{X},
\end{aligned}$$

where $\tilde{X}$ *is the same as in the definition of the BCP.*

6 Properties of the Workload Process

In this section we present some new results concerning properties of the workload process $\tilde{W}$. The proofs of these results are contained in Bramson-Williams [4].

In the examples that have been analyzed to date, the EWF has played a key role in solving the BCP. The following result characterizes the dimension of the workload process in terms of basic model parameters. This is likely to be of use in solving the EWF in general.

Theorem 6.1 *The dimension* $\mathbf{L}$ *of the workload process* $\tilde{W}$ *is given by*

$$\mathbf{L} = \mathbf{I} + \mathbf{K} - \mathbf{B}.$$

Remark. For the special case of $\mathbf{L} = 1$ and $P = 0$ (i.e., a parallel server system with complete resource pooling), this result was proved previously in Harrison-López [9].

The next result justifies the use of the term "workload", which we think of as a non-negative quantity. In those cases where the EWF has been solved, the non-negativity of the components of the workload process $\tilde{W} = \Lambda\tilde{Q}$ has played an important role in finding solutions.

Theorem 6.2 *Suppose that for each buffer* i*, there is a basic activity which serves that buffer. Then, the entries in the matrix* Λ *(whose rows are the linearly independent vectors* $y^1, \ldots, y^{\mathbf{L}}$ *coming from extremal optimal solutions of the dual program DP) are all non-negative.*

Remark. This result is not contradicted by the example in Section 6 of Harrison [7], because Harrison's setup allows an activity to serve more than one buffer (a possibility that is not included in our model).

References

[1] S. L. Bell and R. J. Williams, Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: asymptotic optimality of a threshold policy, submitted to *Ann. Appl. Prob.,* 1999.

[2] S. L. Bell and R. J. Williams, Dynamic scheduling of a parallel server system with complete resource pooling: asymptotic optimality of a threshold policy in heavy traffic, in preparation.

[3] M. Bramson, State space collapse with application to heavy traffic limits for multiclass queueing networks, *Queueing Systems,* **30** (1998), 89–148.

[4] M. Bramson and R. J. Williams, On dynamic scheduling of stochastic networks in heavy traffic, preprint.

[5] J. M. Harrison, Brownian models of queueing networks with heterogeneous customer populations, in *Stochastic Differential Systems, Stochastic Control Theory and Their Applications,* W. Fleming and P.L. Lions (eds.), Springer, 1988, pp. 147-186.

[6] J. M. Harrison, The BIGSTEP approach to flow management in stochastic processing networks, in F. P. Kelly et al. (eds.), *Stochastic Networks: Theory and Applications,* Oxford Univ. Press, 1996, pp. 57–90.

[7] J. M. Harrison, Brownian models of open processing networks: canonical representation of workload, *Ann. Appl. Prob.,* **10** (2000), 75–103.

[8] J. M. Harrison, Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies, *Ann. Appl. Prob.,* **8** (1998), 822-848.

[9] J. M. Harrison and M. J. López, Heavy traffic resource pooling in parallel-server systems, to appear in *Queueing Systems.*

[10] J. M. Harrison and J. A. Van Mieghem, Dynamic control of Brownian networks: state space collapse and equivalent workload formulations, *Ann. Appl. Prob.,* **7** (1997), 747–771.

[11] F. P. Kelly and C. N. Laws, Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling, *Queueing Systems,* **13** (1993), 47–86.

[12] S. Kumar, Two-server closed networks in heavy traffic: diffusion limits and asymptotic optimality, to appear in *Ann. Appl. Prob.*

[13] H. J. Kushner and Y. N. Chen, Optimal control of assignment of jobs to processors under heavy traffic, to appear in *Stochastics.*

[14] H. J. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time,* Springer, 1992.

[15] H. J. Kushner and L. F. Martins, Heavy traffic analysis of a controlled multiclass queueing network via weak convergence methods. *SIAM J. Control and Optimization,* **34** (1996), 1781–1797.

[16] C. N. Laws, Resource pooling in queueing networks with dynamic routing, *Adv. Appl. Prob.,* **24** (1992), 699–726.

[17] L. F. Martins, S. E. Shreve and H. M. Soner, Heavy traffic convergence of a controlled, multi-class queueing system, *SIAM J. Control and Optimization,* **34** (1996), 2133–2171.

[18] R. J. Williams, Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse, *Queueing Systems,* **30** (1998), 27–88.

[19] R. J. Williams, On dynamic scheduling of a parallel server system with complete resource pooling, in *Analysis of Communication Networks: Call Centres, Traffic and Performance,* D. R. McDonald and S. R. E. Turner (eds.), Amer. Math. Soc. 2000.