Large Deviations-based Asymptotics for Inventory Control in Supply Chains[1] [2]

Ioannis Ch. Paschalidis[3] Yong Liu[4]

Abstract

We consider a model of a capacitated single-class supply chain consisting of a tandem of production facilities and propose production policies in two cases: (a) when each facility has access to its local inventory only, and (b) when it has knowledge of the total downstream inventory. In case (a) the proposed policy guarantees stockout probabilities at each stage to stay bounded below given constants (service level constraints). In case (b) we minimize total expected inventory cost subject to service level constraints. In both cases we rely upon large deviations asymptotics to analytically obtain the policy parameters; this leads to huge computational savings compared to simulation. Our model can accommodate autocorrelated demand and service processes, both critical features of modern failure-prone manufacturing systems. We demonstrate that detailed distributional information on demand and service processes, which is incorporated into large deviations asymptotics, is critical in inventory control decisions. Some extensions to a multiclass setting are discussed.

Keywords: Supply chain management, Inventory control, Service levels, Large Deviations.

1 Introduction

Modern manufacturing enterprises have recognized that production can not be viewed separately from the physical distribution of goods. Instead, both activities should be perceived as indispensable parts of a *supply chain*. Moreover, companies are recognizing the significance of *Quality of Service (QoS)* in acquiring and maintaining market share.

Our primary research objective is to develop effective policies for inventory control in supply chains that address the difficulties present in the new manufacturing environment. There is a large literature on production inventory systems (see [2] for a survey). The single-stage, single-class, version of the problem is significantly simpler; it has been shown in a variety of settings that a so called *base-stock* policy is optimal. In multiclass single-stage systems, there have only been results for special cases, or approximations and heuristics for the general case (see [3] and references therein). In a multiple-stage, single-class system, and *without capacity limits*, Clark and Scarf in their seminal paper [4] have shown the optimality of a production policy where each facility follows a base-stock policy based on the total inventory available in the downstream facilities (we will refer to this as *echelon* inventory). Their result has been generalized in several directions [5, 6]. In the general case where capacity limits exist and demand and service processes are autocorrelated, such a policy is not necessarily optimal. Its simplicity, however, makes it attractive. Under a similar echelon policy [7] proposed a perturbation analysis approach to compute the hedging points in a capacitated single-class multi-stage system, and [8] developed asymptotics to approximate stockout probabilities under renewal demand and constant production capacities.

In this paper we propose and analyze two base-stock production policies. Our first policy uses only local inventory information at each stage of the supply chain. Our second policy has similar structure to the policy proposed in [4], that is, each stage makes decisions based on the total downstream inventory. In both cases, we introduce constraints ensuring that stockout probabilities stay bounded below given desirable levels. Such *service-level* constraints provide a more natural representation of customer satisfaction and are closely watched by manufacturing managers.

Our analysis is general enough to accommodate dependencies in demand and production processes. In practice, demand for various products might have strong correlations. Moreover, manufacturing facilities are *stochastic* and *failure-prone*, which creates dependencies in the production process. In this setting, analyzing stockout probabilities exactly is intractable. We will instead rely upon *large deviations* techniques that lead to asymptotically tight approximations. As a result, we will be able to analytically obtain the appropriate base-stock levels for both policies we consider. Related techniques have been recently used in [3] to devise production policies in a multiclass, single-stage setting.

The remainder of this paper is organized as follows: In Section 2, we provide the detailed model of the supply chain, introduce the production policies we will consider, and outline our approach. In Section 3, we analyze our first production policy, which is based on local inventory information. In Section 4, we treat our second policy which uses echelon inventory information. We discuss extensions to the multiclass case in Section 5. Numerical results that assess the accuracy of the proposed analytical approach are in Section 6 and conclusions in Section 7.

2 The Model

Figure 1 depicts the supply chain model we consider in this paper. This system produces a single product class and consists of M production facilities in tandem. We will be referring to these facilities as stages of the supply chain. External demand is met from the finished goods inventory maintained in front of the stage one production facility, and is backordered if inventory is not avail-

[1] The full version of the paper is in [1].

[2] Research partially supported by the NSF under a Career award ANI-9983221 and grants NCR-9706148 and ACI-9873339.

[3] Department of Manufacturing Engineering, Boston University, Boston, MA 02215, USA. E-mail: yannisp@bu.edu, Web: http://ionia.bu.edu/.

[4] Department of Manufacturing Engineering, Boston University, Boston, MA 02215, USA. E-mail: liuyong@bu.edu.

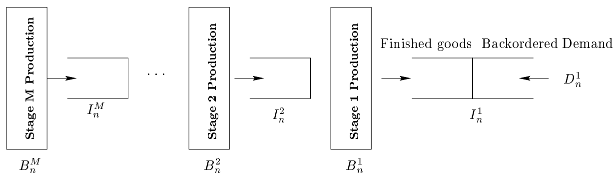

Figure 1: The model of the supply chain.

able. Every production facility is fed by its upstream facility; in particular, to produce one unit facility i, $i = 1, \ldots, M-1$, requires one unit of the product of facility $i+1$. We assume that facility M is fed with an infinite supply of raw material. In front of every facility i, $i = 2, \ldots, M$, there is an inventory buffer which holds the final product of that facility and from which facility $i-1$ draws material for its production. We assume a periodic review policy where time is divided into time slots of equal duration. For all $i = 1, \ldots, M$ and n we let B_n^i denote the amount that the facility at stage i can produce during time slot n (production capacity). We also let D_n^1 denote the amount of external orders arriving at stage one during time slot n. Finally, we let I_n^i, $i = 1, \ldots, M$, denote the inventory in front of stage i at the beginning of time slot n.

The demand process $\{D_n^1;\ n \in \mathbb{Z}\}$ and the production processes $\{B_n^i;\ n \in \mathbb{Z}\}$, $i = 1, \ldots, M$, are arbitrary stationary stochastic processes that satisfy certain mild technical conditions (for details see [9, 10]). For stability purposes we assume that

$$\mathbf{E}[D_n^1] < \min_{i=1,\ldots,M} \mathbf{E}[B_n^i], \tag{1}$$

which by stationarity carries over to all time slots n. Stability can be shown under both base-stock policies we will consider in this paper by using techniques from [11] and [12].

Our objective is to find a production policy that minimizes expected inventory costs and guarantees that the steady-state stockout probability $\mathbf{P}[I_n^1 \leq 0]$, at some arbitrary time slot n, does not exceed a desirable small value ϵ. We will be referring to this as a *service-level* constraint. In this paper, we will propose policies in two separate cases: (a) when each stage i has knowledge of its local inventory I_n^i only, and (b) when each facility i has knowledge of the total downstream inventory $I_n^i + I_n^{i-1} + \ldots + I_n^1$. In both cases, we will implement a base-stock policy.

An exact expression for the stockout probabilities is intractable, especially in view of the rather complicated (autocorrelated) models for the demand and production processes. We will resort to (asymptotic) *large deviations* techniques.

Consider the process $\{X_n;\ n \in \mathbb{Z}\}$, where X_i are identically distributed, possibly autocorrelated, random variables. Let $S_n = \sum_{i=1}^n X_i$. We will refer to $\Lambda(\theta) \triangleq \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}\left[e^{\theta S_n}\right]$ as the *limiting* log-*moment generating function*, and to $\Lambda^*(a) \triangleq \sup_\theta(\theta a - \Lambda(\theta))$ as the *large deviation rate function* of the process X. We also define

$$\Lambda^{*+}(a) \triangleq \begin{cases} \Lambda^*(a) & \text{if } a > m, \\ 0 & \text{if } a \leq m, \end{cases} \tag{2}$$

$$\Lambda^{*-}(a) \triangleq \begin{cases} \Lambda^*(a) & \text{if } a < m, \\ 0 & \text{if } a \geq m, \end{cases} \tag{3}$$

which are large deviations rate functions associated with the tail probabilities $\mathbf{P}[S_n \leq na]$ and $\mathbf{P}[S_n \leq na]$, respectively. The convex duals of these functions will be denoted by $\Lambda^+(\theta)$ and $\Lambda^-(\theta)$, respectively. 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.

3 The Decomposition Approach — A Local Information Case

Next we consider the case where each stage has knowledge of its local inventory only. We start by reviewing the single-stage problem; our analysis of the multi-stage problem will build on this.

3.1 Single-stage problem

In agreement with the notation introduced in Section 2 we will be using I_n, D_n, and B_n, to denote the inventory, demand, and production capacity, respectively, during time slot n. We implement a base-stock policy that maintains a safety stock w in the inventory. The inventory evolves as follows

$$I_{n+1} = \min\{I_n - D_n + B_n, w\}. \tag{4}$$

The objective is to keep the stockout probability $\mathbf{P}[I_n \leq 0] \leq \epsilon$, where $\epsilon > 0$ is some given threshold. This has been done in [3] using large deviations techniques. On a notational remark, in the sequel we will be dropping the reference to the time slot (subscript n) when referring to steady-state quantities.

Theorem 3.1 (Single Stage [3]) *For the single-stage system, the steady-state stockout probability satisfies*

$$\lim_{w\to\infty} \frac{1}{w} \log \mathbf{P}[I \leq 0] = -\theta_s^*, \tag{5}$$

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

$$\Lambda_D^+(\theta) + \Lambda_B(-\theta) = 0. \tag{6}$$

More intuitively, for large enough w we have

$$\mathbf{P}[I \leq 0] \sim e^{-w\theta_s^*},$$

thus, the minimum w that satisfies $\mathbf{P}[I \leq 0] \leq \epsilon$ is

$$w = -\frac{\log \epsilon}{\theta_s^*}. \tag{7}$$

3.2 Multiple Stages

We now return to our original problem with M stages. We propose a base-stock policy that maintains a safety stock equal to w_i for the (local) inventory of every stage i, $i = 1, \ldots, M$. In particular, stage i produces until the local inventory I_n^i reaches the hedging point w_i and idles if $I_n^i \geq w_i$. The amount produced by stage i constitutes demand for the upstream stage $i+1$, for $i = 1, \ldots, M-1$; we will denote it by D_n^{i+1}.

The dynamics for the supply chain are

$$I_{n+1}^i = \min\{I_n^i - D_n^i + B_n^i,\ I_n^i - D_n^i + I_n^{i+1},\ w_i\}, \quad i = 1, \ldots, M-1, \tag{8}$$

$$I_{n+1}^M = \min\{I_n^M - D_n^M + B_n^M,\ w_M\}. \tag{9}$$

We define the inventory shortfall for stage i as follows:

$$L_n^i \triangleq w_i - I_n^i, \quad i = 1, \cdots, M,$$

and the dynamics of the supply chain can be written as

$$\begin{aligned} L_{n+1}^i = &\max\{0, L_n^i + D_n^i + L_n^{i+1} - w_{i+1}, \\ &L_n^i + D_n^i - B_n^i\}, \ i = 1, \ldots, M-1, \quad (10) \\ L_{n+1}^M = &\max\{L_n^M + D_n^M - B_n^M, 0\}. \quad (11) \end{aligned}$$

The major difficulty for analyzing this model and characterizing the stockout probabilities is that the production is constrained not only by its own capacity, but also by the upstream inventory. To bypass this difficulty we will *decouple* the various stages by ignoring the upstream inventory constraint on the downstream production. We can intuitively argue that this decomposition is in fact accurate when the inventory level of the upstream stage is high enough; then the influence of the upstream inventory constraint will be insignificant when compared to the capacity constraint. More specifically, the proposed decomposition amounts to assuming that the system operates according to a policy which satisfies $I_n^{i+1} \geq B_n^i, \quad i = 1, \ldots, M-1$, almost surely for all time slots n. As a result, the dynamics of the supply chain can be simplified as follows:

$$L_{n+1}^i = \max\{L_n^i + D_n^i - B_n^i, 0\}, \quad i = 1, \cdots, M. \quad (12)$$

That is, each stage behaves exactly as a single-stage system. This transformation implies $\mathbf{P}[I_n^i \leq 0] = \mathbf{P}[L_n^i \geq w_i]$, for each stage i, $i = 1, \ldots, M$.

Next note that the dynamics in (12) are exactly the dynamics of M decoupled make-to-order G/G/1 queues. For stage 1, $\{D_n^1; n \in \mathbb{Z}\}$ is the external demand process, whose large deviations rate function is assumed known. For the remaining stages $i = 2, \ldots, M$, recall that D_n^i is the demand for stage i generated by stage $i-1$. In the equivalent make-to-order version of the system D_n^i can be interpreted as the number of departures from the stage $i-1$ queue during time slot n. The following theorem characterizes the large deviations behaviour of the departure process $\{D_n^i; n \in \mathbb{Z}\}$, for all $i = 2, \ldots, M$. The proof, which is based on a result in [9], can be found in [1].

Theorem 3.2 (Departure Process) *The partial sum of the departure process of the $G/G/1$ queue of stage $i-1$ satisfies*

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbf{P}\left[\sum_{j=1}^{n} D_j^i \geq na\right] = -\Lambda_{D^i}^{*+}(a), \quad (13)$$

for all $i = 2, \ldots, M$, where

$$\Lambda_{D^i}^{*+}(a) = \Lambda_{B^{i-1}}^{*+}(a) + \Lambda_{\Gamma^{i-1}}^{*+}(a), \quad (14)$$

$$\Lambda_{\Gamma^{i-1}}^{*+}(a) = \sup_{\{\theta \mid \Lambda_{D^{i-1}}^{+}(\theta) + \Lambda_{B^{i-1}}(-\theta) < 0\}} \left[\theta a - \Lambda_{D^{i-1}}^{+}(\theta)\right].$$

We now have all the ingredients to analyze L_n^i for every stage i in isolation. The result is summarized in the following theorem.

Theorem 3.3 *For every stage $i = 1, \ldots, M$ of the decoupled system, the steady-state queue length L^i satisfies*

$$\lim_{w_i \to \infty} \frac{1}{w_i} \log \mathbf{P}[L^i \geq w_i] = -\theta_{L,i}^*, \quad (15)$$

where $\theta_{L,i}^*$ is the largest root of the equation

$$\Lambda_{D^i}^{+}(\theta) + \Lambda_{B^i}(-\theta) = 0, \quad (16)$$

and $\Lambda_{D^i}^{+}(\theta)$ is the convex dual of $\Lambda_{D^i}^{*+}(a)$.

Assume now that the stockout probability for stage one needs to be upper bounded by some ϵ_1. We can then select the requirement for the stockout probability of stage i, ϵ_i, to be the same as, or an order of magnitude less than, the corresponding requirement, ϵ_{i-1}, for its downstream stage $i-1$. Using the results of this section, we obtain hedging points given by $w_i = -\frac{\log \epsilon_i}{\theta_{L,i}^*}$, $i = 1, \cdots, M$. To improve the accuracy of the asymptotics, especially for fairly large ϵ's, we can introduce a constant c_i and consider the approximation

$$\mathbf{P}[I^i \leq 0] = \mathbf{P}[L^i \geq w_i] \sim c_i \ e^{-w_i \theta_{L,i}^*}, \quad (17)$$

for $i = 1, \ldots, M$. Note that in the decoupled system c_i is independent of w_i's, and can be obtained either by approximations or by simulation (see [3] for details). Hence, the hedging points satisfy

$$w_i = -\frac{\log(\epsilon_i / c_i)}{\theta_{L,i}^*}, \quad i = 1, \cdots, M. \quad (18)$$

Numerical results that help assess the accuracy of the large deviations asymptotics are given in Section 6.

4 The Multi-echelon Approach — A Global Information Case

Next, we consider the case where echelon inventory information is available at every stage $i = 1, \ldots, M$. This will allow us to trade-off inventory between various stages in order to reduce expected inventory costs while maintaining the service level constraints.

We will be using the notation introduced in Section 2. In particular, X_n^i denotes the echelon inventory at time slot n and stage $i = 1, \ldots, M$. We have

$$X_n^i = I_n^i + \ldots + I_n^1 = I_n^i + X_n^{i-1}, \qquad i = 1, \ldots, M.$$

We implement an echelon base-stock production policy that maintains a hedging point or safety stock of w_i for the echelon inventory at stage i. More specifically, the facility at stage i produces until X_n^i reaches w_i and idles otherwise. Clearly, $w_1 \leq w_2 \leq \cdots \leq w_M$. Figure 2 depicts the supply chain model and indicates the stages corresponding to the echelon safety stocks.

As in Section 3, we define the shortfall of echelon i inventory as

$$Y_n^i \triangleq w_i - X_n^i, \quad (19)$$

which implies $\mathbf{P}[X_n^i \leq 0] = \mathbf{P}[Y_n^i \geq w_i]$. In terms of the shortfalls the dynamics of the supply chain can be

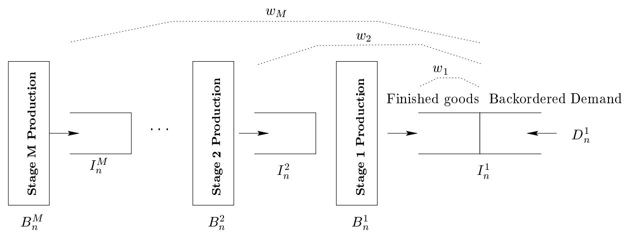

Figure 2: The supply chain model under the echelon base-stock policy.

written as:

$$Y_{n+1}^i = \max\{0, Y_n^{i+1} + D_n^1 - (w_{i+1} - w_i), \quad Y_n^i + D_n^1 - B_n^i\}, \quad i = 1, \ldots, M-1, \tag{20}$$

$$Y_{n+1}^M = \max\{Y_n^M + D_n^1 - B_n^M,\ 0\}. \tag{21}$$

The main result of this section (see [1] for a proof), is a large deviations result for the steady-state probability $\mathbf{P}[Y^1 \geq w_1]$, which is equal to the steady-state stockout probability $\mathbf{P}[X^1 \leq 0]$.

Theorem 4.1 *Assume the hedging points $w_1, w_2, \ldots, w_M$ in the multi-echelon system (cf. (20), (21)) satisfy*

$$w_i = \beta_{i-1} w_1, \qquad i = 2, \ldots, M,$$

where β_i are constants and $1 \leq \beta_1 \leq \cdots \leq \beta_{M-1}$. The steady-state shortfall Y^1 of echelon 1 satisfies

$$\lim_{w_1 \to \infty} \frac{1}{w_1} \log \mathbf{P}[Y^1 \geq w_1] = -\theta_{G,1}^*, \tag{22}$$

where $\theta_{G,1}^*$ is determined by

$$\begin{aligned}\theta_{G,1}^* = \min \Bigg[& \inf_{a>0} \frac{1}{a} \inf_{x_0 - x_1 = a} \left(\Lambda_{D^1}^{*+}(x_0) + \Lambda_{B^1}^{*-}(x_1) \right), \\ & \inf_{a>0} \frac{1}{a} \inf_{\substack{x_0 - \alpha_1 x_1 - \alpha_2 x_2 = a\beta_1 \\ \alpha_1 + \alpha_2 = 1}} \left(\Lambda_{D^1}^{*+}(x_0) \right. \\ & \left. + \alpha_1 \Lambda_{B^1}^{*-}(x_1) + \alpha_2 \Lambda_{B^2}^{*-}(x_2) \right), \ldots, \\ & \inf_{a>0} \frac{1}{a} \inf_{\substack{x_0 - \alpha_1 x_1 - \cdots - \alpha_M x_M = a\beta_{M-1} \\ \alpha_1 + \cdots + \alpha_M = 1}} \left(\Lambda_{D^1}^{*+}(x_0) \right. \\ & \left. + \alpha_1 \Lambda_{B^1}^{*-}(x_1) + \cdots + \alpha_M \Lambda_{B^M}^{*-}(x_M) \right) \Bigg]. \end{aligned} \tag{23}$$

Remarks :

1. More intuitively, Theorem 4.1 asserts
$$\mathbf{P}[X^1 \leq 0] = \mathbf{P}[Y^1 \geq w_1] \sim e^{-\theta_{G,1}^* w_1}. \tag{24}$$

2. The proof of Thm. 4.1 in [1] characterizes the most likely path that leads to stockouts and provides intuition on how they occur. It is shown that
$$\theta_{G,1}^* = \min(\theta_1^*, \beta_1 \theta_2^*, \ldots, \beta_{M-1} \theta_M^*), \tag{25}$$
where θ_i^*, $i = 1, \ldots, M$, is the largest root of the equation $\sup_{a_1 + \cdots + a_i = 1} (\Lambda_{D^1}(\theta) + a_1 \Lambda_{B^1}(-\theta) + \cdots + a_i \Lambda_{B^i}(-\theta)) = 0$. Consider the case $\theta_{G,1}^* = \beta_{i-1} \theta_i^*$ for some $i = 1, \ldots, M$, where $\beta_0 \triangleq 1$. To avoid degenerate cases assume that all production processes B^i have distinct limiting log-moment generating functions and that $1 < \beta_1 < \cdots < \beta_{M-1}$. It can be shown that θ_i^* is also the largest root of the equation $\Lambda_{D^1}(\theta) + \Lambda_{B^i}(-\theta) = 0$ and the stockout probability at stage 1 behaves as the exponential
$$e^{-\beta_{i-1} \theta_i^* w_1} = e^{-\theta_i^* w_i}.$$
Considering the single stage result (cf. Thm. 3.1) we can say that stage i production capacity is the "bottleneck" and characterizes the stockout probability at stage 1.

Thm. 4.1 can be generalized to characterize the stockout probability of the remaining echelon inventories X^i, $i = 2, \ldots, M$. In particular, echelon inventory X^i at stage i, $i = 1, \ldots, M$ is the echelon one inventory of an $(M+1-i)$-stage supply chain starting at the ith stage of the original system.

To improve the accuracy of the approximation, especially for relatively large stockout probabilities (i.e., small safety stock w_1), we will use the following refined approximation

$$\mathbf{P}[Y^1 \geq w_1] \approx f_1(\boldsymbol{\beta}) e^{-\theta_{G,1}^* w_1}, \tag{26}$$

where the prefactor $f_1(\boldsymbol{\beta})$ is a function of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{M-1}) = (\frac{w_2}{w_1}, \ldots, \frac{w_M}{w_1})$. To compute the prefactor, in [1] we evaluate the stockout probability $\mathbf{P}[Y^1 \geq w_1]$ at several sample points $\mathbf{w} = (w_1, \ldots, w_M)$ by simulation and then find a piecewise linear and convex function $f_1(\boldsymbol{\beta})$ so that $f_1(\boldsymbol{\beta}) e^{-\theta_{G,1}^* w_1}$ matches the true value of $\mathbf{P}[Y^1 \geq w_1]$ at those sample points. It should be noted that this heuristic procedure is one potential approach. Alternatively, given a data set consisting of N sample points we can approximate $f_1(\boldsymbol{\beta})$ by some other parametric form.

We next discuss how to approximate inventory costs, which we assume them to be linear. Let h_i be the holding cost for echelon-i inventory for $i = 1, \ldots, M$. Noting that the expected echelon-i inventory is given by $\mathbf{E}[I^i] + \cdots + \mathbf{E}[I^2] + \mathbf{E}[(I^1)^+]$, where $(I^1)^+ = \max(I^1, 0)$, in [1] we show that the total expected inventory cost can be approximated by the following expression

$$\begin{aligned} & \sum_{i=1}^M h_i (w_i - \mathbf{E}[Y^i]) + (h_1 + \cdots + h_M) \mathbf{E}[(Y^1 - w_1)^+] \\ & = \sum_{i=1}^M h_i (w_i - \mathbf{E}[Y^i]) + (h_1 + \cdots + h_M) f_1(\boldsymbol{\beta}) \frac{e^{-\theta_{G,1}^* w_1}}{\theta_{G,1}^*}. \end{aligned} \tag{27}$$

To obtain an analytical approximation for the inventory cost we are now left with computing $\mathbf{E}[Y^i]$. This is hard to do analytically; instead we use an approach similar to the one used in obtaining $f_1(\boldsymbol{\beta})$. In [1] we show that $\mathbf{E}[Y^i]$ is a convex and nonincreasing function of $(w_{i+1} - w_i, \ldots, w_M - w_{M-1})$ in every coordinate. Using sample points obtained from simulation, we construct a piecewise linear convex function $g_i(w_{i+1} - w_i, \ldots, w_M - w_{M-1})$ which approximates $\mathbf{E}[Y^i]$ for $i = 1, \ldots, M-1$.

Summarizing, using (27) and (26) we formulate the problem of minimizing expected inventory cost subject to service level constraints as follows

$$
\begin{array}{ll}
\min & \sum_{i=1}^{M} h_i(w_i - g_i(\boldsymbol{\zeta}_i)) + \left(\sum_{i=1}^{n} h_i \right) f_1(\boldsymbol{\beta}) \frac{e^{-\theta^*_{G,1} w_1}}{\theta^*_{G,1}} \\
\text{s.t.} & \mathbf{P}[Y^i \geq w_i] = f_i(\boldsymbol{\xi}_i)\, e^{-\theta^*_{G,i} w_i} \leq \epsilon_i,\ \forall i, \\
& w_M \geq \cdots \geq w_2 \geq w_1 \geq 0,
\end{array}
\tag{28}
$$

where $\boldsymbol{\zeta}_i = (w_{i+1} - w_i, \ldots, w_M - w_{M-1})$ and $\boldsymbol{\xi}_i = (w_{i+1}/w_i, \ldots, w_M/w_i)$. This problem can be solved analytically using standard nonlinear programming techniques. We will see in Section 6 that the solution predicted by the problem in (28) is accurate. The very significant advantage of our approach is that we can set the proper hedging points analytically, which leads to huge computational savings.

5 Extensions to the Multiclass Case

We extend the model depicted in Figure 1 as follows. We assume that instead of a single product class the system produces K products. We maintain separate inventory buffers for each product in front of each stage.

We will implement a scheduling policy which allocates a constant fraction of the capacity of each facility to every class. In particular, we will let $\phi_{k,i}$ denote the fraction of the stage-i capacity B^i_n allocated to class k during time slot n, for all $k = 1, \ldots, K$ and $i = 1, \ldots, M$, where $\sum_{k=1}^{K} \phi_{k,i} = 1$. Note that $\phi_{k,i}$ is constant for all time slots. This policy will be referred to as the *generalized processor sharing* policy (GPS).

According to the GPS policy the capacity allocated to a class k can be distributed to the remaining classes during times that class k has no work to be done. To facilitate the analysis we will decompose the system across classes and ignore the unutilized capacity allocated to a class during times that other classes are not busy. Hence, the multiclass supply chain is decomposed in K single class chains and the results we have developed in this paper are immediately applicable. In particular, our single class asymptotics and hedging points can be derived for each class k by using capacity $\phi_{i,k} B^i_n$ at each stage i during time slot n.

6 Numerical Results

In this section, we present numerical results to evaluate the performance of the proposed large deviations approximations. We will consider a two-stage system and we will (a) use the decomposition approach developed in Section 3 to derive a base-stock policy for each stage under a variety of service level requirements, and (b) use the echelon base-stock policy analyzed in Section 4 to optimize the expected inventory cost subject to service level constraints.

In both parts (a) and (b) we consider Markov-modulated demand and production processes. Figure 3 depicts the model of the demand and production processes in a two-stage supply-chain. We denote by $\mathbf{r}$ the vector of demand or production amounts at each state of the corresponding Markov chain.

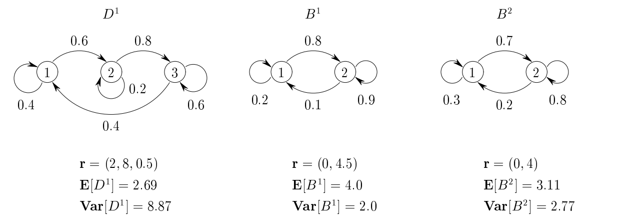

Figure 3: The models of demand and production processes in a two-stage system.

6.1 The decomposition approach

Results for the example are shown in Table 1. The first two columns list the desired service level requirements for stages 1 and 2, respectively. The third and fourth column list the analytically computed hedging points, for stages 1 and 2, respectively. Finally, in the 5th column we report the actual simulated stockout probability at stage one (we simulated the system with integer hedging points computed by rounding the values obtained analytically).

ϵ_1	ϵ_2	w_1	w_2	$\mathbf{P}[I^1 \leq 0]$
10^{-2}	10^{-2}	16.0	46.6	1.36×10^{-2}
10^{-2}	10^{-3}	16.0	71.3	1.01×10^{-2}
10^{-3}	10^{-3}	24.9	71.3	1.14×10^{-3}
10^{-3}	10^{-4}	24.9	96.0	0.98×10^{-3}
10^{-4}	10^{-4}	33.8	96.0	0.97×10^{-4}
10^{-4}	10^{-5}	33.8	120.7	0.93×10^{-4}
10^{-5}	10^{-5}	42.7	120.7	1.00×10^{-5}
10^{-5}	10^{-6}	42.7	145.4	0.92×10^{-5}

Table 1: Numerical results from the decomposition approach for the example.

We selected the service level requirement of the second stage to be same as or one order of magnitude less than ϵ_1. The numerical results suggest that this suffices to make the decomposition approach valid. In particular, we observe that the proposed large deviations asymptotics are fairly accurate, the error being less than 5% (they capture the exponent of the stockout probability and get fairly close in the first significant digit). Of course, there are many combinations of w_1 and w_2 that would lead to the same service level. Our decomposition approach yields one possible such combination. Next, we explore how we can select the best such combination to minimize expected inventory costs.

6.2 The multi-echelon approach

We apply the multi-echelon approach to the example considered above. Using the results of Thm. 4.1, its generalization for other echelons, and the characterization of $\theta^*_{G,1}$ in Eq. (25) we obtain $\theta^*_1 = 0.2584, \theta^*_2 = 0.0932$, and $\theta^*_{G,1} = \min\left(\theta^*_1, \frac{w_2}{w_1}\theta^*_2\right)$. For echelon 2, we obtain $\theta^*_{G,2} = 0.0932$. Solving the optimization problem in (28) we obtain the results reported in Table 2. The analytical solution is very close to the one obtained by simulation. We compare the multi-echelon policy with the decomposition policy in Table 3. As expected the multi-echelon policy leads to more economic solutions.

			Analytical Results				Simulation Results			
ϵ_1	h_1	h_2	w_1^*	w_2^*	$\mathbf{P}[X^1 \leq 0]$	$\mathbf{E}[\text{Cost}]$	w_1^*	w_2^*	$\mathbf{P}[X^1 < 0]$	$\mathbf{E}[\text{Cost}]$
10^{-2}	1	1	21.8	54.5	1.0×10^{-2}	64.2	21	55	1.0×10^{-2}	64.1
10^{-2}	5	1	16.6	66.4	1.0×10^{-2}	127.7	17	67	1.0×10^{-2}	130.5
10^{-2}	1	10	26.6	52.9	1.0×10^{-2}	460.2	25	53	1.0×10^{-2}	460.1
10^{-3}	1	1	29.2	81.0	1.0×10^{-3}	98.4	30	80	1.0×10^{-3}	98.2
10^{-4}	1	1	38.1	105.7	1.0×10^{-4}	132.1	40	103	1.0×10^{-4}	131.3

Table 2: Numerical results for the system under the multi-echelon policy.

6.3 Significance of distributional information

As our final example we present a two-stage supply chain model operated under the multi-echelon inventory policy. We will demonstrate that distributional infor-

Analytical Results				Simulation Results					
ϵ_1	ϵ_2	w_1	w_2	$\mathbf{E}[(I^1)^+]$	$\mathbf{E}[I^2]$	$\mathbf{P}[I^1 \leq 0]$	h_1	h_2	$\mathbf{E}[\text{Cost}]$
10^{-2}	10^{-3}	16.0	71.3	13.6	62.8	1.01×10^{-2}	1	1	90.0
10^{-2}	10^{-3}	16.0	71.3	13.6	62.8	1.01×10^{-2}	5	1	144.3
10^{-2}	10^{-3}	16.0	71.3	13.6	62.8	1.01×10^{-2}	1	10	777.4
10^{-3}	10^{-3}	24.9	71.3	22.6	62.8	1.14×10^{-3}	1	1	108.0
10^{-3}	10^{-4}	24.9	96.0	22.6	87.8	0.98×10^{-3}	1	1	133.0
10^{-4}	10^{-4}	33.8	96.0	31.6	87.8	0.97×10^{-4}	1	1	151.0
10^{-4}	10^{-5}	33.8	120.7	32.0	112.8	0.93×10^{-4}	1	1	176.0

Table 3: The expected inventory costs under the decomposition policy.

mation on the demand and service processes is critical in making inventory control decisions.

The demand and production processes are all discrete-time Markov modulated processes. Letting $\mathbf{P}$ and $\mathbf{r}$ denote the transition probabilities and the vector of demand or production amounts in each state of the corresponding Markov chain we set:

$$\mathbf{r}_D = (5, 10), \quad \mathbf{P}_D = \begin{bmatrix} 0.2 & 0.8 \\ 0.4 & 0.6 \end{bmatrix}, \quad \mathbf{E}[D] = 8.33,$$

$$\mathbf{r}_{B^1} = (0, 25), \quad \mathbf{P}_{B^1} = \begin{bmatrix} 0.2 & 0.8 \\ 0.3 & 0.7 \end{bmatrix}, \quad \mathbf{E}[B^1] = 18.18,$$

$$\mathbf{r}_{B^2} = (0, 14), \quad \mathbf{P}_{B^2} = \begin{bmatrix} 0.15 & 0.85 \\ 0.05 & 0.95 \end{bmatrix}, \quad \mathbf{E}[B^2] = 13.22.$$

Applying the results of Section 4 (cf. (25)) we obtain $\theta_1^* = 0.1785$ and $\theta_2^* = 0.2029$.

We compute $\theta_{G,1}^* = 0.1785$, which according to the discussion in Remark 2 of Section 4 implies that the "bottleneck" is stage 1 in the sense that the process B^1 and not B^2 characterizes the stockout probability at stage one. This seems to contradict naive intuition that the "bottleneck" is stage two since $\mathbf{E}[B^1] > \mathbf{E}[B^2]$! The conclusion that the "bottleneck" is stage one is explained by noting that B^1 is more bursty than B^2.

7 Conclusions

We have developed two production policies for inventory control in a multi-stage single-class supply chain. Demand and service processes are general, potentially autocorrelated processes, which makes it possible to model complex demand scenarios and failure-prone production facilities. Both policies emphasize quality of service, which is becoming important in modern manufacturing, by maintaining desirable service level constraints. Our analysis under the echelon-base stock policy provides particular insight on how stockouts occur. In particular, it identifies a "bottleneck" stage whose production capacity is "responsible" for stockouts at stage 1. But, this "bottleneck" stage is not necessarily the one with the smallest mean production capacity; it depends on the full distribution of the production processes. We provided a simple numerical example to underline this observation, which at first sight might appear counterintuitive.

The echelon base-stock policy enables optimization among all possible hedging point vectors that satisfy the service level constraints; by solving a nonlinear optimization problem we select the one with minimum expected inventory cost. Numerical results show that the solutions obtained by analysis are very close to the ones obtained by brute force simulation. Our analytic approach for selecting appropriate hedging points leads to dramatic computational savings when compared to the time needed to obtain them by simulation.

Acknowledgments

We would like to thank Dimitris Bertsimas for several useful discussions.

References

[1] I.Ch. Paschalidis and Y. Liu, "Large deviations-based asymptotics for inventory control in supply chains," Tech. Rep., Dept. of Manufacturing Eng., Boston University, April 2000, submitted for publication, available at `http://ionia.bu.edu`.

[2] R. Kapuscinski and S.R. Tayur, "Optimal policies and simulation based optimization for capacitated production inventory systems," in *Quantitative Models for Supply Chain Management*, S.R. Tayur, R. Ganeshan, and M. Magazine, Eds., chapter 2, pp. 7–40. Kluwer, 1999.

[3] D. Bertsimas and I. Ch. Paschalidis, "Probabilistic service level guarantees in make-to-stock manufacturing systems," Tech. Rep., Department of Manufacturing Engineering, Boston University, February 1999; Revised September 1999, to appear *Operations Research*, available at `http://ionia.bu.edu`.

[4] A.J. Clark and H. Scarf, "Optimal policies for a multi-echelon inventory problem," *Management Science*, vol. 6, pp. 475–490, 1960.

[5] A. Federgruen and P. Zipkin, "Computational issues in an infinite horizon multi-echelon inventory model," *Operations Research*, vol. 32, pp. 818–836, 1984.

[6] F. Chen and J.-S. Song, "Optimal policies for multi-echelon inventory problems with nonstationary demand," Preprint, Graduate School of Business, Columbia University, 1997.

[7] P. Glasserman and S. Tayur, "Sensitivity analysis for base-stock levels in multiechelon production-inventory systems," *Management Science*, vol. 45, no. 2, pp. 263–281, 1995.

[8] P. Glasserman, "Bounds and asymptotics for planning critical safety stocks," *Operations Research*, vol. 45, no. 2, pp. 244–257, 1997.

[9] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis, "On the large deviations behaviour of acyclic networks of G/G/1 queues," *The Annals of Applied Probability*, vol. 8, no. 4, pp. 1027–1069, 1998.

[10] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis, "Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach," *IEEE Transactions on Automatic Control*, vol. 43, no. 3, pp. 315–335, 1998.

[11] F. Baccelli and Z. Liu, "On a class of stochastic recursive sequences arising in queueing theory," *Annals of Probability*, vol. 20, pp. 350–374, 1992.

[12] P. Glasserman and S. Tayur, "The stability of a capacitated, multi-echelon production-inventory system under a base-stock policy," *Operations Research*, vol. 42, no. 5, pp. 913–925, 1994.