PROBABILISTIC SERVICE LEVEL GUARANTEES IN MAKE-TO-STOCK MANUFACTURING SYSTEMS

DIMITRIS BERTSIMAS

Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, dbertsim@mit.edu

IOANNIS CH. PASCHALIDIS

Department of Manufacturing Engineering, Boston University, Boston, Massachusetts 02215, yannisp@bu.edu, URL: http://ionia.bu.edu

(Received February 1999; revision received September 1999; accepted September 1999)

We consider a model of a multiclass make-to-stock manufacturing system. External demand for each product class is met from the available finished goods inventory; unsatisfied demand is backlogged. The objective is to devise a production policy that minimizes inventory costs subject to guaranteeing stockout probabilities to stay bounded above by given constants ϵ_j, for each product class j (*service level guarantees*). Such a policy determines whether the facility should be producing (*idling decisions*), and if it should, which product class (*sequencing decisions*). Approximating the original system, we analyze a corresponding *fluid model* to make sequencing decisions and employ *large deviations techniques* to make idling ones. We consider both linear and quadratic inventory cost structures to obtain a *priority-based* and a *generalized longest queue first-based* production policy, respectively. An important feature of our model is that it accommodates autocorrelated demand and service processes, both critical features of modern failure-prone manufacturing systems.

1. INTRODUCTION

Make-to-stock manufacturing is the norm for a very large variety of industries. In such systems, demand is met from a *finished goods inventory* (FGI) and the production facility strives to maintain this inventory nonempty to avoid stockouts which lead either to backordered demand or simply lost sales. Most of the retail products, cars, appliances, silicon chips, and jet engines, are just a few examples in the huge set of products manufactured in this fashion. In make-to-stock manufacturing the fundamental trade-off is between *producing,* which accumulates inventory and incurs inventory costs, and *idling,* which leads to stockouts and unsatisfied demand. In *multiclass* systems, where a single facility produces several products, an additional control action is *sequencing* or *scheduling*, that is, what product to produce, if any. We will refer to the set of rules that determine both (a) sequencing decisions, and (b) idling decisions, as *production policy*. The objective is to devise a production policy which optimizes some measure of the system's performance. This measure should incorporate both *inventory costs* and *backorder costs*, i.e., costs associated with not being able to meet demand at the time it arrives.

The single-class version of the problem has been studied extensively in the literature (Evans 1967, Sobel 1982, Gavish and Graves 1980, Federgruen and Zipkin 1986). (For a more extensive literature review see Kapuscinski and Tayur 1999.) In an M/M/1 setting it has been established that a *base-stock policy* (produce when inventory falls below a cetain threshold and idle otherwise) is optimal (Gavish and Graves 1980, Sobel 1982). The same is true for renewal demand and deterministic production capacity (Federgruen and Zipkin 1986). In Akella and Kumar (1986) the optimality of a similar policy has been established when the demand is deterministic and the production capacity modulated by a two-state Markov chain. In Zheng and Zipkin (1990) the M/M/1 version of a two-class system is analyzed under the longest queue first policy. In Wein (1992) the multiclass case is analyzed in heavy traffic and an idling policy is proposed where production stops when a weighted sum of the expected backlogs for different classes exceeds a certain threshold value. When the facility should be working it uses a scheduling policy which resembles one of the policies (the priority-based policy) we will propose in this paper. The priority-based policy is further justified by arguments in Peña Perez and Zipkin (1997) under some cost assumptions ("weakly-cost-ordered" model). This latter paper also proposes a heuristic sequencing policy coupled with a base-stock idling policy. Additional heuristic policies are compared in Veatch and Wein (1996) and a heuristic sequencing policy coupled with a heavy-traffic-based idling policy is proposed. In Ha (1997) the optimality of a priority policy is established in a two-class system, and monotone switching curves are numerically obtained. A more detailed, but still partial, characterization of the optimal policy in some two-class systems is given in de Véricourt, Karaesmen and Dallery (1998). The hedging point we propose in this paper is in accordance with the result in de Véricourt, Karaesmen and Dallery (1998). Finally, a static allocation policy is considered in Glasserman (1996)

Subject classifications: Inventory/Production: multi-item, make-to-stock systems. Queues, approximations: large deviations, fluid models.
Area of review: MANUFACTURING OPERATIONS.

0030-364X/01/4901-0119 $05.00
1526-5463 electronic ISSN

Operations Research © 2001 INFORMS
Vol. 49, No. 1, January–February 2001, pp. 119–133

for the multiclass system and asymptotically analyzed for renewal arrivals and services by essentially decomposing the system to single-class systems. In the latter paper, the author considers probabilistic service level constraints and uses asymptotics which are similar in spirit to ours but apply only under renewal assumptions (for related work see also Glasserman and Wang 1998 and Glasserman 1997).

Our main contributions in this paper are:

(1) *Probabilistic constraints to capture Quality of Service* (QoS). Most of the work in the literature considers minimizing expected *linear* inventory and backorder costs. In practice, and at least for big manufacturing facilities which maintain relatively large quantities of finished goods inventory, linear inventory costs are a good approximation of reality and data to estimate the slope of the cost function are readily available. However, the same cannot be said about backorder costs. The assumption of linear backorder costs, that is often made in the literature, appears to serve analytical tractability rather than an accurate representation of reality. Moreover and more importantly, it is hard to obtain data which will help quantify customer satisfaction via linear backorder costs. To address this need, we introduce constraints that ensure that probabilities of stockout for different products stay bounded below given desirable levels. We believe that such constraints result in a more natural representation of customer satisfaction. Thus, we formulate the performance objective as: *minimize expected inventory costs subject to stockout probabilities being bounded above by given constants*.

(2) *Dependencies in demand and service processes.* In practice, demand for various products might have strong correlations with a variety of phenomena such as: sales events (e.g., sales and discount events increase demand while they last), weather patterns (e.g., severe rains can increase the demand for umbrellas), state of the economy (e.g., economic prosperity leads to more demand for consumer products), technological advances, demand for other products, etc. An additional complication is that manufacturing facilities are *stochastic* and *failure-prone*. In particular, when a machine breaks down, it is very likely to stay down for some period of time, which disrupts the output flow and creates dependencies. To accommodate such phenomena, our analysis will allow demand and production to be modeled by *autocorrelated* stochastic processes. We will demonstrate that such *distributional information* (vs. knowledge of the first two moments only) on demand and service processes is critical in optimizing the performance of the make-to-stock system.

(3) *Fluid and large deviations techniques.* On the methodological side, we combine recently developed techniques in *fluid models* and *large deviations.* We decompose the derivation of the proposed production policy in two parts: (a) sequencing or scheduling decisions, and (b) idling decisions. For part (a) we ignore stochasticities and consider a fluid model of the problem. Our motivation is that fluid models have been shown to provide good scheduling policies in a variety of settings (see for example Avram, Bertsimas and Ricard 1995 and Meyn 1996). For part (b) we employ large deviations techniques to obtain provably tight asymptotics on the stockout probabilities.

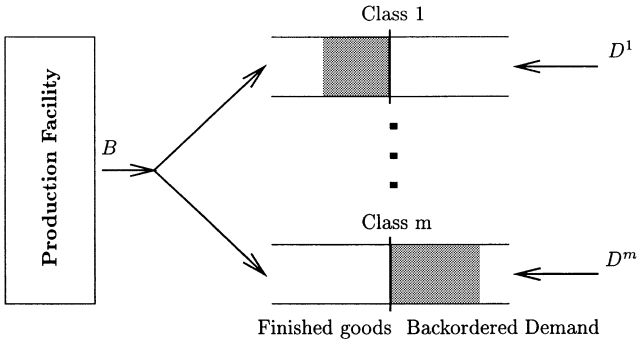
The remainder of this paper is organized as follows: In §2 we provide a detailed model of the make-to-stock manufacturing system we will analyze. In §3 we summarize background material on large deviations. In §4, and to motivate the subsequent analysis, we consider the simpler single-class case and derive the production policy using our large deviations asymptotics. We show that this policy is optimal in the M/M/1 case. In §5 we introduce a fluid model for the problem and solve it to obtain the structure of the scheduling policy under two different cost functionals. In §6 we utilize the structure obtained from the fluid analysis and depending on inventory cost assumptions we propose two distinct production policies: a *priority-based policy* and a *generalized longest queue first-based* (GLQF) policy. In both policies idling depends on a *hedging point* or vector of *safety stocks.* We analyze these policies in the large deviations regime in §7 and use the latter analysis in §8 to select an appropriate hedging point. This hedging point and inventory cost assumptions completely determine the operation of the GLQF-based production policy. This is not the case with the priority policy: the optimal priority ordering is determined in §9 so that total expected inventory cost is minimized. To that end, large deviations asymptotics are again employed. Finally, in §10 we present numerical results to assess the performance of the proposed policies. Concluding remarks are in §11.

2. THE MODEL

We consider the *multiclass make-to-stock manufacturing system* depicted in Figure 1. The production facility produces m products to be stocked in finished goods inventory. Demand is met from the available finished goods inventory, and it is backordered if inventory is not available. We assume a periodic review policy where time is divided into time slots of equal duration and the system is examined at the beginning of each time slot. We let D_i^j denote the amount of class j orders arriving during time slot i, for i in the set of integers $\mathbb{Z}$ and $j = 1, \ldots, m$. We will be measuring D_i^j in production time units when the facility is working at a production rate of 1. That is, D_i^j is equal to the time that the production facility requires to produce, at a production rate of 1, the amount of class j orders arriving during time slot i. We let also B_i denote the amount of work, measured in the same production time units, that the production facility can complete during time slot i. Finally, let x_i^j denote the class j inventory, measured in the same production time units, which is available at the beginning of time slot i (without taking into account D_i^j and B_i). We allow the inventory to take negative values to denote backordering; when x_i^j is negative $-x_i^j$ is equal to the amount of work backordered from class j. We will be using the notation $\mathbf{x}_i = (x_i^1, \ldots, x_i^m)$.

All the demand processes $\{D_i^j; i \in \mathbb{Z}, j = 1, \ldots, m\}$ and the production process $\{B_i; i \in \mathbb{Z}\}$ are arbitrary stationary

Figure 1. A multiclass make-to-stock manufacturing system.



stochastic processes that satisfy certain mild technical conditions. These conditions are satisfied by renewal processes, Markov-modulated processes, and, in general, stationary processes with mild mixing conditions (for details see Bertsimas, Paschalidis and Tsitsiklis 1998a, 1999). For stability purposes we assume that

$$\sum_{j=1}^{m} \mathbf{E}[D_1^j] < \mathbf{E}[B_1], \tag{1}$$

which by stationarity carries over to all time slots i.

As discussed in the Introduction, our objective is to devise a production policy that minimizes finished goods inventory costs and guarantees that the steady-state stockout probabilities $\mathbf{P}[x_i^j \leqslant 0]$ do not exceed some desired small values ϵ_j, for each class j. We will refer to these latter constraints as *service level* constraints. Of course, with such constraints we can only control the fraction of stockouts. This does not penalize a policy where some customers end up waiting much longer than others when backlogged, and can lead to unfair practices (see Spearman and Zhang 1999). Nevertheless, the production policies we will propose are "fair" in that they do not discriminate between customers of the same class.

Outline of Our Approach

We next briefly outline the approach we will use to achieve the objective outlined above. We will first ignore stochasticities and consider a deterministic version of the problem for which we will obtain an optimal sequencing policy. For the purposes of this deterministic version of the model we will assume that the backlog at time slot i incurs cost at a rate of $\sum_{j=1}^{m} f_j(x_i^j)$ per time slot, where the cost function $f_j(x_i^j)$ can have one of the following two forms (*linear* or *quadratic*):

$$f_j^L(x_i^j) = \begin{cases} h_j x_i^j, & x_i^j \geqslant 0, \\ b_j |x_i^j|, & x_i^j \leqslant 0, \end{cases} \tag{2}$$

or

$$f_j^Q(x_i^j) = c_j (x_i^j)^2, \tag{3}$$

where $h_j, b_j, c_j, j = 1, \ldots, m$, are nonnegative constants. We will obtain optimal sequencing policies that minimize

$$\sum_{i=1}^{T} \sum_{j=1}^{m} f_j(x_i^j),$$

under both cost assumptions, where T is the time horizon of interest. To that end, we will introduce a continuous-time fluid model (i.e., the continuous-time analog of the deterministic version of the problem) and use calculus of variations techniques. In view of our earlier comments in the Introduction on the linear backorder cost assumptions, we note that we make the cost assumption in (2) just for the purposes of deriving the structure of the sequencing policy.

The fluid model evolves in continuous time and ignores stochasticities in the demand and service processes which lead to stockouts. To accommodate these effects we will consider a deterministic analog of the optimal "fluid" sequencing policies and enhance them with an appropriate idling policy. For relatively large inventories, or equivalently for small stockout probabilities, stockouts are rare events; hence, we will use large deviations theory to analyze the resulting production policies and tune the parameters that characterize the corresponding idling policy. We will finally provide evidence (analytical and numerical) that the large deviations asymptotics are *relevant*; i.e., the analytically calculated parameters of the proposed production policies are very close to simulated values.

3. BACKGROUND MATERIAL ON LARGE DEVIATIONS

Before we proceed with our agenda we first review some basic results, which will also help in establishing some of our notation. Consider a sequence of i.i.d. random variables $X_i, i \geqslant 1$, with mean $\mathbf{E}[X_1] = \overline{X}$. The strong law of large numbers asserts that $(\sum_{i=1}^{n} X_i)/n$ converges to $\overline{X}$, as $n \to \infty$, w.p.1. Thus, for large n the event $\sum_{i=1}^{n} X_i > na$, where $a > \overline{X}$ (or $\sum_{i=1}^{n} X_i < na$, for $a < \overline{X}$) is a rare event. More specifically, its probability behaves as $e^{-nr(a)}$, as $n \to \infty$, where the function $r(\cdot)$ determines the rate at which the probability of this event is diminishing. Cramér's theorem (Cramér 1938) determines $r(\cdot)$, and is considered the first Large Deviations statement. In particular,

$$r(a) = \sup_{\theta} \big(\theta a - \log \mathbf{E}[e^{\theta X_1}]\big).$$

Consider next a sequence $\{S_1, S_2, \ldots\}$ of random variables, with values in $\mathbb{R}$ and define

$$\Lambda_n(\theta) \stackrel{\Delta}{=} \frac{1}{n} \log \mathbf{E}[e^{\theta S_n}]. \tag{4}$$

For the applications we have in mind, S_n is a partial sum process. Namely, $S_n = \sum_{i=1}^{n} X_i$, where $X_i, i \geqslant 1$, are identically distributed, possibly dependent, random variables. Let

$$\Lambda(\theta) \stackrel{\Delta}{=} \lim_{n\to\infty} \Lambda_n(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{\theta S_n}]. \tag{5}$$

(We assume that the limit exists for all θ, where $\pm\infty$ are allowed both as elements of the sequence $\Lambda_n(\theta)$ and as limit points.) We will refer to $\Lambda(\cdot)$ as the *limiting* log-*moment generating function*. Let us also define

$$\Lambda^*(a) \stackrel{\Delta}{=} \sup_{\theta}(\theta a - \Lambda(\theta)), \tag{6}$$

which will be referred to as the *large deviation rate function.* Under a technical assumption (see Dembo and Zeitouni 1993) it has been established (Gärtner-Ellis Theorem) that for large enough n and for small $\epsilon > 0$,

$$\mathbf{P}[S_n \in (na - n\epsilon, na + n\epsilon)] \sim e^{-n\Lambda^*(a)}.$$

This can be viewed as an extension of Cramér's theorem to autocorrelated stochastic processes. We say that $\{S_n\}$ satisfies a *Large Deviations Principle* (LDP) with *rate function* $\Lambda^*(\cdot)$. The notation "$\sim$" should be interpreted as "asymptotically behaves"; more rigorously, the logarithm of the probability divided by n converges to $-\Lambda^*(a)$, as $n \to \infty$. (For a more rigorous statement of the Gärtner-Ellis theorem see Dembo and Zeitouni (1993).)

In the sequel, we will be denoting by $\Lambda_X(\cdot)$ and $\Lambda_X^*(\cdot)$ the limiting log-moment generating function and the large deviations rate function, respectively, of the process X.

4. THE SINGLE CLASS CASE

To motivate the subsequent multiclass analysis, in this section we focus on the simpler single-class case. In particular, we assume that the manufacturing facility produces only a single product with demand $\{D_i; i \in \mathbb{Z}\}$. The modeling assumptions are as in §2. In this case, there are no scheduling decisions to be made. Thus, we are interested in an *idling policy* which guarantees that the steady-state stockout probability $\mathbf{P}[x_i \leqslant 0]$ does not exceed a desired small value ϵ.

For the single-class system it has been shown in a variety of settings (Evans 1967, Sobel 1982, Gavish and Graves 1980, and Federgruen and Zipkin 1986) that a so-called *base stock* policy is optimal. According to such a policy, the system produces if the inventory x_i falls below a certain threshold value w, and idles otherwise. We will refer to w as the *hedging point* or *safety stock*. Hence, our objective is to determine w to satisfy

$$\mathbf{P}[x_i \leqslant 0] \leqslant \epsilon. \tag{7}$$

We first observe that under the above base stock policy the make-to-stock system can be transformed to an equivalent make-to-order system. In particular, consider the transformation

$$L_i = w - x_i, \tag{8}$$

and note that L_i can be interpreted as the queue length during time slot i in a discrete-time $G/G/1$ queue with D_i arrivals and at most B_i services during time slot i. Consequently, the stockout probability in the make-to-stock system is equal to the overflow probability in the corresponding make-to-order system; i.e.,

$$\mathbf{P}[x_i \leqslant 0] = \mathbf{P}[L_i \geqslant w], \tag{9}$$

and thus it suffices to bound the latter one with ϵ.

Calculating the above overflow probability exactly is particularly hard in view of the complicated stochastic nature of the arrival and service process (processes with dependencies). To that end, we will use *large deviations* asymptotics. Under very general assumptions that include renewal processes, Markov-modulated processes, and in general arbitrary stationary processes with mild mixing conditions (see Dembo and Zeitouni 1993), the following proposition has been established (Glynn and Whitt 1994, Bertsimas, Paschalidis and Tsitsiklis 1998b).

In preparation for the result, consider a convex function $g(u)$ with the property $g(0) = 0$. We define the *largest root* of $g(u)$ to be the solution of the optimization problem $\sup_{u:g(u)<0} u$. If $g(\cdot)$ has negative derivative at $u = 0$, there are two cases: either $g(\cdot)$ has a single positive root or it stays below the horizontal axis $u = 0$, for all $u > 0$. In the latter case, we will say that $g(\cdot)$ has a root at $u = \infty$.

PROPOSITION 4.1 (SINGLE CLASS). *The steady-state queue length process L_i satisfies*

$$\lim_{w\to\infty} \frac{1}{w} \log \mathbf{P}[L_i \geqslant w] = -\theta^*, \tag{10}$$

where $\theta^* > 0$ is the largest root of the equation

$$\Lambda_D(\theta) + \Lambda_B(-\theta) = 0. \tag{11}$$

More intuitively, for large enough w we have

$$\mathbf{P}[x_i \leqslant 0] = \mathbf{P}[L_i \geqslant w] \sim e^{-w\theta^*}.$$

Thus, approximating the stockout probability with the right-hand side of the above, we conclude that the minimum value of w (hedging point) guaranteeing (7) is given by

$$w = -\frac{\log \epsilon}{\theta^*}. \tag{12}$$

If $\theta^* = \infty$, then no stockouts occur and a hedging point equal to zero should be used (*Just in Time* (JIT) policy).

The *M/M/1* Case

To assess the accuracy of the asymptotics in Proposition 4.1 we next consider the $M/M/1$ case. We assume that demand for products arrives according to a Poisson process with rate λ. To produce a single product requires an exponentially distributed amount of time with parameter μ.

To cast these assumptions into our discrete-time model we let the duration of a time slot be equal to δ; we will later

take the limit as $\delta \to 0$. According to our modeling conventions (i.e., measuring demand in time units), the demand during time slot i is given by

$$D_i = \begin{cases} Y, & \text{with probability } \lambda\delta, \\ 0, & \text{with probability } 1-\lambda\delta, \end{cases} \tag{13}$$

where Y is exponentially distributed with parameter μ. Assuming that the service facility works at a production rate of 1, the service process during time slot i is characterized by

$$B_i = \delta. \tag{14}$$

To apply the result of Proposition 4.1 we calculate the *log*-moment generating functions of the demand and service processes. It particular,

$$\begin{aligned} \Lambda_D(\theta) &= \log \mathbf{E}[e^{\theta D_i}] \\ &= \log\left(\lambda\delta \frac{\mu}{\mu-\theta} + (1-\lambda\delta)\right), \quad \text{for } \theta < \mu, \end{aligned}$$

and

$$\Lambda_B(\theta) = \delta\theta.$$

We now take $\delta \to 0$ and solve Equation (11) ignoring $O(\delta^2)$ terms to obtain that the largest positive root is $\theta^* = \mu - \lambda$. Hence, using Equation (9),

$$\mathbf{P}[x_i \leqslant 0] = \mathbf{P}[L_i \geqslant w] \sim e^{-w(\mu-\lambda)}. \tag{15}$$

The important observation is that the expression above is *exact*, and thus it leads to an exact calculation of the hedging point. To see that, notice that the system time in an $M/M/1$ queue is exponentially distributed with parameter $\mu - \lambda$. In summary, we established that in the single class $M/M/1$ case large deviations asymptotics lead to an exact derivation of the optimal idling policy.

5. A FLUID CONTROL PROBLEM

As discussed in §2, in order to make sequencing decisions we consider a fluid model of the problem: time becomes continuous and orders as well as finished goods flow into the inventory as continuous fluids. Let $x_j(t)$ denote the class j backlog for the fluid model. We will write $\mathbf{x}(t) = (x_1(t), \ldots, x_m(t))$ for the vector of the backlogs. Let also $u_j(t), T_j(t)$ be the instantaneous and cumulative service effort allocated to class j at time t respectively. Clearly,

$$T_j(t) = \int_0^t u_j(s)\, ds.$$

We will be using the notation $\mathbf{u}(t) = (u_1(t), \ldots, u_m(t))$ and $\mathbf{T}(t) = (T_1(t), \ldots, T_m(t))$. Let finally $\mathbf{d} = (d_1, \ldots, d_m) \stackrel{\Delta}{=} (\mathbf{E}[D_1^1], \ldots, \mathbf{E}[D_1^m])$. Then the fluid control problem becomes:

$$\begin{aligned} \text{minimize} \quad & \int_0^T \sum_{j=1}^m f_j(x_j(t))\, dt \\ \text{subject to} \quad & \dot{\mathbf{x}}(t) = \mathbf{u}(t) - \mathbf{d}, \\ & \mathbf{e}'\mathbf{u}(t) \leqslant \mathbf{E}[B_1], \\ & \mathbf{x}(0): \text{ given}, \\ & \mathbf{u}(t) \geqslant \mathbf{0}, \end{aligned} \tag{16}$$

where $\mathbf{e}$ is the vector of all ones, $\mathbf{0}$ the vector of all zeros, and prime denotes transpose. Changing variables, the problem becomes:

$$\begin{aligned} \text{minimize} \quad & \int_0^T \sum_{j=1}^m f_j(x_j(0) + T_j(t) - d_j t)\, dt \\ \text{subject to} \quad & \mathbf{e}'\mathbf{T}(t) \leqslant \mathbf{E}[B_1]t, \\ & T_j(t): \text{ nondecreasing for all } j. \end{aligned} \tag{17}$$

Suppose we have found the optimal policy until time s, where $s \in [0, T)$. In the next proposition we derive the structure of the optimal policy under the linear and quadratic cost assumptions of (2) and (3) in the infinitesimal time interval $[s, s+\delta]$, where $\delta \to 0$. We assume that the initial state $\mathbf{x}(s)$ for this interval is given and that $x_j(s) \neq 0$ for all j.

PROPOSITION 5.1. *Let $u_j^*(t)$ and $x_j^*(t)$ denote the optimal control and state trajectories, respectively, for all $j = 1, \ldots, m$ and $t \in [s, s+\delta]$.*

(1) (LINEAR COST). *Assume $f_j(x_j(t)) = f_j^L(x_j(t))$, for all $j = 1, \ldots, m$ and $t \in [s, s+\delta]$, and that $b_j \neq b_{j'}$ for all $j \neq j'$. Then the optimal policy is as follows: If all $x_j(s) > 0$, then idle (i.e., $u_j^*(t) = 0$ for all j and $t \in [s, s+\delta]$). Otherwise, among the classes that have negative $x_j(s)$, work on the class that has the highest index b_j (i.e., $j^* = \arg\max_{\{j|x_j(s)<0\}} b_j$, $u_{j^*}^*(t) = \mathbf{E}[B_1]$, and $u_j^*(t) = 0$ for all $j \neq j^*$ and $t \in [s, s+\delta]$).*

(2) (QUADRATIC COST). *Assume $f_j(x_j(t)) = f_j^Q(x_j(t))$, for all $j = 1, \ldots, m$ and $t \in [s, s+\delta]$, and that $c_j|x_j(s)| \neq c_{j'}|x_{j'}(s)|$ for all $j \neq j'$. Then the optimal policy is as follows: If all $x_j(s) > 0$, then idle (i.e., $u_j^*(t) = 0$ for all j and $t \in [s, s+\delta]$). Otherwise, among the classes that have negative $x_j(s)$, work on the class that has the highest index $c_j|x_j(s)|$ (i.e., $j^* = \arg\max_{\{j|x_j(s)<0\}} c_j|x_j(s)|$, $u_{j^*}^*(t) = \mathbf{E}[B_1]$, and $u_j^*(t) = 0$ for all $j \neq j^*$ and $t \in [s, s+\delta]$).*

PROOF. We will apply standard calculus of variations techniques (see Bertsekas 1995, Chapter 3] on Pontryagin's minimum principle). The Hamiltonian is given by

$$H(\mathbf{x}, \mathbf{u}, \mathbf{p}) = \sum_{j=1}^m f_j(x_j) + \sum_{j=1}^m p_j(u_j - d_j),$$

where $\mathbf{x}, \mathbf{u}, \mathbf{p} \in \mathbb{R}^m$. The optimal state trajectory in $[s, s+\delta]$ satisfies

$$\dot{\mathbf{x}}^*(t) = \mathbf{u}^*(t) - \mathbf{d}, \qquad \mathbf{x}^*(\mathbf{s}) = \mathbf{x}(\mathbf{s}): \text{ given},$$

and $\mathbf{p}(t)$ satisfies the adjoint equation

$$\dot{\mathbf{p}}(t) = -\nabla_x H(\mathbf{x}^*(t), \mathbf{u}^*(t), \mathbf{p}(t)) = -\nabla_x \sum_{j=1}^{m} f_j(x_j^*(t)), \qquad \mathbf{p}(s+\delta) = 0.$$

The optimal control trajectory is given by

$$\mathbf{u}^*(t) = \arg\min_{\mathbf{u}\in\mathcal{U}} H(\mathbf{x}^*(t), \mathbf{u}, \mathbf{p}(t)),$$

where $\mathcal{U} = \{\mathbf{u} \geqslant \mathbf{0} \mid \mathbf{e}'\mathbf{u} \leqslant \mathbf{E}[B_1]\}$. Since we are dealing with a convex objective function, linear state dynamics, and we are optimizing over a convex set $\mathcal{U}$, Pontryagin's minimum principle provides necessary and sufficient conditions for optimality. From the structure of the Hamiltonian it can be seen that

$$u_j^*(t) = \begin{cases} 0, & \text{if } p_j(t) > 0 \ \forall j, \\ \mathbf{E}[B_1], & \text{if } \exists k \text{ with } p_k(t) < 0, \\ & \quad j^* = \arg\min_j \{p_j(t)\}, \text{ and } j = j^*, \\ 0, & \text{if } \exists k \text{ with } p_k(t) < 0, \\ & \quad j^* = \arg\min_j \{p_j(t)\}, \text{ and } j \neq j^*. \end{cases}$$

Let us next focus on the case of linear cost. The adjoint equation takes the form

$$\dot{p}_j(t) = \begin{cases} -h_j, & \text{if } x_j^*(t) > 0, \\ b_j, & \text{if } x_j^*(t) < 0, \end{cases}$$

with terminal condition $p_j(s+\delta) = 0$. We can distinguish two cases:

(1) $x_j(s) > 0$ for all j. Then for all j and $t \in [s, s+\delta]$ we have $p_j(t) = -h_j(t-s-\delta)$ and $u_j^*(t) = 0$.

(2) There exists at least one negative $x_j(s)$ and $j^* = \arg\max_{\{j|x_j(s)<0\}} b_j$. Then for all j with $x_j(s) < 0$ we have $p_j(t) = b_j(t-s-\delta)$. According to the optimality condition we work on the class with the most negative $p_j(t)$, which is j^* for all $t \in [s, s+\delta]$.

A similar argument holds for the quadratic cost case. The adjoint equation takes the form

$$\dot{p}_j(t) = -2c_j x_j^*(t),$$

with terminal condition $p_j(s+\delta) = 0$. Again, we have two cases:

(1) $x_j(s) > 0$ for all j. Then $p_j(t) > 0$ for all j and $t \in (s, s+\delta)$ and the optimal policy is to idle.

(2) There exists at least one negative $x_j(s)$ and $j^* = \arg\max_{\{j|x_j(s)<0\}} c_j|x_j(s)|$. Then for those j with $x_j(s) < 0$, $p_j(t)$ is quadratic in $[s, s+\delta]$ and the most negative is the one with the largest slope at $s+\delta$, which is the same as the one with the largest slope at s (since $\delta \to 0$), that is, j^*. The optimality condition implies that we should work on class j^*. □

The proposition above implies that the optimal policy at time s depends only on the state at that time and not explicitly on s. By discretizing time, using induction, and applying Bellman's principle of optimality, it can be shown that Proposition 5.1 characterizes the structure of the optimal policy in $[0, T]$. In particular, we work on the class j that has the highest index b_j (for the linear cost case) or $c_j|x_j(t)|$ (for the quadratic cost case) among those classes that have negative inventory; if all classes have positive inventory we idle.

Notice, that we have not spelled out here the detailed implementation of the optimal fluid policy, because this does not impact our later large deviations analysis. In particular, we have not specified what the optimal policy does if $x_j^*(t) = 0$, for some j, or if there is a tie in the definition of j^* in the statement of Proposition 5.1. It can be shown that in both these cases the server should split its capacity between classes. In particular, if $x_j^*(t) = 0$, for some j, the server should keep the backlog of class j at zero by allocating $u_j^*(t) = d_j$. Similarly, if there is a tie in the definition of j^* the server should split its capacity between all classes j with $j = j^*$ to keep them at the same level of backlog.

6. THE PROPOSED PRODUCTION POLICY

The analysis of the previous section led to optimal sequencing policies in the fluid model. However, since stochasticities are completely ignored, it does not provide any information on the idling policy. Motivated by the optimality results for the single class case (Evans 1967, Sobel 1982, Gavish and Graves 1980, and Federgruen and Zipkin 1986) we will enhance the "fluid" policy to *hedge* against stochasticity and focus on a *base stock* class of policies. In particular, we will consider the following idling policy that utilizes a so-called *hedging point* or *safety stock* $\mathbf{w} = (w_1, \ldots, w_m)$:

- idle during time slot i when $\mathbf{x}_i \geqslant \mathbf{w}$, and
- work on the classes j that satisfy $x_i^j < w_j$, without exceeding the corresponding safety stock w_j.

In the latter case, one of the sequencing policies determined by the optimal fluid policy will be followed, as listed, below:

PRIORITY-BASED POLICY: We define a fixed ordering $(\chi(1), \ldots, \chi(m))$ of the set of classes $\{1, \ldots, m\}$ and we work on the class j which has the highest rank $\chi(j)$ and satisfies $x_i^j < w_j$.

GENERALIZED LONGEST QUEUE FIRST-BASED POLICY (GLQF): We define scalars $c_1, \ldots, c_m$ and we work on the class j that maximizes $c_j(w_j - x_i^j)$, assuming that there exists at least one j that satisfies $x_i^j < w_j$.

Some comments on these policies are in order. The priority-based policy with ordering determined by the indices b_j is identical to one of the policies proposed in Peña Perez and Zipkin (1997). Under the weakly-cost-ordered assumption of Peña Perez and Zipkin (1997) (i.e., $\arg\min_j b_j = \arg\min_j h_j$) this policy is of the type proposed in Wein (1992) and is asymptotically optimal in heavy traffic (see Peña Perez and Zipkin 1997, §3] for a rigorous statement and an interesting discussion). Furthermore, in the two-class $M/M/1$ case the authors in de Véricourt et al.

(1998) prove the optimality of a priority policy in a part of the state space. The GLQF-based policy can be seen as a generalization of the longest-queue-first policy analyzed in a symmetric two-class $M/M/1$ case in Zheng and Zipkin (1990). If the objective is to minimize expected inventory and backorder costs, the proposed production policies are not optimal. The optimal policy is in fact unknown (expect in very limited special cases) and only heuristics (occasionally based on asymptotics) have been proposed to date. As outlined above, though, the objective we pursue in this paper is to minimize inventory costs subject to probabilistic service level guarantees. The fluid analysis narrows down the choices for sequencing policies; it provides two classes of interesting policies, which have optimality properties at least in the fluid limiting regime. We will later select a policy within each class (i.e., select the "best" priority policy and the appropriate GLQF policy) based on inventory cost considerations.

Given the two proposed production policies, we are interested in determining a hedging point $\mathbf{w}$ which guarantees that the steady-state stockout probabilities $\mathbf{P}[x_i^j \leqslant 0]$ do not exceed some desired small values ϵ_j, for each class j. As in the single-class case, the system can be transformed to an equivalent make-to-order one. More specifically, let $\mathbf{L}_i$ denote the vector $(L_i^1, \ldots, L_i^m)$ for each time slot i and consider the transformation

$$\mathbf{L}_i := \mathbf{w} - \mathbf{x}_i. \tag{18}$$

Note that under the proposed idling policy x_i^j is in the interval $(-\infty, w_j]$ for all time slots i and classes j. Hence, L_i^j is in the interval $[0, \infty)$, with zero corresponding to the hedging point. The vector $\mathbf{L}_i$ evolves exactly as the queue length vector of a discrete-time multiclass queue with dedicated buffers for each class, arrival processes $\{D_i^j; i \in \mathbb{Z}, j = 1, \ldots, m\}$, and service process $\{B_i; i \in \mathbb{Z}\}$. Therefore, based on this equivalence the stockout probability is equal to the overflow probability in the make-to-order system, i.e.,

$$\mathbf{P}[x_i^j \leqslant 0] = \mathbf{P}[L_i^j \geqslant w_j], \qquad i \in \mathbb{Z}, j = 1, \ldots, m. \tag{19}$$

Since the exact calculation of these probabilities is intractable we will resort to asymptotics. Recently, some new asymptotic techniques for calculating such tail probabilities have been developed by the authors (Paschalidis 1996, Bertsimas et al. 1998a, b, 1999) based on ideas from large deviations theory, optimal control, and optimization. We will extend these results to analytically approximate the stockout probabilities.

7. ASYMPTOTIC EXPRESSIONS FOR THE STOCKOUT PROBABILITIES

In this section we extend the results from Paschalidis (1996), Bertsimas et al. (1998a, 1999) to obtain asymptotic expressions for the stockout probability of each class. Based on the equivalence with the corresponding make-to-order system that we obtained in §6 it suffices to obtain asymptotics on the overflow probabilities in multiclass queues with dedicated buffers operating under a strict priority policy, and the GLQF policy (cf. Equation (19)).

The results in this section hold under fairly general assumptions on the stochastic processes involved (arrival and service processes). As we mentioned in §2, these assumptions are satisfied by renewal processes, Markov-modulated processes, and in general, stationary processes with mild mixing conditions (for details see Bertsimas et al. 1998a, 1999).

7.1. Priority Policy

Without loss of generality we reorder the classes so that strict priority is given to classes $1, \ldots, m$ in this order. The main result on the overflow probabilities of the corresponding make-to-order system is summarized in the following theorem.

THEOREM 7.1 (PRIORITY POLICY). *Under the policy which gives priority to class j over class j' for all $j < j'$, the steady-state queue length L^1 of class 1 satisfies*

$$\lim_{w_1 \to \infty} \frac{1}{w_1} \log \mathbf{P}[L^1 \geqslant w_1] = -\theta_{P_1}^*, \tag{20}$$

where $\theta_{P_1}^*$ is given by

$$\theta_{P_1}^* = \inf_{a>0} \frac{1}{a} \Lambda_1^*(a), \tag{21}$$

and where

$$\Lambda_1^*(a) \stackrel{\Delta}{=} \inf_{x_1 - y = a} [\Lambda_{D^1}^*(x_1) + \Lambda_B^*(y)]. \tag{22}$$

Moreover, the steady-state queue length L^j of class j, for $j = 2, \ldots, m$, satisfies

$$\lim_{w_j \to \infty} \frac{1}{w_j} \log \mathbf{P}[L^j \geqslant w_j] = -\theta_{P_j}^*, \tag{23}$$

where $\theta_{P_j}^*$ is given by

$$\theta_{P_j}^* = \inf_{a>0} \frac{1}{a} \Lambda_j^*(a), \tag{24}$$

and where

$$\Lambda_j^*(a) \stackrel{\Delta}{=} \inf_{\substack{x_j + z - y = a \\ \sum_{i=1}^{j-1} x_i = z \\ z \leqslant y}} \left[\Lambda_{D^j}^*(x_j) + \sum_{i=1}^{j-1} \Lambda_{D^i}^*(x_i) + \Lambda_B^*(y) \right]. \tag{25}$$

PROOF. The main technical result on which this theorem is based has been established in Bertsimas et al. (1999). Equation (20) is obtained by direct application of Corollary 7.1 in Bertsimas et al. (1999), which provides the large deviations exponent of the overflow probabilities in a two-class priority queue. For Equation (23) we group classes $1, \ldots, j-1$ into one superclass and use the result of Corollary 7.1 in Bertsimas et al. (1999) to obtain the tail of the queue length process for class j. Note that since the policy

of interest is a priority policy, the queue length of class j is not affected by classes $j+1,\ldots,m$ and by the scheduling policy used for classes $1,\ldots,j-1$, as long as it is a work-conserving one. Moreover, it is not hard to verify that the limiting log-moment generating function of the demand process for the superclass constructed by bundling together classes $1,\ldots,j-1$ is given by $\sum_{i=1}^{j-1}\Lambda_{D^i}(\theta)$. By standard convex analysis properties (Rockafellar 1970) the corresponding large deviations rate function is given by

$$\inf_{\sum_{i=1}^{j-1}x_1=z}\sum_{i=1}^{j-1}\Lambda^*_{D^i}(x_i).$$

Using the above expression in the result of Corollary 7.1 in Bertsimas et al. (1999) yields Equation (23). □

REMARKS.

(1) Intuitively, the result is that the overflow probability for class $j=1,2,\ldots,m$ (hence, due to (19) the stockout probability as well) satisfies

$$\mathbf{P}[x^j\leqslant 0]=\mathbf{P}[L^j\geqslant w_j]\sim e^{-w_j\theta^*_{P_j}}, \tag{26}$$

asymptotically, as $w_j\to\infty$. We will be referring to $\theta^*_{P_j}$ as the *asymptotic decay rate* of the stockout probability.

(2) To establish this result, in Bertsimas et al. (1999) we have combined techniques from large deviations and *optimal control.* In particular, we introduced an optimal control formulation that provides a tight (up to first degree in the exponent) lower bound on the overflow probability. A matching upper bound was proved via large deviations techniques, thus, establishing the result. To develop the lower bound we considered all scenarios (sample paths) that lead to an overflow. We showed that the probability of each of these scenarios ω asymptotically behaves as $e^{-w_j\theta(\omega)}$, for some function $\theta(\omega)$. This probability is a lower bound on $\mathbf{P}[x^j_i\leqslant 0]$ for all ω. We selected the tightest lower bound by minimizing $\theta(\omega)$ over all scenarios ω, which amounts to solving a deterministic optimal control problem. Optimal trajectories (paths) of the control problem correspond to *most likely* overflow scenarios. The interesting conclusion from this discussion is that the control problem formulation provides both the answer (the asymptotic decay rate of the overflow probability), along with an intuitive understanding of how the overflow occurs and how different classes interact.

(3) It is instructive to characterize these *most likely* overflow scenarios. The proof of the theorem suggests that it suffices to consider a two-class system. Figure 2 depicts the situation for the high-priority class. We denote by L^{high} and L^{low} the properly normalized queue lengths of the high- and low-priority classes, respectively. In this normalized space, overflow means reaching a value of 1, thus, a most likely scenario is one which drives an empty system $((L^{\text{high}},L^{\text{low}})=(0,0))$ to any point on the threshold line $L^{\text{high}}=1$ and has the largest probability among all such scenarios. It can be proven that in this scenario L^{high} grows linearly, and thus receives all the capacity. Consequently, the low-priority queue overflows as well (also linearly).

Figure 2. The most-likely overflow scenario for the high-priority class.

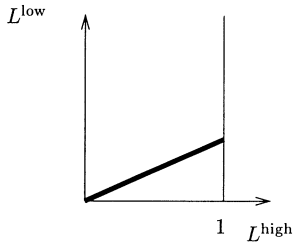
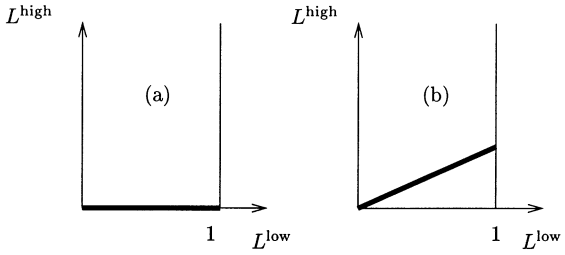


Figure 3. The most-likely overflow scenarios for the low-priority class.



The situation is different for the overflow of the low-priority class (see Figure 3). There are two most likely scenarios and which one occurs depends on the distributions of the arrival and service processes. Scenario (b) is similar to the one for the overflow of the high-priority class, that is, to have an overflow of the low-priority class both classes have to overflow. It is also possible however, that the high-priority class does not consume all the capacity but the residual capacity is not enough for the low-priority class. This latter case is depicted in Figure 3(a).

As the result indicates, the calculation of the overflow probabilities involves the solution of an optimization problem. The next theorem establishes that due to the special structure this optimization problem exhibits, it is equivalent to finding the maximum root of a convex function. Such a task might be easier to perform in some cases, analytically or computationally.

THEOREM 7.2. *$\theta^*_{P_1}$ is the largest positive root of the equation*

$$\Lambda_{D^1}(\theta)+\Lambda_B(-\theta)=0. \tag{27}$$

Also, $\theta^*_{P_j}$ is the largest positive root of the equation

$$\Lambda_{D^j}(\theta)+\inf_{0\leqslant u\leqslant\theta}\left[\sum_{i=1}^{j-1}\Lambda_{D^i}(\theta-u)+\Lambda_B(-\theta+u)\right]=0. \tag{28}$$

PROOF. We use the argument used in the proof of Theorem 7.1 and the result of Corollary 7.3 in Bertsimas et al. (1999), which provides the large deviations exponents of the overflow probabilities in a two-class priority queue as largest roots of nonlinear equations. □

7.2. The GLQF policy

We next turn our attention to the GLQF policy. Recall that, according to this policy, among all classes with backlogs below their hedging point, during time slot i, we work on the one that maximizes $c_j(w_j - x_i^j)$. Thus, due to (18), in the corresponding make-to-order system we serve the queue with maximum $c_j L_i^j$.

The following theorem is from Bertsimas et al. (1998a) and holds under the same general assumptions as Theorem 7.1. The theorem establishes asymptotics for two-class systems. The general case appears much harder and large deviations results are not available. One can use instead tight (but not asymptotically tight in the sense that there exists a lower bound with the same exponent) upper bounds on the overflow probabilities developed in Paschalidis (1998).

THEOREM 7.3 (GLQF POLICY). *Let $\beta = c_1/c_2$. Under the GLQF policy the steady-state queue length L^1 of class 1 satisfies*

$$\lim_{w_1\to\infty} \frac{1}{w_1} \log \mathbf{P}[L^1 > w_1] = -\theta^*_{GLQF_1}, \tag{29}$$

where $\theta^*_{GLQF_1}$ is given by

$$\theta^*_{GLQF_1} = \min\left[\inf_{a>0} \frac{1}{a}\Lambda^{I*}_{GLQF_1}(a), \inf_{a>0} \frac{1}{a}\Lambda^{II*}_{GLQF_1}(a)\right], \tag{30}$$

and the functions $\Lambda^{I*}_{GLQF_1}(\cdot)$ and $\Lambda^{II*}_{GLQF_1}(\cdot)$ are defined as follows

$$\Lambda^{I*}_{GLQF_1}(a) \stackrel{\Delta}{=} \inf_{\substack{x_1-x_3=a \\ x_2\leqslant\beta(x_1-x_3)}} [\Lambda^*_{D^1}(x_1) + \Lambda^*_{D^2}(x_2) + \Lambda^*_B(x_3)], \tag{31}$$

and

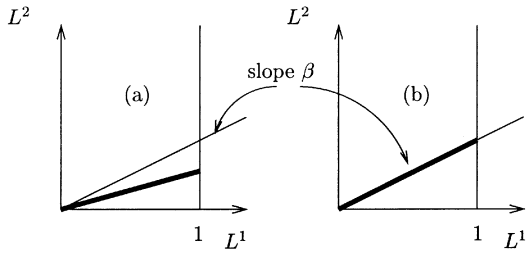
$$\Lambda^{II*}_{GLQF_1}(a) \stackrel{\Delta}{=} \inf_{\substack{x_1-\phi x_3=a \\ x_2-(1-\phi)x_3=\beta a \\ 0\leqslant\phi<1}} [\Lambda^*_{D^1}(x_1) + \Lambda^*_{D^2}(x_2) + \Lambda^*_B(x_3)]. \tag{32}$$

REMARKS.

(1) Our analysis also provides the *most likely* overflow scenarios, which we depict in Figure 4. In scenario (a) L^1 builds up linearly while staying larger than $(1/\beta)L^2$, which forces the server to attend only to L^1 during the busy period of overflow. In scenario (b), L^1 builds up also linearly but with $L^2 = \beta L^1$ holding. This implies that the server splits its capacity between the two queues during the busy period of overflow.

(2) By symmetry, the results can be easily adapted (it suffices to substitute everywhere $1 := 2, 2 := 1$, and $\beta := 1/\beta$) to estimate the overflow probability of the second queue and characterize the most likely ways that it builds up.

Figure 4. The most-likely overflow scenarios for class one under the GLQF policy.



8. AN ANALYTIC EXPRESSION FOR THE HEDGING POINT

In this section we provide refined asymptotics on the stockout probabilities and we use them to analytically determine the appropriate hedging point $\mathbf{w}$ such that the stockout probability of each class j is upper bounded by some desired ϵ_j.

Notice that in §7 we considered the large deviations asymptotics of the stockout probabilities as $w_j \to \infty$, for each class j. It has been observed that in some cases one might need a large value of w_j (or equivalently a small value of ϵ_j) to obtain an accurate estimate of the stockout probability. To avoid such accuracy problems we will consider a refined asymptotic. More specifically, we will be estimating the stockout probability of class j by using the expression

$$\mathbf{P}[x^j \leqslant 0] \approx \alpha_j e^{-w_j\theta^*_j}, \tag{33}$$

where θ^*_j is the decay rate of the stockout probability obtained from the results of §7 (either from Theorem 7.1 or Theorem 7.3 depending on the policy implemented). Thus, the hedging point satisfies

$$w_j = -\frac{\log(\epsilon_j/\alpha_j)}{\theta^*_j}, \qquad j = 1, \ldots, m. \tag{34}$$

If $\theta^*_j = \infty$, then stockouts for class j do not occur and a hedging point equal to zero should be used.

An estimate of the constant α_j can be obtained by using an idea from Abate, Choudhury and Whitt (1995) and assuming that the above expression provides the *exact* distribution of the backlog process. Using the transformation in (18) it also provides the distribution $\mathbf{P}[L^j \geqslant w_j]$ for the queue length process in the corresponding make-to-order system operated under the priority policy or the GLQF policy, respectively. Matching the expectation of the distribution in (33) with $\mathbf{E}[L^j]$ we obtain

$$\alpha_j = \theta^*_j \mathbf{E}[L^j]. \tag{35}$$

Thus, to find the asymptotic constant we need the asymptotic exponent θ^*_j and the expectation of the queue length process L^j in a multiclass $G/G/1$ queue. The latter one can be obtained either by approximations or by simulation. We will argue in §10 that simulating to obtain the expectation incurs no additional computational cost to the one

required to estimate a model for the demand and service processes. As an alternative to simulation we next provide approximate expressions for $\mathbf{E}[L^j]$ in the case of the priority policy.

Approximations for $\mathbf{E}[L^j]$ under the Priority Policy

We let $c^2_{D^j}$ denote the squared coefficient of variation (i.e., $c^2_{D^j} = \mathrm{Var}(D_1^j)/(\mathbf{E}[D_1^j])^2$) of the demand process for class j, and c_B^2 the same quantity for the service process B. Moreover, we define

$$\rho_j \stackrel{\Delta}{=} \frac{\sum_{k=1}^j \mathbf{E}[D_1^k]}{\mathbf{E}[B_1]}, \tag{36}$$

and

$$c^2_{D^1+\cdots+D^j} \stackrel{\Delta}{=} \frac{\sum_{k=1}^j \mathrm{Var}(D_1^k)}{(\sum_{k=1}^j \mathbf{E}[D_1^k])^2}. \tag{37}$$

We will next use an approximate formula for the waiting time in a $GI/G/1$ queue to approximate $\mathbf{E}[L^j]$. Note that as defined in §6, $\mathbf{L}_i$ evolves as the queue length process in a discrete-time $G/G/1$ queue with arrival process D^j for class j and service process B. The scheduling policy is a priority policy and as in §7 we reorder, without loss of generality, the classes so that priority is given to classes $1, \ldots, m$ in this order. As a consequence, L_i^j evolves as the queue length process in a single class $G/G/1$ queue with $\sum_{k=1}^j D_i^k$ arrivals at time slot i and service process B. Using the Lindley equation for the queue length of class j at time slot i we obtain

$$\sum_{k=1}^j L_i^k = \left[\sum_{k=1}^j L_{i-1}^k + \sum_{k=1}^j D_i^k - B_i\right]^+, \tag{38}$$

where $[x]^+$ denotes $\max(x, 0)$. Consider now a continuous-time single class $GI/G/1$ queue with interarrival times $\{A_i; i \in \mathbb{Z}\}$ and service times $\{S_i; i \in \mathbb{Z}\}$. Using the Lindley equation the waiting time of customer i is given by

$$W_i = [W_{i-1} + S_{i-1} - A_i]^+. \tag{39}$$

Comparing Equations (38) and (39) we conclude that an approximate formula for the expected waiting time in the latter queue will approximate $\sum_{k=1}^j \mathbf{E}[L^k]$ when we make the substitutions $S_{i-1} := \sum_{k=1}^j D_i^k$ and $A_i := B_i$. Using the Krämer and Langenbach-Belz (1976) approximations as reported in Tijms (1986) we conclude that depending on the value of c_B^2 the quantities $\mathbf{E}[L^j]$, $j = 1, \ldots, m$, can be obtained as the solution of one of the following two triangular systems of linear equations. More specifically,

if $c_B^2 \leqslant 1$ then $\mathbf{E}[L^j]$, $j = 1, \ldots, m$, solve the system of linear equations:

$$\sum_{k=1}^j \mathbf{E}[L^k] = \frac{\rho_j \sum_{k=1}^j \mathbf{E}[D_1^k]}{2(1-\rho_j)}(c_B^2 + c^2_{D^1+\cdots+D^j}) \cdot \exp\left\{\frac{-2(1-\rho_j)(1-c_B^2)^2}{3\rho_j(c_B^2 + c^2_{D^1+\cdots+D^j})}\right\},\ j = 1, \ldots, m, \tag{40}$$

else, if $c_B^2 > 1$ then $\mathbf{E}[L^j]$, $j = 1, \ldots, m$, solve the system of linear equations:

$$\sum_{k=1}^j \mathbf{E}[L^k] = \frac{\rho_j \sum_{k=1}^j \mathbf{E}[D_1^k]}{2(1-\rho_j)}(c_B^2 + c^2_{D^1+\cdots+D^j}) \cdot \exp\left\{\frac{-(1-\rho_j)(c_B^2-1)}{c_B^2 + 4c^2_{D^1+\cdots+D^j}}\right\},\ j = 1, \ldots, m. \tag{41}$$

Although the above expressions were derived from analysis of the $GI/G/1$ queue we will use them in the general case that both demand and service processes have dependencies; we will establish via numerical results that they perform adequately.

9. SELECTING THE "BEST" PRIORITY POLICY

Knowing the structure of the production policy from the fluid analysis, and the hedging point from the large deviations analysis we have all the ingredients to implement the proposed production policies: the priority-based policy and the GLQF-based policy. In the latter case, the costs coefficients c_j are assumed to be given and reflect inventory cost considerations. In the former case, however, the fluid analysis determined that priority classes are ordered according to the backorder cost coefficients. We argued in the Introduction that these coefficients cannot be assumed as given, since it is very hard to quantify backorder costs (especially linear). Thus, to completely characterize the proposed production policies we are left with selecting the priority ordering. We will do that with the objective of minimizing the expected inventory cost $\sum_{j=1}^m h_j \mathbf{E}[(x^j)^+]$.

To that end, we next provide an analytic approximation for the expected inventory cost. Let $1, \ldots, m$ be the priority ordering and let w_j and a_j be as determined by the results in §§7 and 8. We have

$$\begin{aligned} h_j \mathbf{E}[(x^j)^+] &= h_j \mathbf{E}[(w_j - L^j)^+] \\ &= h_j \mathbf{E}[\max(w_j - L^j, 0)] \\ &= h_j\big(w_j - \mathbf{E}[L^j] + \mathbf{E}[\max(0, L^j - w_j)]\big), \end{aligned} \tag{42}$$

where we used Equation (18). Using the asymptotic in (33) we obtain

$$\begin{aligned} \mathbf{E}[\max(0, L^j - w_j)] &= \int_0^\infty \mathbf{P}[\max(0, L^j - w_j) > y]\, dy \\ &= \int_0^\infty \mathbf{P}[L^j - w_j > y]\, dy \\ &\approx a_j e^{-w_j \theta^*_{P_j}} \int_0^\infty e^{-x\theta^*_{P_j}}\, dx \\ &= a_j \frac{e^{-w_j\theta^*_{P_j}}}{\theta^*_{P_j}}. \end{aligned} \tag{43}$$

Thus, using (35) we arrive at the following approximation for the expected inventory cost

$$\sum_{j=1}^m h_j \mathbf{E}[(x^j)^+] \approx \sum_{j=1}^m h_j \left(w_j - \mathbf{E}[L^j] + \mathbf{E}[L^j] e^{-w_j\theta^*_{P_j}}\right). \tag{44}$$

Therefore, we will select the priority ordering that minimizes the right-hand side of (44). This requires searching over $m!$ priority orderings which is computationally prohibitive for a large number of classes. In practice, however, each product j can be thought of as a class of similar products, which implies that m is a fairly small number. In addition, observing that $\theta^*_{P_j}$ increases as class j is elevated to a higher priority class one can develop heuristic search algorithms (similar to the bubble sort algorithm) that can substantially decrease the search time.

10. NUMERICAL RESULTS AND IMPLEMENTATION ISSUES

To assess the performance of the proposed production policies, in this section we present numerical results and discuss issues related to the implementation of the required algorithmic procedures. In particular, we present two examples: (i) a three-class example with priority scheduling, and (ii) a two-class example with GLQF scheduling.

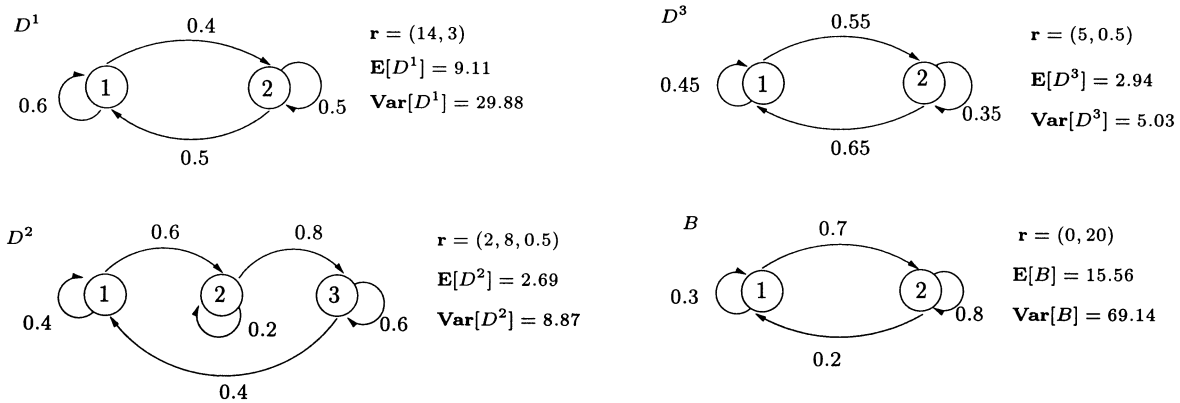
10.1. A Three-Class Example with Priority Scheduling

Demand and service processes are *deterministic Markov-modulated* processes. That is, D_i^j is a function of an underlying discrete-time Markov chain which makes one transition at each time slot. The value of D_i^j is equal to a constant r_k when the Markov chain is in state k. The value of B_i is determined by the evolution of a similar Markov chain. For the particular example we will analyze the models for the demand and service processes are depicted in Figure 5.

We first want to examine the accuracy of the procedure we proposed in §9 to select the priority policy which minimizes the expected inventory cost and satisfies the service level constraints. Table 1 presents some numerical results. In all cases reported, the optimal ordering $\boldsymbol{\chi}$ given in the table is identical to the one obtained by comparing simulated values for the expected inventory costs when the hedging point is fixed to the value $\mathbf{w}$ given in the table. This suggests that the procedure in §9 predicts the optimal priority ordering *accurately*. It is interesting to note that in all cases our approximation for the expected inventory cost (Equation (44)) was within 9% of the simulated value (in fact from 5.5% to 9%), but the error was "ordering preserving," that is, for each case the approximation was in fact an upper bound on the simulated value for the best two orderings.

In Cases 1, 2 and 3 (notice that Cases 1 and 2 are symmetrical) the result for the optimal ordering is intuitively obvious: higher inventory cost h_j and smaller ϵ_j results in higher priority for class j. The situation in Cases 4 and 5 is more subtle. In Case 4, although all classes have the same ϵ's and class 1 has higher inventory cost than class 2, it is optimal to give top priority to class 2. The reason is that class 2 demand arrives in smaller quantities (mean 2.69 with a peak of 8 versus a mean of 9.11 and a peak of 14), which the server can handle "almost immediately," thus, when given top priority it requires very small safety stock. In addition, since it consumes only a small fraction of the capacity, serving it with top priority does not substantially affect the required safety stock for class 1 ($(w_{\chi(1)}, w_{\chi(2)}) = (w_2, w_1) = (4.8, 73.49)$ for classes 2 and 1, respectively). If on the other hand top priority is given to class 1, the required safety stocks are $(w_1, w_2) = (36.53, 64.58)$ for classes 1 and 2, respectively, which justifies that it is preferable to serve class 2 with higher priority. Similarly, in Case 5 it is optimal to give top priority to class 2, although inventory costs are identical across classes and class 1 has smaller ϵ than class 2. Of course, when we bring the ϵ of class 1 sufficiently down (to 0.001) it becomes optimal to serve it with top priority (Case 3). But for this to happen the difference in the ϵ's between the two classes has to be substantial (0.001 vs. 0.01). The important conclusion is that the optimal priority ordering may depend on subtle *distributional* differences between classes that cannot be captured by just the first few moments. It is interesting that these differences are cap-

Figure 5. The models for demand and service processes in the three-class example with priority scheduling.



Note. By $\mathbf{r}$ we denote the vector of demand or production amounts at each state of the corresponding Markov chain. The demand processes are fairly "bursty," meaning that the work arriving per time slot fluctuates substantially as the corresponding Markov chain evolves. The service facility can be in two states: ON (state 2) where it works at a constant production rate, and OFF (state 1) where it is broken down.

Table 1. We follow the notation established so far: $\mathbf{h} = (h_1, h_2, h_3)$ denotes the vector of inventory costs for the products 1,2,3, respectively, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3)$ are the desirable service levels for the stockout probabilities, $\boldsymbol{\chi} = (\chi(1), \chi(2), \chi(3))$ denotes the optimal priority ordering as derived by the procedure in §9, and $\mathbf{w} = (w_{\chi(1)}, w_{\chi(2)}, w_{\chi(3)})$ is the hedging point required to maintain the service levels (given by Equation (34)). To obtain the asymptotic constants a_j (cf. Equation (35)) required to calculate $\boldsymbol{\chi}$ and $\mathbf{w}$, we used the approximation for $\mathbf{E}[L^j]$ reported in §8.

	Inventory costs ($\mathbf{h}$)	Service levels ($\boldsymbol{\epsilon}$)	Optimal ordering ($\boldsymbol{\chi}$)	Hedging Point ($\mathbf{w}$)
Case 1	(1, 2, 3)	(0.05, 0.01, 0.005)	(3, 2, 1)	(3.89, 10.77, 231.72)
Case 2	(3, 2, 1)	(0.005, 0.01, 0.05)	(1, 2, 3)	(44.37, 64.58, 214.26)
Case 3	(1, 1, 1)	(0.001, 0.01, 0.05)	(1, 2, 3)	(88.58, 64.58, 214.26)
Case 4	(3, 2, 1)	(0.01, 0.01, 0.01)	(2, 1, 3)	(4.8, 73.49, 349.39)
Case 5	(1, 1, 1)	(0.007, 0.01, 0.05)	(2, 1, 3)	(4.8, 80.43, 214.26)

Table 2. Comparison of analytically calculated versus simulated hedging points. The priority ordering for all entries is fixed to (1,2,3).

	Calculated $\mathbf{w}$			Simulated $\mathbf{w}'$			Error ($\boldsymbol{\delta}$) (%)		
$\boldsymbol{\epsilon}$	w_1	w_2	w_3	w_1'	w_2'	w_3'	δ_1	δ_2	δ_3
0.1	17.14	35	181.8	17	36	181	0.82	2.78	0.44
0.05	24.97	48.10	240	27	50	239	7.52	3.8	0.42
0.01	43.16	78.53	375.1	44	81	372	1.91	3.05	0.83
$5 \cdot 10^{-3}$	50.99	91.63	432.9	52	94	430	1.94	2.52	0.67
10^{-3}	69.18	122.05	567.6	70	125	563	1.17	2.36	0.82
$5 \cdot 10^{-4}$	77.02	135.15	625.6	78	139	621	1.26	2.77	0.74
10^{-4}	95.20	165.58	760.2	96	169	759	0.83	2.02	0.16

tured by our method, since the large deviations behaviour depends on the whole distribution through the limiting log-moment generating functions.

We next turn our attention to the accuracy of the analytically estimated hedging point (Equation (34)). Table 2 compares the hedging point calculated by Equation (34) (denoted by $\mathbf{w}$ in the table) with the one obtained by simulation (denoted by $\mathbf{w}'$ in the table) for a range of service levels $\boldsymbol{\epsilon}$. For example, if the desired service levels are $(0.01, 10^{-3}, 5 \cdot 10^{-4})$, then from the table we read that the analytically calculated hedging point is $\mathbf{w} = (43.16, 122.05, 625.6)$, while the one obtained from the simulation is $\mathbf{w}' = (44, 125, 621)$. Notice that to find the required hedging point via simulation, one needs to simulate the stockout probabilities for every possible value of the hedging point, which is a very computationally intensive task. The advantage of an analytically obtained hedging point is apparent. The error we report in the table is defined as

$$\delta_j \stackrel{\Delta}{=} \frac{|w_j - w_j'|}{w_j'} \times 100\%,$$

for class j. We observe that the analytically obtained hedging point is fairly accurate with the error being mostly less than 3%, with the exception of two cases where the error is 3.8% and 7.52%, respectively. Recall that we measure everything in production time units, thus, the hedging point is a real number. In the simulation, however, since it is impossible to simulate for every possible value of the hedging point, we considered only integer values. As a result, the reported error includes this "quantization error," and hence underestimates the accuracy of the calculated hedging point (this "quantization" error is up to 1 time unit, that is 3.7% and 2% for the cases where we report errors of 7.52% and 3.8%, respectively).

Finally, Table 3 reports the stockout probabilities achieved with the analytically calculated hedging point $\mathbf{w}$ of Table 2. We observe that the actual stockout probabilities are close to the desired target $\boldsymbol{\epsilon}$. More specifically, in all cases the simulated values have the same order of magnitude with $\boldsymbol{\epsilon}$ and the first significant digit is very close.

The calculation of $\mathbf{w}$ via Equation (34) requires the derivation of an asymptotic constant which depends on the expectation of a queue length process in the corresponding make-to-order system (cf. Equation (35)). We used simulation to obtain this latter value for the calculations in Tables 2 and 3. It is instructive at this point to consider how the proposed computational procedure could be implemented. One requires a detailed model of the demand and service processes which can be obtained by on-line estimating Markov-modulated models from real observations. Given these models, the analytical calculations can be performed and a production policy can be obtained. This approach allows for frequent updates of the Markov-modulated models (and hence the production policy) to accommodate changes in the demand and service conditions. Notice, that since demand and service observations will be made on-line to infer the appropriate stochastic models, the queue length expectations required for the calculation of the hedging point can also be observed at the same time with no additional computational cost.

Table 3. We denote by $\epsilon' = (\epsilon'_1, \epsilon'_2, \epsilon'_3)$ the actual stockout probability obtained by simulation of the system when the hedging point is fixed to the analytically calculated value. For example, if we use the hedging point (50.99, 91.63, 432.9) the simulated values for the stockout probabilities are $(5.14 \cdot 10^{-3}, 5.49 \cdot 10^{-3}, 4.76 \cdot 10^{-3})$. The priority ordering for all entries is fixed to (1,2,3).

	Calculated **w**			Simulated ϵ'		
ϵ	w_1	w_2	w_3	ϵ'_1	ϵ'_2	ϵ'_3
0.1	17.14	35	181.8	$8.79 \cdot 10^{-2}$	$10.45 \cdot 10^{-2}$	$9.93 \cdot 10^{-2}$
0.05	24.97	48.10	240	$5.34 \cdot 10^{-2}$	$5.32 \cdot 10^{-2}$	$4.89 \cdot 10^{-2}$
0.01	43.16	78.53	375.1	$10.43 \cdot 10^{-3}$	$10.74 \cdot 10^{-3}$	$9.54 \cdot 10^{-3}$
$5 \cdot 10^{-3}$	50.99	91.63	432.9	$5.14 \cdot 10^{-3}$	$5.49 \cdot 10^{-3}$	$4.76 \cdot 10^{-3}$
10^{-3}	69.18	122.05	567.6	$10.89 \cdot 10^{-4}$	$11.74 \cdot 10^{-4}$	$9.28 \cdot 10^{-4}$
$5 \cdot 10^{-4}$	77.02	135.15	625.6	$5.24 \cdot 10^{-4}$	$6.00 \cdot 10^{-4}$	$4.61 \cdot 10^{-4}$
10^{-4}	95.20	165.58	760.2	$1.03 \cdot 10^{-4}$	$1.25 \cdot 10^{-4}$	$9.13 \cdot 10^{-5}$

10.2. A Two-Class Example with GLQF Scheduling

We finally consider a two-class system operated under the GLQF-based policy. The demand and service processes are depicted in Figure 6. The inventory costs are $\mathbf{c} = (3, 2)$ which implies that β, as defined in the statement of Theorem 7.3, is equal to 1.5.

As in the previous example, we compare the hedging point obtained analytically (Equation (34)) with the one obtained by simulation. The results are reported in Table 4. Again, we conclude that the analytically obtained hedging point is very accurate. The same comments as in the previous example regarding implementation issues and "quantization" error in the simulation apply.

Furthermore, in Table 5 we report the actual stockout probabilities achieved when the system operates with the analytically calculated hedging point **w** of Table 4.

As in §10.1, we observe that the simulated stockout probabilities have the same exponent with the target ϵ with the first significant digits being very close.

11. CONCLUSIONS

We have combined fluid and large deviations techniques to derive production policies for multiclass make-to-stock manufacturing systems under realistic modeling assumptions. Our analysis can handle linear or quadratic inventory costs and ensures that the stockout probability for each product stays bounded below a desirable threshold. This leads to manufacturing systems with *Quality of Service guarantees*, a feature which we view as being increasingly important in today's competitive and service-oriented environment.

To provide such guarantees we require detailed *distributional* information on the stochastic processes involved (demand and production processes). We demonstrated through numerical results that such information is critical in optimizing the performance of the system. Ignoring it can lead to substantial performance loss. The spread of information technology in manufacturing plants and the capabilities in data collection and in the implementation of sophisticated production policies that it provides, enhance the practical significance of our techniques.

ACKNOWLEDGMENTS

The first author's research was partially supported by the NSF under Grant DMI-9610486 and the MIT-Singapore alliance. The secord author's research was partially supported by the NSF under a CAREER Award ANI-9983221 and under Grants NCR-9706148 and ACI-9873339.

Figure 6. The models for the demand and service processes in the two-class example with GLQF scheduling.

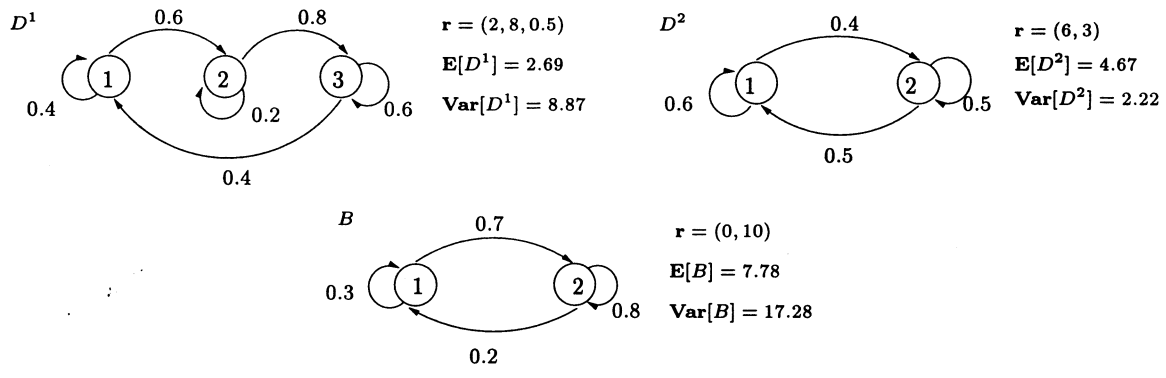


Table 4. Comparison of analytically calculated versus simulated hedging points.

	Calculated $\mathbf{w}$		Simulated $\mathbf{w}'$		Error ($\boldsymbol{\delta}$) (%)	
ϵ	w_1	w_2	w_1'	w_2'	δ_1'	δ_2'
0.15	27.01	42.93	28	44	3.54	2.43
0.1	32.91	51.78	33	52	0.27	0.42
0.05	43.00	66.91	43	67	0	0.13
0.01	66.41	102.04	67	102	0.88	0.04
$5 \cdot 10^{-3}$	76.5	117.16	77	117	0.65	0.14
10^{-3}	99.91	152.29	100	152	0.09	0.19
$5 \cdot 10^{-4}$	110.00	167.42	110	167	0	0.25

Table 5. Comparing the simulated stockout probabilities $\boldsymbol{\epsilon}' = (\epsilon_1', \epsilon_2')$ with the desired target $\boldsymbol{\epsilon}$, when the hedging point is fixed to the analytically calculated value $\mathbf{w}$.

	Calculated $\mathbf{w}$		Simulated $\boldsymbol{\epsilon}'$	
ϵ	w_1	w_2	ϵ_1'	ϵ_2'
0.15	27.01	42.93	$1.51 \cdot 10^{-1}$	$1.58 \cdot 10^{-1}$
0.1	32.91	51.78	$1.04 \cdot 10^{-1}$	$1.05 \cdot 10^{-1}$
0.05	43.00	66.91	$5.04 \cdot 10^{-2}$	$5.27 \cdot 10^{-2}$
0.01	66.41	102.04	$1.04 \cdot 10^{-2}$	$1.01 \cdot 10^{-2}$
$5 \cdot 10^{-3}$	76.5	117.16	$5.04 \cdot 10^{-3}$	$5.10 \cdot 10^{-3}$
10^{-3}	99.91	152.29	$1.04 \cdot 10^{-3}$	$1.03 \cdot 10^{-3}$
$5 \cdot 10^{-4}$	110.00	167.42	$5.04 \cdot 10^{-4}$	$5.15 \cdot 10^{-4}$

REFERENCES

Avram, F., D. Bertsimas, M. Ricard. 1995. Optimization of multiclass fluid queueing networks: a linear control approach. Proceedings of the IMA (F. P. Kelly and R. Williams, eds.), 199–234.

Abate, J., G. L. Choudhury, W. Whitt. 1995. Exponential approximations for tail probabilities in queues, I: Waiting times. *Oper. Res.* **43** (5) 885–901.

Akella, R., P. R. Kumar. 1986. Optimal Control of production rate in a failure prone manufacturing system. *IEEE Trans. Automat. Control* **AC-31** 116–126.

Bertsekas, D. P. 1995. *Dynamic Programming and Optimal Control*, vol. I. Athena Scientific, Belmont, MA.

Bertsimas, D., I. Ch. Paschalidis, J. N. Tsitsiklis. 1998a. Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach. *IEEE Trans. Automat. Control* **43** (3) 315–335.

——, ——, ——. 1998b. On the large deviations behaviour of acyclic networks of $G/G/1$ queues. The *Ann. Appl. Probab.* **8** (4) 1027–1069.

——, ——, ——. 1999. Large deviations analysis of the generalized processor sharing policy, *Queueing Systems* **32** 319–349.

Cramér, H. 1938. Sûr un nouveau théorème-limite de la théorie des probabilités, In Actualiés Scientifique et Industrielles, no. 736 in Colloque consacré à la théorie des probabilités, pages 5–23, Hermann, Paris.

de Véricourt, F., F. Karaesmen, Y. Dallery. 1998. Dynamic scheduling in a make-to-stock system: A partial characterization of optimal policies, *Oper. Res.* **48** (5) 811–819.

Dembo, A., O. Zeitouni. 1993. *Large Deviations Techniques and Applications*. Jones and Bartlett, Boston, MA.

Evans, R. 1967. Inventory control of a multiproduct system with a limited production resource. *Naval Res. Logist.* 173–184.

Federgruen, A., P. Zipkin. 1986. An inventory model with limited production capacity and uncertain demands I. The average cost criterion. *Math. Oper. Res.* **11** (2) 193–207.

Gavish, B., S. Graves. 1980. A one-product production/inventory problem under continuous review policy. *Oper. Res.* **28** 1228–1236.

Glasserman, P. 1996. Allocating production capacity among multiple products. *Oper. Res.* **44** (5) 724–734.

——. 1997. Bounds and asymptotics for planning critical safety stocks. *Oper. Res.* **45** (2) 244–257.

——, Y. Wang. 1998. Fill-rate bottlenecks in production-inventory networks. Working paper.

Glynn, P. W., W. Whitt. 1994. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Probab.* **31A** 131–156.

Ha, A.Y. 1997. Optimal dynamic scheduling policy for a make-to-stock production system. *Oper. Res.* **45** (1) 42–53.

Krämer, W., M. Langenbach-Belz. 1976. Approximate formulae for the delay in the queueing system $GI/G/1$. Proceedings 8th International Teletraffic Congress (Melbourne).

Kapuscinski, R., S. R. Tayur. 1999. Optimal policies and simulation based optimization for capacitated production inventory systems. S. R. Tayur, R. Ganeshan, and M. Magazine, eds. *Quantitative Models for Supply Chain Management*. Kluwer, The Netherlands, 7–40.

Meyn, S. P. 1996. Stability and optimization of queueing networks and their fluid models, Proceedings of the Summer Seminar on The Mathematics of Stochastic Manufacturing Systems (Williamsburg, VA) 17–21.

Paschalidis, I. Ch. 1996. Large deviations in high speed communication networks. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA.

——. 1998. A multiclass queue under the generalized longest queue first policy: Bounds on the overflow probabilities. Unpublished manuscript.

Peña Perez, A., P. Zipkin. 1997. Dynamic scheduling rules for a multiproduct make-to-stock queue. *Oper. Res.* **45** (6) 919–930.

Rockafellar, R. T. 1970. *Convex Analysis*. Princeton University Press, Princeton, NJ.

Sobel, M. 1982. The optimality of full-service policies. *Oper. Res.* **30** 636–649.

Spearman, M. L., R. Q. Zhang. 1999. Optimal lead time policies. *Management Sci.* **45** (2) 290–295.

Tijms, H. C., 1986. *Stochastic Modelling and Analysis: A Computational Approach.* Wiley, New York.

Veatch, M. H., L. M. Wein. 1996. Scheduling a make-to-stock queue: Index policies and hedging points. *Oper. Res.* **44** (4) 643–647.

Wein, L. M., 1992. Dynamic scheduling of a multiclass make-to-stock queue. *Oper. Res.* **40** 724–735.

Zheng, Y.-S., P. Zipkin. 1990. A queueing model to analyze the value of centralized inventory information. *Oper. Res.* **38** (2) 296–307.