Production, Manufacturing and Logistics

The influence of demand variability on the performance of a make-to-stock queue

Zied Jemaï [a], Fikri Karaesmen [b],*

[a] *Laboratoire Génie Industriel, Ecole Centrale Paris, Grande Voie des Vignes, 92295 Chatenay-Malabry Cedex, France*
[b] *Department of Industrial Engineering, Koç University, 34450 Sariyer, Istanbul, Turkey*

Abstract

Variability, in general, has a deteriorating effect on the performance of stochastic inventory systems. In particular, previous results indicate that demand variability causes a performance degradation in terms of inventory related costs when production capacity is unlimited. In order to investigate the effects of demand variability in capacitated production settings, we analyze a make-to-stock queue with general demand arrival times operated according to a base-stock policy. We show that when demand inter-arrival distributions are ordered in a stochastic sense, increased arrival time variability indeed leads to an augmentation of optimal base-stock levels and to a corresponding increase in optimal inventory related costs. We quantify these effects through several numerical examples.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Production/Inventory; Make-to-stock; Base-stock; Stochastic comparisons; GI/M/1

1. Introduction

We consider a single-item, single-stage production/inventory system operating in a make-to-stock mode. A plausible production control policy in this setting is a base-stock policy which drives the inventories to a predetermined base-stock (target inventory) level. Policy optimization, in order to minimize inventory holding and back-ordering costs for example, then reduces to the optimization of the base-stock level. Depending on the complexity of the underlying modeling assumptions, this optimization can be performed analytically, numerically or through simulation. While simulation or numerical analysis may enable a case-by-case comparison of different systems in terms of their optimal performance (inventory levels, costs etc.), it is impossible to state general structural properties through these approaches. In this paper, we pursue an analytical approach that leads to a structural comparison related to the variability of the demand inter-arrival times.

It is known that variability, in general, has a deteriorating effect on the performance of stochastic inventory systems. There are, however, relatively few papers that investigate variability

---

* Corresponding author. Tel.: +90-212-338-17-18; fax: +90-212-338-15-48.
E-mail addresses: jemai@lgi.ecp.fr (Z. Jemaï), fkaraesmen@ku.edu.tr (F. Karaesmen).

from an analytical point of view. Most of this research has focused on uncapacitated systems (exogenous lead times). Gerchak and Mossman [4] showed that, in a single-period newsvendor setting, the optimal replenishment quantity and the optimal cost are both increasing (under reasonable conditions) in the demand variability when demand is transformed using a mean-preserving transformation. Ridder et al. [10] presented comparison results based on demand variability in the identical setting emphasizing at the same time that depending on the definition of variability, some counter-examples can be found. For a continuous review single-item inventory system with exogenous lead times, Song [12,13] proved that increased lead time variability causes an increase in the optimal base-stock levels and the optimal costs. It is important to underline that all of these previous results hold under precise definitions of variability and that simple measures of variability such as "the coefficient of variation of lead time demand" may not suffice for an ordering of optimal base-stock levels or optimal costs.

Inter-arrival time variability has a negative effect on the performance of queueing systems as well (see for example, [1] for some analytical evidence through approximations and [8] for a numerical investigation). There are also some supporting numerical results for this negative effect on capacitated inventory systems – also called "production/inventory systems" – (e.g. for instance [9]). On the other hand, there are few purely analytical results on the effects of variability in capacitated systems. Such a result is presented in Güllü [6] where a single-item, periodic-review production/inventory problem under a base-stock policy is investigated and appropriate conditions on the demand distribution under which the optimal performance measures can be ordered are presented.

In this paper, we study a continuous review single-item single-stage make-to-stock type production system where demand inter-arrival times and processing times are random. Production capacity is explicitly modeled as a limited resource represented as the server of a queue (a detailed treatment of such models can be found in [1]). Unlike [6], our underlying system is a continuous review capacitated production/inventory system.

In order to capture some the effects of variability on the two key performance measures (base-stock levels and costs), the arrival process is modeled by a general renewal process. The system lends itself to almost-explicit analysis when processing times are exponentially distributed. The resulting model is a GI/M/1 make-to-stock queue.

Our contributions can be summarized as follows: we compare the optimal base-stock levels and optimal costs of two GI/M/1 make-to-stock queues with identical demand arrival and processing rates. We show that, if the demand inter-arrival times are ordered according to the (stochastic) convex order, then the optimal base-stock levels and the optimal average costs are ordered in the same direction. At the same time, our analysis indicates that, in either case, the convex order is essential for the results, and that weaker comparisons of variability (such as the coefficient of variation) do not suffice in general. Finally, we complement the theoretical results with a numerical investigation which enables us to quantify the effects of inter-arrival time variability.

The paper is structured as follows. In Section 2, we introduce the model, the employed notation, and some definitions that will be used later. Section 3 presents our main results on the effects of variability on optimal base-stock levels and optimal costs in GI/M/1 make-to-stock queues. A short numerical investigation is presented in Section 4 and the concluding remarks in Section 5.

2. Model and preliminaries

2.1. The model and notations

We consider a single-stage production system where demands arrive in single units (we discuss a special case related to batch arrivals later). Demand inter-arrival times are independent and identically distributed random variables. The production stage is modeled by a single server whose processing times are exponentially distributed. We denote by T the demand inter-arrival time, $\lambda = 1/E[T]$ the demand arrival rate, μ the (exponential) processing rate of the server and define $\rho = \lambda/\mu$.

The production system is controlled according to a base-stock policy with a base-stock (or target inventory) level S (see [2] for a detailed description in the context of production/inventory systems). The server produces whenever the inventory level is under the target level S and stops when the inventory level reaches S. We assume that demand is backlogged whenever inventory is not available and assume the standard cost structure: h is the inventory holding cost per part per unit time and b is the backorder cost per backorder per unit time. Let $X(t)$ denote the inventory level at time t, X be the corresponding stationary random variable and let $p_x = \text{Probability}\{X = x\}$. Under the above cost structure, the optimization problem is to select the base-stock parameter S which minimizes the expected average cost

$$\min_S \left\{ C(S) = \sum_{x=0}^{S} hxp_x + \sum_{x=-\infty}^{0} -bxp_x \right\}. \tag{1}$$

Let us now define $N(t) = S - X(t)$, the shortfall with respect to the base-stock level S at time t. $N(t)$ is the underlying queueing process in the production stage. In particular with general demand inter-arrival times and exponential processing times, the process $N(t)$ is equivalent to the number of customers at time t in a GI/M/1 queue. The analysis of the production/inventory system can then be performed through the corresponding queueing system. To this end, let $p_n = \text{Probability}\{N = n\}$, be the stationary probability that there are n customers in this queue. The objective function (1) can then be expressed as

$$\min_S \left\{ C(S) = \sum_{n=0}^{S} h(S-n)p_n + \sum_{n=S+1}^{\infty} b(n-S)p_n \right\}. \tag{2}$$

Minimizing the above expression with respect to S leads to the discrete version of the familiar critical-fractile formula for the optimal base-stock level. In particular, let F_N be the cumulative distribution function of N, then the optimal base-stock level S^* is given by (see [15] for example)

$$F_N(\widetilde{S}^* - 1) = \frac{b}{h+b} \quad \text{and} \quad S^* = \lfloor \widetilde{S}^* \rfloor, \tag{3}$$

where $\lfloor y \rfloor$ denotes the greatest integer that is less than or equal to y (a real number).

Note that in (3), $\widetilde{S}^*$ is the value where the first-order optimality condition (i.e. the first equation) is satisfied with equality. $\widetilde{S}^*$ itself is, in general, not an integer but the integer base-stock level is easily obtained from $\widetilde{S}^*$ by the second equation in (3). Since $\widetilde{S}^*$ subsumes all important qualitative characteristics of the system, we refer to it frequently in the rest of the paper as the *continuous approximation of the optimal base-stock level.*

2.2. Definitions and properties of stochastic comparisons

The principal tool of analysis in the rest of the paper will be stochastic comparisons of random variables. We provide below the definitions and properties of these comparisons that are used in the paper. These definitions and further details on stochastic comparison methods can be found in Stoyan [14] and Shaked and Shanthikumar [11].

Let X_1 and X_2 two random variables, F_1 and F_2 their cumulative distribution functions, f_1 and f_2 their probability density functions, and L_1, L_2 their Laplace transforms.

Definition 1 (*stochastic order*). The random variable X_1 is stochastically greater than a random variable X_2, denoted $X_1 \geqslant_{\text{st}} X_2$, if $1 - F_1(x) \geqslant 1 - F_2(x) \forall x$.

Definition 2 (*convex order*). For two random variables X_1 and X_2, $X_1 \geqslant_{\text{c}} X_2$ ($X_1 \geqslant_{\text{ic}} X_2$) if and only if $E[f(x_1)] \geqslant E[f(x_2)] \forall f$ convex (non-decreasing and convex).

Definition 3 (*Laplace transform order*). For two random variables X_1 and X_2, $X_1 \geqslant_L X_2$ if $E[\mathrm{e}^{-sX_1}] \leqslant E[\mathrm{e}^{-sX_2}]$. In what follows, we summarize some properties of the comparisons previously defined

$$X_1 \geqslant_{\text{st}} X_2 \Rightarrow X_1 \geqslant_{\text{ic}} X_2 \Rightarrow X_1 \geqslant_L X_2.$$

Note: In the increasing convex comparison, for two non-negative random variables having identical

means, the condition "non-decreasing" is not necessary.

Finally, in addition to the comparisons presented above, we frequently refer to a simple aggregate measure of variability: the coefficient of variation (CV) which is the ratio of the standard deviation to the mean of a random variable.

3. The influence of variability

3.1. The optimal base-stock level

Our objective in this section is to analyze the effects of demand inter-arrival variability on optimal base-stock levels. Song [12] studied uncapacitated systems where the corresponding variable of interest is the lead time demand. Song shows that under a so called "variability" ordering of the lead time demand, the optimal base-stock levels are ordered. We compare two GI/M/1 type make-to-stock queues that are identical except for their demand arrival processes with associated stationary inter-arrival time random variables A_1 and A_2. In order to isolate the effects of variability from those of utilization rate, we focus on the case where $E[A_1] = E[A_2]$ (the systems compared are then equivalent in their utilization rates). For a different (capacitated but periodic-review) model, Güllü [6] presents a comparison result which holds under a regular stochastic order (Definition 1). This order is, however, rather strong and does not allow, for instance, comparing two inter-arrival time distributions with the same mean.

Recall that the optimal base-stock level is obtained through the distribution function of the shortfall queue from the equation: $F_N(\widetilde{S}^* - 1) = b/(h + b)$, $S^* = \lfloor \widetilde{S}^* \rfloor$. Let us note that, for G/M/1 queues, the distribution function is a function of the parameter r, the root of the characteristic equation: $r = L_A((1 - r)\mu)$, where L_A is the Laplace transform of the inter-arrival time distribution F_A (see [5]). This leads to the following expression for the distribution function of N, the number of customers in the queue

$$F_N(x) = P[N \leqslant x] = (1 - \rho) + \rho(1 - r^x)$$
$$= 1 - \rho r^x \quad x = 0, 1, 2 \ldots \quad (4)$$

Lemma 1. *Consider two* GI/M/1 *queues with identical arrival and processing rates and with respective parameters* r_1 *and* r_2 *such that* $r_1 \geqslant r_2$, *then* $N_1 \geqslant_{st} N_2$ *where* N_1 *and* N_2 *are the number of customers in the queues* 1 *and* 2 *respectively.*

Proof. Let the function $G_x(r) = F(x) = (1 - \rho \times r^x)$, with parameter r defined on $(0, 1)$. $G_x(r)$ is decreasing and concave, then for all x, $r_1 \geqslant r_2 \Rightarrow G_x(r_1) \leqslant G_x(r_2)$. This implies: $F_1(x) \leqslant F_2(x) \Longleftrightarrow N_1 \geqslant_{st} N_2$. Consequently, $r_1 \geqslant r_2 \Rightarrow N_1 \geqslant_{st} N_2$. □

Lemma 2. *Consider two* GI/M/1 *make-to-stock queues such that* $N_1 \geqslant_{st} N_2$, *then* $S_1^* \geqslant S_2^*$ *where* S_1^* *(respectively* S_2^**) is the optimal base-stock level of queue* 1 (2).

Proof. The optimality condition is such that: $F_1(\widetilde{S}_1^* - 1) = F_2(\widetilde{S}_2^* - 1) = b/(h + b)$. By definition of a the stochastic order, $N_1 \geqslant_{st} N_2$ implies that $F_1(n) \leqslant F_2(n)$, for all n, which implies that $\widetilde{S}_1^* \geqslant \widetilde{S}_2^*$ and consequently that $S_1^* \geqslant S_2^*$. □

The next lemma is taken from Wolff [19].

Lemma 3. *Consider two* GI/M/1 *queues with identical service rates and with respective inter-arrival time random variables* A_1 *and* A_2 *such that* $E[A_1] = E[A_2]$ *and* $A_1 \leqslant_c A_2$, *then* $A_1 \geqslant_L A_2$ *and consequently* $r_1 \leqslant r_2$.

Putting together Lemmas 1–3, the following property is established for two GI/M/1 queues with respective inter-arrival times A_1 and A_2 such that $E[A_1] = E[A_2]$.

Proposition 1. *Consider two* GI/M/1 *make-to-stock queues with identical cost parameters, service rates, and with respective demand inter-arrival-times* A_1 *and* A_2 *such that* $E[A_1] = E[A_2]$ *and* $A_1 \leqslant_c A_2$, *then* $S_1^* \leqslant S_2^*$ *where* S_1^* *(respectively* S_2^**) is the optimal base-stock level of system* 1 *(system* 2).

Proof. Let r_1 and r_2 be the respective parameters of systems 1 and 2. Using Lemma 3 we have: $A_1 \leqslant_{ic} A_2 \Rightarrow r_1 \leqslant r_2$. Lemma 1 states that $r_1 \leqslant r_2 \Rightarrow N_1 \leqslant_{st} N_2$, and finally employing Lemma 2, we have the desired result. □

This result states the effect of the demand distribution on the base-stock level S^* via the parameter r. The optimal base-stock levels are increasing in with respect to the convex order of inter-arrival time distributions when all other parameters are held constant.

Remark. Note that, one can alternatively directly compute the optimal base-stock level using Eqs. (3) and (4). This leads to

$$S^* = \left\lfloor \frac{\log\left[r\left(1 - \frac{\alpha - (1-\rho)}{\rho}\right)\right]}{\log r} \right\rfloor, \qquad (5)$$

where $\alpha = b/(b+h)$. The effect of the coefficient r on S^* can also be inferred from expression (5): S^* is increasing in r. However, this direct computation does not give any insights on the underlying conditions. For instance, we can observe from Lemma 2 that the desired result imposes a strong stochastic order condition on the variable N. The key point is that the inter-arrival time comparison should induce a stochastic order of the queue lengths (a rather strong and non-trivial condition). Unfortunately, there are few existing comparison results for the queue length process based on weaker orders – such as the convex order – of arrival or service processes (see [18] for some known cases). This implies that immediate extension of the above approach to more general make-to-stock queues is difficult.

3.2. The optimal cost

The previous section focused on the relationship between the base-stock level and the demand variability. In this section, we study the influence of this variability on the optimal cost. For that, we start by pointing out the expression of the cost as well as the results previously obtained.

Let us note that, in uncapacitated systems, corresponding comparison results are usually expressed in terms of the demand distribution (for single-period models) and in terms of the distribution of lead-time demand (infinite-horizon models). Song [12] shows that, for an (infinite-horizon) uncapacitated continuous-time system, the ordering of the demand during lead time (or the ordering of the lead time itself) induces an ordering of the corresponding costs: $D_1 \leqslant_{\mathrm{ic}} D_2 \Rightarrow C_1(S_1^*) \geqslant C_2(S_2^*)$. Ridder et al. [10] use a weaker condition called "2_variability" (see [10]), showing for a single-period model that $D_1 \geqslant_2 D_2 \Rightarrow C_1(S_1^*) \leqslant C_2(S_2^*)$ (where D is the demand random variable).

As in the corresponding uncapacitated model [12], the comparison of the optimal cost function is more delicate than the comparison of optimal base-stock levels. In particular, our main result will require a continuous relaxation of the base-stock level. This is a frequently made assumption (see [12] or [13]) in the literature. To outline the procedure, recall from Section 2.1 that, the optimal cost function is given by Eq. (2) where the base-stock level is taken to be $S^* = \left\lfloor \widetilde{S}^* \right\rfloor$. We ignore the integrality correction temporarily and first focus on the continuous variable $\widetilde{S}^*$. Based on Proposition 1, we can then obtain the following lemma:

Lemma 4. *Let $\widetilde{S}^*$ be the continuous approximation of the optimal base-stock level (see the remark following Eq. (3)), the corresponding optimal cost (defined in Eq. (2)) $C(\widetilde{S}^*)$ is equal to $h(\widetilde{S}^* - ((\rho - r)/(1 - r)))$.*

Proof. Using the explicit form of the stationary queue length distribution, we can express the expected cost as a function of the base-stock level as

$$C(S) = h\left(S - \rho - \rho r \frac{1 - r^{S-2}}{1 - r} - \rho r^{S-1}\right) + b\left(\rho \frac{r^S}{1 - r}\right). \qquad (6)$$

The proof follows by a direct insertion of $\widetilde{S}^*$ in the above cost function. $\square$

It now follows from Proposition 1 and Lemma 4 that, for GI/M/1-type make-to-stock queues, the parameter r induces an order on the optimal cost.

Proposition 2. *Consider two GI/M/1 make-to-stock queues with identical service rates and cost parameters, and with inter-arrival-times A_1 and A_2 such that $E[A_1] = E[A_2]$ and $A_1 \leqslant_{\mathrm{c}} A_2$, then $C_1(\widetilde{S}_1^*) \leqslant C_2(\widetilde{S}_2^*)$, where $C_1(\widetilde{S}_1^*)$ (respectively $C_2(\widetilde{S}_2^*)$) is the optimal cost under the base-stock policy of system 1 (respectively system 2).*

Proof. Referring to Lemmas 3, 2, and 1, we obtain: $A_1 \leqslant_{\mathrm{c}} A_2 \Rightarrow r_1 \leqslant r_2$ (by Lemma 3), $r_1 \leqslant r_2 \Rightarrow N_1 \leqslant_{\mathrm{st}} N_2$ (by Lemma 1) and $N_1 \leqslant_{\mathrm{st}} N_2 \Rightarrow \widetilde{S}_1^* \leqslant \widetilde{S}_2^*$ (by Lemma 2). Finally, by virtue of Lemma 4, $\widetilde{S}_1^* \leqslant \widetilde{S}_2^*$ implies that $C_1(\widetilde{S}_1^*) \leqslant C_2(\widetilde{S}_2^*)$. $\quad\square$

3.3. Comparisons of some commonly used arrival processes

Below, we present some probability distributions that are frequently used in modeling inventory and queueing processes. Stochastic convex order results for some of these distributions are also presented. Whenever this type of order is available, the ordering of optimal inventory levels and costs follow directly from Propositions 1 and 2. Detailed definitions of the distributions and their parameters can be found in Appendix A.

3.3.1. Gamma/Weibull distributions

For certain frequently used probability distributions, convex stochastic order has been established in terms of the parameters of the distribution. The gamma distribution frequently used in queueing applications is such a case. The following result is taken from Stoyan [14]. Consider two gamma distributions $G_1(\lambda, \alpha, x)$ and $G_2(\mu, \beta, x)$ with respective densities $g_1(x)$ and $g_2(x)$ (see Appendix A), if $\alpha > \beta$ and $\alpha/\lambda \leqslant \beta/\mu$, then $G_1 \leqslant_{\mathrm{ic}} G_2$. Propositions 1 and 2 then imply the following: for two GI/M/1 make-to-stock queues with identical service rates and cost parameters, and with inter-arrival-times A_1 (with distribution $G_1(\lambda, \alpha, x)$) and A_2 (with distribution $G_2(\mu, \beta, x)$) such that $E[A_1] = E[A_2]$, if $\alpha > \beta$ and $\alpha/\lambda \leqslant \beta/\mu$ we have the following ordering of the optimal base-stock levels and the optimal costs: $S_1^* \leqslant S_2^*$ and $C_1(\widetilde{S}_1^*) \leqslant C_2(\widetilde{S}_2^*)$.

As an important special case of the above result, if we have two Erlang-distributed random variables, A_1 and A_2 with identical means and respectively with k_1 (k_2) stages such that $k_1 > k_2$ then $A_1 \leqslant_{\mathrm{c}} A_2$. Propositions 1 and 2 then lead to the following result: for two GI/M/1 make-to-stock queues with identical service rates and cost parameters, and with inter-arrival-times A_1 (with an Erlang distribution of k_1 stages) and A_2 (with an Erlang distribution of k_2 stages) such that $E[A_1] = E[A_2]$, if $k_1 > k_2$ we have the following ordering of the optimal base-stock levels and the optimal costs: $S_1^* \leqslant S_2^*$ and $C_1(\widetilde{S}_1^*) \leqslant C_2(\widetilde{S}_2^*)$.

Weibull distributions are frequently used in reliability/maintenance applications and are pertinent for spare parts inventory management. From [14] we have the following comparison result: let two Weibull distributions $W_1(\lambda, \alpha, x)$ and $W_2(\mu, \beta, x)$ with respective density functions $f_1(x)$ and $f_2(x)$ (see Appendix A) and respective means m_{f_1} and m_{f_2}, if $\alpha > \beta$ and $m_{f_1} \leqslant m_{f_2}$, then $W_1 \leqslant_{\mathrm{ic}} W_2$. Using Propositions 1 and 2, we then have then following result: consider two GI/M/1 make-to-stock queues with identical service rates and cost parameters, and with inter-arrival-times A_1 (with distribution $W_1(\lambda, \alpha, x)$) and A_2 (with distribution $W_2(\mu, \beta, x)$) such that $E[A_1] = E[A_2]$. If $\alpha > \beta$, then: $S_1^* \leqslant S_2^*$ and $C_1(\widetilde{S}_1^*) \leqslant C_2(\widetilde{S}_2^*)$.

3.3.2. Erlang distributions with unidentical stages

In this section, we consider Erlang distributions consisting of k different stages with different means. This class of distributions can cover coefficients of variation ranging between $1/\sqrt{k}$ and 1. Because of the rational form of Laplace transform $L_A(S) = \prod_{i=1}^{k} (\lambda_i/(\lambda_i + S))$ (with λ_i the rate of stage i), the calculation of r for the Erlang distributions with k stages amounts to solving a $(k+1)$th degree equation.

Even though numerical analysis is relatively easy, explicit results of stochastic comparisons do not seem to exist for this class of distributions. Hereon, we concentrate on two-stage generalized-Erlang distributions and thus cover CV's ranging from $1/\sqrt{2}$ to 1. In this case, the calculation of r requires solving a third degree equation (see Appendix B).

It can be verified that for two Erlang distributions with the same mean and different coefficients of variation such that $\mathrm{CV}_1 \geqslant \mathrm{CV}_2$, we have $r_1 \geqslant r_2$.

Using Propositions 1 and 2 (along with Lemmas 1 and 2) then leads to on ordering of the optimal base-stock levels and costs in the corresponding GI/M/1 make-to-stock queues. In addition, when A_1 (an Erlang random variable) is compared with A_2 (an exponential random variable) having the identical mean, we obtain: $r_1 \leqslant r_2$ (in fact, it is known that $r_2 = \rho$, see [19]). Using Propositions 1 and 2, it is immediately seen that an Erlang inter-arrival time distribution induces lower optimal base-stock levels and costs than an exponential inter-arrival time distribution with the same mean (in the GI/M/1 make-to-stock queue setting).

3.3.3. Two-stage hyper-exponential distributions

Two-stage hyper-exponential distributions cover the domain $CV \geqslant 1$ and are frequently used to model high-variability arrival processes. For this class of distributions the parameter r can be explicitly obtained.

The Laplace transform of a H_2 distributed random variable with parameter q and rates λ_1 and λ_2 (see Appendix A) is

$$L_A(s) = \frac{q\lambda_1}{\lambda_1 + s} + \frac{(1-q)\lambda_2}{\lambda_2 + s},$$

therefore

$$r = \frac{q\lambda_1}{\lambda_1 + \mu(1-r)} + \frac{(1-q)\lambda_2}{\lambda_2 + \mu(1-r)}$$

leading to

$$r = 0.5 + \frac{\lambda_1 + \lambda_2}{2\mu}$$
$$- \sqrt{0.25 + \left(\frac{\lambda_1 - \lambda_2}{2\mu}\right)^2 + (q - 0.5)\left(\frac{\lambda_2 - \lambda_1}{\mu}\right)}$$

with the stability condition: $q/\rho_1 + (1-q)/\rho_2 > 1$.

It can easily be verified that, that r_1 of an H_2/M/1 queue is always greater than r_2 ($= \rho$) of a corresponding M/M/1 queue. Using Propositions 1 and 2, we can then establish that the optimal base-stock levels and the costs are higher in H_2/M/1 make-to-stock queues than in a corresponding M/M/1 make-to-stock queue (with the identical mean inter-arrival time).

3.3.4. Modeling batch arrivals: The general-exponential distribution

For high-variability ($CV \geqslant 1$) arrival processes the general-exponential distribution (see Appendix A for a precise definition) constitutes a modeling tool which covers all ranges of the coefficient of variation.

A useful feature of this distribution is that as a model, it is equivalent to a batch-arrival process where batches arrive according a Poisson distribution with rate λ and where the batch size X is geometrically distributed with parameter q.

The Laplace transform of a *GE* distribution of parameter q and rate λ is $L_A(s) = 1 - q + q\lambda/(\lambda + s)$, therefore:

$$r = 1 - q + \frac{q\lambda}{\lambda + \mu(1-r)} \quad \text{then} \quad r = \frac{\lambda}{\mu} + 1 - q$$

with the stability condition $q > \lambda/\mu$.

The comparison of an M/M/1 and a GE/M/1 with the same mean yields that r_1 (of the GE/M/1) is always greater than r_2 ($= \rho$) of the corresponding M/M/1 system. Therefore, by Propositions 1 and 2 a batch-arrival demand process requires a higher optimal base-stock level and generates higher costs than the unit-arrival demand process with the identical arrival rate.

Similarly, the comparison of two GE/M/1 queues with the same mean $q_1/\lambda_1 = q_2/\lambda_2$ and with different CV's (where $CV_1 \geqslant CV_2$) implies that $r_1 \geqslant r_2$, thereby leading to an ordering to of the optimal costs and the base-stock levels.

4. Numerical examples

In this section, we investigate some numerical examples of different GI/M/1 make-to-stock queues in order to quantify the effects of variability. Our theoretical results in the previous sections are based on a precise definition of variability that stems from the convex stochastic order. Because this order is not easily quantifiable, we present the numerical results based on a simple aggregate measure of variability: the coefficient of variation (CV). It is important to note that, as a

comparison, the ordering of CVs is weaker than the convex order (in fact, it is implied by the convex order for identical means). This allows us to numerically verify whether the convex order condition can be relaxed.

Our investigation then consists of studying S^* and $C(S^*)$ as functions of the coefficient of variation for given distributions. For this purpose, we take a fixed value of ρ and compute the parameters of the different inter-arrival time distributions in order to obtain the same average arrival rate. We then compute the parameter r and the optimal base-stock level S^* and the associated cost $C(S^*)$ using formulas (5) and (6). We then plot these values as a function of the coefficient of variation of the inter-arrival time distribution. Appendix B outlines the procedure that is used to modify the CV for different inter-arrival time distributions.

The first set of results is based on two-stage generalized-Erlang distributions. Fig. 1 depicts the variation of optimal base-stock levels and costs as a function of the coefficient of variation in

Fig. 1. S^* and $C(S^*)$ as a function of CV for $\rho = 0.9$, $b = 10$.

$Er(\lambda_1, \lambda_2)/M/1$ make-to-stock queues with $\rho = 0.9$, $b = 10$.

The optimal cost as a function of CV shows that $C(S^*)$ increasing in the coefficient of variation even though there are discontinuities due to the discrete nature of S^*. These discontinuities are more significant when the backlog cost increases. These first results demonstrate that optimal base-stock levels and optimal costs are increasing in the CV of the inter-arrival time distribution. On the other hand, our analytical results require the stochastic convex order definition of variability which is much stricter than a simple CV order. The question then is: are there cases where the simple CV order fails? This question will be investigated in the next example.

As a second example, let us investigate GE/M/1 and H_2/M/1 make-to-stock queues (both inter-arrival time distributions have CV's greater than 1). Indeed, within each class the base-stock level and the associated cost increase as a function of the coefficient of variation as shown in Fig. 2 for $\rho = 0.9$, $h = 1$, $b = 10$.

We can observe from Fig. 2 that the optimal base-stock level is almost a linear function of the CV when GE and H_2 distributions are considered separately.

More interestingly however, note that optimal base-stock levels and the optimal costs increase faster in GE/M/1 make-to-stock systems than in corresponding H_2/M/1 systems as shown in Fig. 3.

Fig. 3 underlines the limitations of comparisons based only on the coefficient of variation. For identical values of CV, a higher base-stock level is required for GE/M/1 systems than in H_2/M/1 systems. This difference becomes more pro-
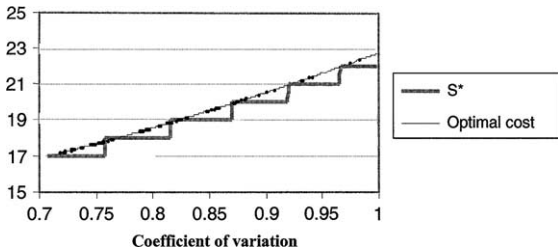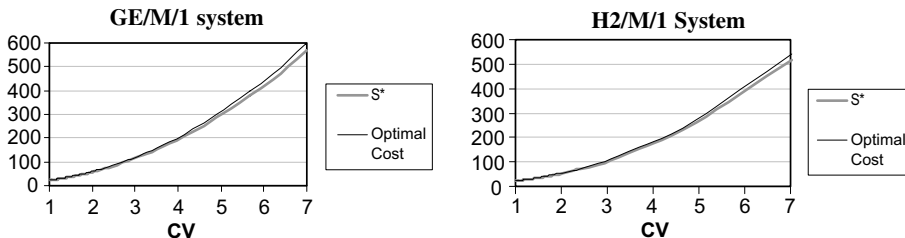

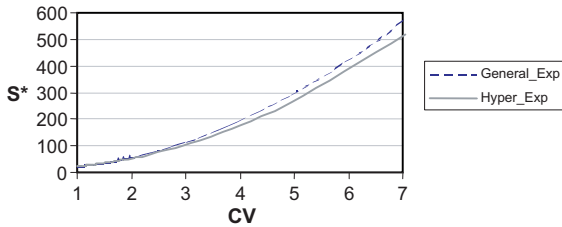Fig. 2. S^* and $C(S^*)$ as a function of the CV for $\rho = 0.9$, $h = 1$, $b = 10$.

Fig. 3. Comparison between the optimal base-stock levels of the GE/M/1 and the H_2/M/1 make-to-stock queues ($\rho = 0.9$, $h = 1$, $b = 10$).

nounced as the coefficient of variation increases (for instance when CV > 3). For instance, the H_2/M/1 system with a CV of 4 has a lower optimal base-stock level than a GE/M/1 system with a CV of 3.9. Obviously, an increased coefficient of variation alone does not lead to increased base-stock levels in this case. Furthermore, by Proposition 2, the same arguments apply to the optimal costs: an H_2/M/1 system with a higher coefficient of variation can have lower optimal costs than a less variable (in terms of CV) GE/M/1 system. In other words, as in Ridder et al. [10], increased demand variability (in terms of coefficient of variation) can lead to lower base-stock levels and to lower costs in some cases.

Fig. 4 explains why CV alone cannot suffice, in general, to compare optimal base-stock levels and costs. Going back to Lemma 1 (which then leads to Propositions 1 and 2), the key comparison parameter is r (a higher value of r leads to higher (in a non-strict sense) optimal base-stock levels and costs for the same ρ). As the respective CVs
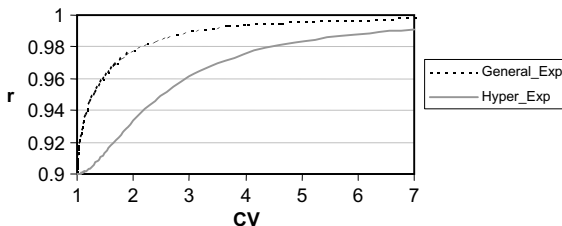
are varied according to the rule explained in Appendix A, it can be seen from Fig. 4 that the GE distribution always has a higher r value for the identical CV level. By Propositions 1 and 2, it follows then that for the same CV, the GE inter-arrival time distribution will generate higher optimal base-stock levels and optimal costs. The difference between the optimal base-stock levels of the two systems (observed in Fig. 3) becomes especially pronounced as the CV increases. This can be explained as follows: from Fig. 4, it can be seen that as the CV increases both r values approach 1 while the r value of the GE distribution continues to stay above that of the HE distribution. From Eq. (5), it is known that the optimal base-stock level is very sensitive to small changes in r when r is close to 1 (note that the denominator of Eq. 5, $\log(r)$, approaches 0 as r approaches 1). For high CVs (greater than 3.5 in Fig. 4), the relatively small differences in the respective r parameters translate into significant differences in the optimal base-stock levels. Unfortunately a general relationship between the CV and the parameter r does not seem to exist (r is the root of a non-linear equation related to the Laplace transform of the inter-arrival time distribution).

5. Conclusion

The degrading effects of variability on the performance of production and inventory systems are well known. We attempted to provide a precise and general description of the effects of variability for make-to-stock queues. Our investigation here is limited to GI/M/1-type systems. A parallel technical note [7] extends – approximately – some of these results to M/G/1 and G/G/1 type systems. However, even the analysis of these special cases underline the difficulty of obtaining general conditions for more complicated systems. Results on increasing optimal base-stock levels and costs require very strong stochastic order relationships on queue length distributions as a function of inter-arrival (or processing) time distributions, which may not hold under very general circumstances. On the other hand, a couple of general conclusions

Fig. 4. The r parameters of the GE/M/1 and the H_2/M/1 make-to-stock queues as a function of the inter-arrival time coefficient of variation (for $\rho = 0.9$).

can be extracted from our analysis. First, the coefficient of variation alone is not a sufficient measure of variability for ordering base-stock levels and optimal costs in general. In certain cases, increased coefficient of variation can lead to decreased inventories and costs. Second, the convex order is a valuable condition which guarantees the ordering of optimal costs and base-stock levels. Both our analytical and numerical results indicate that production/inventory systems that have the same average load can behave completely differently depending on the second order characteristics of the underlying processes. This implies that careful modeling of underlying demand and production processes is critical in order to capture finer properties of these systems.

This paper was limited to the analysis of a single-class make-to-stock queue. The analysis of multi-class make-to-stock queues pose several additional challenges like the scheduling of production and the allocation of inventories and is an on-going investigation (see [16,17]). A recent paper by Benjaafar and Kim [3] generalizes some of the results in this paper to a multi-class GI/M/1 make-to-stock queue (under first come first served order scheduling). It would be interesting to verify whether such multi-class results can be extended to more complicated scheduling/allocation policies as in [16] or [17].

Acknowledgements

Appendix A

This appendix summarizes the parameters of the probability distributions used as well as the approach used to vary the coefficient of variation as a function of the parameters.

Distribution	Density function	CV variation
Gamma $G(\lambda, \alpha)$	$f(x) = \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}$	
Weibull $W(\lambda, \alpha)$	$f(x) = \lambda \alpha x^{\alpha-1} e^{-\lambda x^\alpha}$	
Erlang $\text{Er}(\lambda_1, \lambda_2, \ldots, \lambda_n)$	$f(x) = \frac{\left(\sum_i \lambda_i\right)^n x^{n-1}}{(n-1)!} \exp\left(-\sum_i \lambda_i x\right)$	In order to vary the coefficient of variation while keeping the same average of the $\text{Er}(\lambda_1, \lambda_2)$ distributions, we vary λ_1 and calculate corresponding λ_2
$\text{HE}(q_i, \lambda_i)$	$f(x) = \sum_i q_i \lambda_i e^{-\lambda_i x}$	For $\text{HE}(q_1, q_2, \lambda_1, \lambda_2)$ distributions, we vary q_1 and calculate the other parameters of the distribution
$\text{GE}(q, \lambda)$	$f(0) = 1 - q; \ f(x) = q \lambda e^{-\lambda x}, \ x > 0$	For the $\text{GE}(q, \lambda)$ distributions, we vary q and calculate the corresponding λ

Appendix B. Calculation of r for an Erlang distribution with two stages

The Laplace transform of an Erlang distribution with K different stages of rate λ_i is

$$L_A(S) = \prod_{i=1}^{k} \frac{\lambda_i}{\lambda_i + S}.$$

Recall that $r = L_A((1 - r) \times \mu)$, then for the two-stage Erlang we have

$$r = \frac{\lambda_1}{\lambda_1 + \mu(1 - r)} \frac{\lambda_2}{\lambda_2 + \mu(1 - r)}$$

and finally

$$r = \frac{1}{2} + \frac{\rho_1 + \rho_2}{2} - \sqrt{\left(\frac{1 + \rho_1 + \rho_2}{2}\right)^2 - \rho_1 \rho_2},$$

with $\rho_1 \rho_2 / (\rho_1 + \rho_2) < 1$ the stability condition.

Consider two Erlang distributions with the same mean $(\lambda_1 + \lambda_2)/\lambda_1 \lambda_2 = (\lambda_1' + \lambda_2')/\lambda_1' \lambda_2'$ and different coefficients of variation $CV_1 \geqslant CV_2$ $(\lambda_1 \lambda_2 \leqslant \lambda_1' \lambda_2'$ et $\lambda_1 + \lambda_2 \leqslant \lambda_1' + \lambda_2')$, then we have $r_1 \geqslant r_2$.

The calculation leads us to investigate a function f of the form

$$f(x_1, x_2, y_1, y_2) = \frac{x_1 - y_1}{2} + \sqrt{\left(\frac{1 + y_1}{2}\right)^2 - y_2}$$
$$- \sqrt{\left(\frac{1 + x_1}{2}\right)^2 - x_2}$$

on the domain: $x_1/x_2 = y_1/y_2 < 1$; $x_1 > y_1$ and $x_2 > y_2$. We verify numerically that f is always positive on its domain and consequently that $r_1 \geqslant r_2$.

References

[1] J.A. Buzacott, J.G. Shanthikumar, Stochastic Models of Manufacturing Systems, Prentice Hall, New Jersey, 1993.
[2] J.A. Buzacott, S.M. Price, J.G. Shanthikumar, Service levels in multi-stage MRP and base stock control systems, in: T. Fandel, T. Gulledge, A. Jones (Eds.), New Directions for Operations Research in Manufacturing Systems, Springer, New York, 1993, pp. 445–463.
[3] S. Benjaafar, J. Kim, When does higher demand variability lead to lower safety stocks? Working Paper, University of Minnesota, 2002.
[4] Y. Gerchak, D. Mossman, On the effect of demand randomness on inventories and costs, Operations Research 40 (1992) 804–807.
[5] D. Gross, C.M. Harris, Fundamentals of Queuing Theory, Wiley Series in Probability and Mathematical Statistics USA, 1983.
[6] R. Güllü, Base stock policies for production/inventory problems with uncertain capacity levels, European Journal of Operational Research 105 (1998) 43–51.
[7] Z. Jemai, F. Karaesmen, A note on the influence of variability in make-to-stock queues, Technical Report, Ecole Centrale Paris and Koç University, 2003.
[8] F. Karaesmen, S.M. Gupta, The finite capacity GI/M/1 queue with server vacations, Journal of the Operational Research Society 47 (1996) 817–828.
[9] F. Karaesmen, J.A. Buzacott, Y. Dallery, Integrating advance order information in make-to-stock production systems, IIE Transactions 34 (8) (2002) 649–662.
[10] A. Ridder, E. Van Der Laan, M. Solomon, How larger demand variability may lead to lower costs in the newsvendor problem, Operations Research 46 (6) (1998) 934–936.
[11] M. Shaked, J.G. Shanthikumar, Stochastic Orders and their Applications, Academic Press, London, 1994.
[12] J.S. Song, The effect of lead time uncertainty in a simple stochastic inventory model, Management Science 40 (5) (1994) 603–613.
[13] J.S. Song, Understanding the lead-time effects in stochastic inventory systems with discounted costs, Operations Research Letters 15 (1994) 85–93.
[14] D. Stoyan, Comparison Methods for Queues and Other Stochastic Models, John Wiley and Sons, Berlin, 1983.
[15] M. Veatch, L.M. Wein, Scheduling a make-to-stock queue: Index policies and hedging points, Operations Research 44 (1996) 634–647.
[16] F. de Véricourt, F. Karaesmen, Y. Dallery, Dynamic scheduling in a make-to-stock system: A partial characterization of optimal policies, Operations Research 48 (2000) 811–819.
[17] F. de Véricourt, F. Karaesmen, Y. Dallery, Optimal stock allocation for a capacitated supply system, Management Science 48 (2002) 1486–1501.
[18] W. Whitt, Comparing counting processes and queues, Advances in Applied Probability 13 (1981) 207–220.
[19] R.W. Wolff, Stochastic Modelling and the Theory of Queues, Prentice-Hall International Series in Industrial and Systems Engineering, New Jersey, 1989.