Available online at www.sciencedirect.com

SCIENCE DIRECT

Int. J. Production Economics 96 (2005) 213–232

international journal of production economics

www.elsevier.com/locate/dsw

Tradeoffs between base stock levels, numbers of kanbans, and planned supply lead times in production/inventory systems with advance demand information

George Liberopoulos*, Stelios Koukoumialos

Department of Mechanical and Industrial Engineering, University of Thessaly, Volos, Greece

Received 12 March 2002; accepted 27 April 2004

Available online 8 June 2004

Abstract

We numerically investigate tradeoffs between near-optimal base stock levels, numbers of kanbans, and planned supply lead times in base stock policies and hybrid base stock/kanban policies with advance demand information used for the control of multi-stage production/inventory systems. We report simulation-based computational experience regarding such tradeoffs and the managerial insights behind them for single-stage and two-stage production/inventory systems.

Keywords: Production/inventory control; Advance demand information; Planned supply lead time; Base stock; Kanban

1. Introduction

Recent developments in information technology and the emphasis on supply chain system integration have significantly reduced the cost of obtaining end-item *advance demand information*—henceforth referred to as ADI—in the form of actual orders, order commitments, forecasts, etc., and diffusing it among all stages of the system. This has created opportunities for developing effective production/inventory control policies that exploit such information. The implementation of such policies may result in significant cost savings in the entire system through inventory reductions and improvements in customer service (Bourland et al., 1996; Buzacott and Shanthikumar, 1994; Chen, 2001; DeCroix and Mookerjee, 1997; Gallego and Özer, 2001; Gilbert and Ballou, 1999; Güllü, 1996; Hariharan and Zipkin, 1995; Karaesmen et al., 2003, 2004; Milgrom and Roberts, 1988; Van Donselaar et al., 2001; Wijngaard, 2004).

In this paper, we investigate policies that use ADI for production/inventory control of a multi-stage serial system that produces a single type of parts in a make-to-stock mode. We make the following specific assumptions. Every stage in the system consists of a facility, where unfinished parts

*Corresponding author. Tel.: +30-24210-74056.

E-mail addresses: glib@mie.uth.gr (G. Liberopoulos), skoukoum@mie.uth.gr (S. Koukoumialos).

0925-5273/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpe.2004.04.006

are processed, and an output store, where finished parts are stored. Parts in the facility are referred to as *work-in-process* (WIP), and parts in the output store are referred to as *finished goods* (FG). FG of the last stage are referred to as *end-items*. There is an infinite supply of raw parts feeding the first stage. Customer demands arrive randomly for one end-item at a time, with a constant *demand lead time* in advance of their due dates. Once a customer demand arrives, it cannot be cancelled, i.e. the ADI is assumed to be perfect. Demands that cannot be satisfied on their due dates are backordered and are referred to as *backordered demands* (BD). The arrival of a customer demand for an end-item triggers the placement of a production order to replenish FG inventory at every stage. FG inventory levels are followed continuously at all stages, and replenishment production orders may be placed at any time. There is no setup cost or setup time for placing a production order and no limit on the number of orders that can be placed per unit time. Under the above assumptions, there is no incentive to replenish FG inventory by anything other than a continuous review, one-for-one replenishment policy. The model described above is simple but captures some of the basic operational elements of a serial, capacitated production/inventory system.

When there is no ADI, demand due dates coincide with demand arrival times. In this case, the replenishment production orders at every stage, which are triggered by the arrival of customer demands, may be placed only at or after the demand due dates. The simplest production/inventory control policy, in case there is no ADI, is the base stock policy. The base stock policy was originally developed for non-capacitated inventory systems and is better suited for such systems (see, e.g., Zipkin, 2000). In the base stock policy, a production order to replenish FG inventory at each stage is placed and authorized to be released as soon as a customer demand arrives to the system. A policy that has attracted considerable attention and is particularly suited for a JIT capacitated manufacturing environment is the kanban policy (see, e.g., Buzacott and Shanthikumar, 1993). In the kanban policy, a production order to replenish FG inventory at a particular stage is placed and authorized to be released only when a finished part of this stage is consumed by its downstream stage. In the case of a single-stage system, the kanban policy is equivalent to a make-to-stock CONWIP policy (Spearman et al., 1990). Base stock and kanban policies may be combined to form more sophisticated, hybrid, base stock/kanban policies, such as the generalized kanban policy (Buzacott, 1989; Buzacott and Shanthikumar, 1993; Zipkin, 1989, 2000) and the extended kanban policy (Dallery and Liberopoulos, 2000). In the generalized kanban policy, a production order to replenish FG inventory at a particular stage is placed only when the inventory in this stage is below a given inventory-cap level. In the extended kanban policy, a production order to replenish FG inventory at each stage is placed as soon as customer demand arrives to the system; however, it is authorized to be released into a particular stage only when the inventory in this stage is below a given inventory-cap level. A detailed description of these and other policies can be found in Liberopoulos and Tsikis (2003) and Liberopoulos and Dallery (2000, 2003). There exist several studies that analyze and compare different production/inventory control policies in the case where there is no ADI. Some of the methods used for the performance evaluation and comparison of these policies are stochastic ordering arguments (e.g., Spearman, 1992), dynamic programming (e.g., Karaesmen and Dallery, 2000; Veach and Wein, 1994), queuing theory (e.g., Rubio and Wein, 1996), approximate analysis (e.g., Duri et al., 2000; Frein et al., 1995; Zipkin, 2000, Section 8.8.2), and simulation (e.g., Bonvik et al., 1997).

When there is ADI, the production orders to replenish FG inventory at every stage, which are triggered by the arrival of a customer demand, may be placed before the due date of the demand. Base stock and hybrid base stock/kanban policies can be easily modified to take advantage of ADI by offsetting the due date of each demand by a constant *planned supply lead time* at each particular stage in order to determine the time of placing the resulting production replenishment order at this stage, as is done in the time-phasing step of the MRP procedure. The planned supply lead time of each stage is a fixed parameter of the control policy, which, in an MRP system, is typically set so as to guarantee that the actual flow time (a random variable) of a part through the facility of this stage falls within the planned supply lead time a certain percentage of time (e.g., 95% of the time) (Karaesmen et al., 2002). The kanban policy cannot exploit ADI, because in the kanban policy a production order is placed after a part in FG inventory is consumed and therefore at (or after) the due date of the demand that triggered it. When ADI is available, it is therefore reasonable to consider only base stock and hybrid base stock/kanban policies and not pure kanban policies. Hybrid base stock/kanban policies with ADI are of particular interest because they fuse together reorder-point policies, JIT, and MRP, three widely practiced approaches for controlling the flow of material in multi-stage production/inventory systems.

The aim of our investigation in this paper is to reveal tradeoffs between optimal base stock levels, numbers of kanbans, and planned supply lead times in multi-stage base stock and hybrid base stock/kanban policies with ADI. Some of the more specific issues that we will address are the following.

In both base stock and hybrid base stock/kanban policies with ADI, the base stock level of FG inventory represents FG that have been produced before any demands have arrived to the system in order to satisfy the expected demand during the supply lead time and protect the system against possible stockouts. Intuitively, there should be a tradeoff between the demand lead time and the optimal base stock level of FG inventory at each stage. Namely, as the demand lead time increases, the optimal base stock level should decrease or at least stay the same. But is there a structure to this tradeoff? More specifically, does the optimal base stock level decrease at a constant rate or at a diminishing rate, and if so, until which point, as the demand lead time increases? Do the optimal base stock levels at different stages all decrease at the same time until they drop to zero or to some constant level, or do they decrease one after the other in a certain order as the demand lead time increases? If the latter is true, as the demand lead time increases, is it more beneficial to first lower the base stock level of upstream stages, where FG inventory is usually less expensive to hold but also less important with respect to customer service, or of downstream stages, where FG inventory is usually more expensive to hold but also more important with respect to customer service?

The planned supply lead times are control parameters that determine how much (if any) to delay the placement of production replenishment orders that are triggered by the arrival of customer demands. Intuitively, if the demand lead time is short, the placement of production replenishment orders should not be delayed, whereas if the demand lead time is long, the placement of production replenishment orders should be delayed. But what is the maximum critical demand lead time below which the placement of production replenishment orders should not be delayed? Does it make sense to delay the placement of production replenishment orders and at the same time have positive base stock levels at some stages?

In hybrid base stock/kanban policies, the number of kanbans at each stage sets an inventory-cap and determines the production capacity of this stage. Intuitively, there should be a tradeoff between the number of kanbans and the base stock level at each stage. Namely, as the number of kanbans decreases, the production capacity should decrease, the production replenishment time should increase, and consequently the base stock level of FG inventory should also increase. But what is the optimal number of kanbans and therefore the optimal base stock level, and how are they affected by the demand lead time? More specifically, as the demand lead time increases, should the optimal number of kanbans decrease, and if so, by how much?

Since exact analytical tools for evaluating the performance of multi-stage base stock and hybrid base stock/kanban policies with ADI are limited and approximation-based analytical tools may yield fairly accurate but systematically biased results, which may be misleading when trying to reveal tradeoffs between parameters, we use simulation and brute-force optimization to investigate such tradeoffs and report the results of this investigation for single-stage and two-stage systems. The main contribution of this paper is the managerial insights that these results bring to light.

The rest of the paper is organized as follows. In Sections 2 and 3, we numerically investigate single-stage base stock and hybrid base stock/kanban policies, respectively. In Sections 4 and 5, we numerically investigate two-stage base stock and hybrid base stock/kanban policies, respectively. Finally, in Section 6, we draw conclusions.

2. Single-stage base stock policy with ADI

In this section, we consider a single-stage base stock policy with ADI, similar to the system considered in Karaesmen et al. (2002). Customer demands arrive for one end-item at a time according to a Poisson process with rate λ, with a constant demand lead time, T, in advance of their due dates. The arrival of every customer demand eventually triggers the consumption of an end-item from FG inventory and the placement of a production order to the facility of the sole stage to replenish FG inventory. More specifically, the consumption of an end-item from FG inventory is triggered T time units after the arrival time of the demand. If no end-items are available at that time, the demand is backordered.

The control policy depends on two design parameters, namely, the target or base stock level of end-items in FG inventory, denoted by S, and the stage planned supply lead time, denoted by L. We should point out that we use the word ‘‘base stock’’ for terminological simplicity but with caution, because in the presence of ADI, the inventory position can actually exceed the base stock level. In fact, Hariharan and Zipkin (1995) and Chen (2001) use the term ‘‘order base stock’’ instead of ‘‘base stock’’ to describe the target FG inventory in the presence of ADI. The stage planned supply lead time L has the same meaning as the fixed lead time parameter in an MRP system.

Initially, the system starts with a base stock of S end-items in FG inventory. The time of placing the replenishment production order triggered by the arrival of a demand is determined by offsetting the demand due date by the stage planned supply lead time, L, as is done in the time-phasing step of the MRP procedure. This means that the order is placed immediately, i.e. with no delay, if $L \geqslant T$ (in this case, the order is already late), or with a delay equal to $T - L$ with respect to the demand arrival time, if $L < T$. In other words, the delay in placing an order is equal to $\max(0, T - L)$. When the order is placed, a new part is immediately released into the facility. If there is no ADI, i.e. if $T = 0$, both the consumption of an end-item from FG inventory and the replenishment production order are triggered at the demand arrival time, and the resulting policy is the classical base stock policy. A queuing network model of the base stock policy with ADI is shown in Fig. 1.

The symbolism used in Fig. 1 (and all other similar figures that follow in the rest of the paper) is the same as that used in Dallery and Liberopoulos (2000), Duri et al. (2000), Karaesmen et al. (2002), Frein et al. (1995), Liberopoulos and Tsikis (2003), and Liberopoulos and Dallery (2000, 2003) and has the following interpretation. The oval represents the facility, and the circles represent constant time delays. The queues followed by vertical bars represent *synchronization stations*. A synchronization station is a server with zero service time that instantaneously serves customers as soon as there is at least one customer in each of the queues that it synchronizes. Queues are labeled according to their content, and their initial value is indicated inside a parenthesis. Queue OH stands for replenishment *orders on hold*. In the single-stage base stock policy, this queue is always equal to zero, because of the assumption that there is an infinite number of raw parts. Recall that BD stands for *backordered demands*.

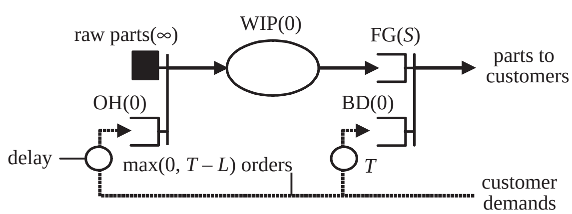

Fig. 1. Single-stage base stock policy with ADI.

We consider a classical optimization problem whose objective is to find the values of S and L that minimize the long-run expected average cost of holding and backordering inventory,

$$C(S,L) = hE[\text{WIP} + \text{FG}(S,L)] + bE[\text{BD}(S,L)], \tag{1}$$

where h is the unit cost of holding WIP + FG inventory per unit time and b is the unit cost of backordering FG inventory per unit time. Here, we are more interested in the effect of ADI on inventory holding and backordering costs than on the cost of ADI itself, so we assume that there is no cost of obtaining ADI. The effect of buying ADI is considered in Karaesmen et al. (2003). It is not difficult to see that control parameters S and L affect only the expected average FG and BD and not the expected average WIP. We explicitly express these dependencies in the cost function (1). In what follows, we will study the above optimization problem, first for the case where there is no ADI and then for the case where there is ADI.

2.1. The case where there is no ADI

If there is no ADI, i.e. if $T = 0$, the planned supply lead time L is irrelevant, because a replenishment order is always placed immediately upon the arrival of a customer demand. After some algebraic manipulations, the long-run expected average cost (1) can be expressed as a function of S as follows:

$$C(S) = (h+b)\left\{E[\text{WIP}] - \sum_{n=0}^{S} nP(\text{WIP}=n) + S[P(\text{WIP} \leqslant S) - b/(b+h)]\right\}. \tag{2}$$

Moreover, the optimal base stock level, S^*, is given by the well-known critical fraction rule of the newsvendor problem, i.e. it is the smallest integer that satisfies (see Rubio and Wein, 1996)

$$P(\text{WIP} \leqslant S^*) \geqslant b/(b+h). \tag{3}$$

If the facility consists of a single-server station with exponential service rate μ, the long-run expected average cost (excluding the cost of WIP, which is equal to $h\rho/(1-\rho)$ and is therefore independent of the design parameters S and L) is given by (see Buzacott and Shanthikumar, 1993, Section 4.3.1; Rubio and Wein, 1996)

$$C(S) = h[S - \rho(1-\rho^S)/(1-\rho)] + b[\rho^{S+1}/(1-\rho)],$$

where $\rho = \lambda/\mu$. In this case, $S^* = \lfloor \hat{S} \rfloor$, where $\hat{S} = \ln[h/(h+b)]/\ln\rho$ and $\lfloor x \rfloor$ denotes the largest integer which is smaller than or equal to x.

If the facility consists of a Jackson network of servers, S^* satisfies a non-closed-form expression that can be solved numerically. For instance, in the case of a balanced Jackson network consisting of M identical single-server stations, each server having an exponential service rate μ, S^* is the smallest integer that satisfies (3), where the WIP has a negative binomial steady-state distribution given by (see Rubio and Wein, 1996)

$$P(\text{WIP}=n) = \binom{M+n-1}{n}(1-\rho)^M\rho^n, \tag{4}$$

where $\rho = \lambda/\mu$.

2.2. The case where there is ADI

If there is ADI, i.e. if $T > 0$, there is a time lag between placing an order and demanding an end-item from FG inventory. This time lag is equal to $T - \max(0, T-L) = \min(T,L)$ (see Fig. 1). This implies that any system with demand lead time $T > L$ behaves exactly like a system with demand lead time $T = L$.

2.2.1. The case of a single-server exponential station

If the facility consists of a single-server station with exponential service rate μ, the long-run expected average cost (excluding the cost of WIP, which is equal to $h\rho/(1-\rho)$) is given by (see Buzacott and Shanthikumar, 1993, Section 4.5.2; Buzacott and Shanthikumar, 1994; Karaesmen et al., 2004)

$$C(S,L) = h[S + \lambda\min(L,T) - \rho/(1-\rho)] + (h+b)[\rho^{S+1}/(1-\rho)]e^{-\mu(1-\rho)\min(L,T)},$$

where $\rho = \lambda/\mu$. One can optimize $C(S,L)$ with respect to parameters S and L to gain insight into the behavior of the system under the optimal parameters. Specifically, it is shown in Karaesmen et al. (2004) that for a fixed L, the optimal-based stock level, $S^*(L)$, is given by

$$S^*(L) = \begin{cases} \lfloor \hat{S}(L) \rfloor & \text{if } L \leqslant L^*, \\ 0 & \text{if } L \geqslant L^*, \end{cases}$$

where

$$\hat{S}(L) = \ln[(h+b)/h]/\ln \rho) - [(\mu - 1)/\ln \rho]L$$

and the optimal planned supply lead time, L^*, is given by

$$L^* = \ln[(h+b)/h]/(\mu - \lambda).$$

The overall optimal base stock, $S^* = S^*(L^*)$, is then equal to the integer $\lfloor \hat{S} \rfloor$, where

$$\hat{S} = \max\{0, \ln[(h+b)/h]/\ln \rho - [(\mu - \lambda)/\ln \rho]T\}.$$

The above analysis implies that L^* is independent of T and is equal to $cE[W]$, where W is the waiting (or flow) time of a part in the facility if the system were operated in make-to-order mode (recall that for an $M/M/1$ queue, $E[W] = 1/(\mu - \lambda)$), and c is a factor equal to $\ln[(h+b)/h]$. $\hat{S}$ and consequently S^*, on the other hand, are functions of T. More specifically, $\hat{S}$ decreases linearly with T and reaches zero at $T = L^*$. Thus, for demand lead times T, such that $T < L^*$, $\hat{S} > 0$ and production orders are placed upon the arrival of demands with no delay. For demand lead times T, such that $T > L^*$, however, $\hat{S} = 0$ and production orders are placed upon the arrival of demands with a delay of $T - L^*$. This means that for $T > L^*$, the optimal operation mode of the system switches from "make-to-stock" to "make-to-order". The minimum long-run expected average cost $C(S^*, L^*)$ decreases with T and attains its minimum value at $T = L^*$. Since L^* is the smallest value of T for which $\hat{S} = 0$, and $S^* = \lfloor \hat{S} \rfloor$, it follows that the smallest value of T for which $S^* = 0$, is slightly smaller than L^*.

2.2.2. *The case of a Jackson network of servers*

If the facility consists of a Jackson network of servers, there are no general analytical results available for the optimal parameter values. Intuitively, we would expect that as T increases, the optimal base stock level should decrease, as is the case with the single-server station. The question is how exactly does it decrease? Does it decrease linearly until it drops to zero, as in the case of a single-server station, or does it decrease in some sort of non-linear way (e.g., in a diminishing way)? What is the smallest value of T, for which the optimal base stock level becomes zero, indicating a switch in the optimal operation mode of the system switches from make-to-stock to make-to-order? Is it equal to the average flow time of a part through the facility multiplied by the factor $c = \ln[(h+b)/h]$, as in the case of a single-server station?

The only general analytical result related to the latter question is Proposition 1 in Karaesmen et al. (2004). That proposition states that if a supply system operates in a make-to-order mode (i.e., with zero base stock level) and satisfies Assumption 1, given below, then the optimal planned supply lead time, L^*, is the smallest real number that satisfies

$$P(W \leqslant L^*) \geqslant b/(b+h), \tag{5}$$

where W is the order replenishment time, i.e. the waiting or flow time of a part in the facility.

Assumption 1. All replenishment orders enter the supply system one at the time, remain in the system until they are fulfilled (there is no blocking, balking or reneging), leave one at a time in the order of arrival (FIFO) and do not affect the flow time of previous replenishment orders (lack of anticipation).

The implication of the above result is that if the system in Fig. 1 satisfies Assumption 1, then when $T \geqslant L^*$, where L^* is given by (5), the optimal operation mode of the system switches from make-to-stock to make-to-order with optimal planned supply lead time L^*. Notice the similarity between expressions (3) and (5). These two expressions demonstrate explicitly the interchangeability of safety stock, which is related to the base stock level, and safety time, which is related to the planned supply lead time.

To summarize, when $T = 0$, S^* is given by (3), and when $T \geqslant L^*$, $S^* = 0$. A question that remains unanswered is what happens when $0 < T < L^*$? To shed some light into this issue, we numerically investigated a particular but representative instance of the system, in which the facility consists of a Jackson network of four identical single-server stations in series, each server having a mean service time $1/\mu$. Thus, the system instance that we considered satisfies Assumption 1. For this instance, we considered four sets of system parameter values shown in Table 1. In cases 1 and 2, the service time distribution of each machine is exponential, whereas in cases 3 and 4, it is Erlang with two phases. The optimal control parameter values for the four cases are as follows.

For $T=0$, L is irrelevant, and S^* can be determined from (3). In cases 1 and 2, $P(\text{WIP}=n)$ can be computed analytically from (4) for $M=4$. S^* can then be substituted into (2) to determine $C(S^*)$. The results are: $S^*=8$ and $C(S^*)=90.8954$, for case 1, and $S^*=68$ and $C(S^*)=83.6966$, for case 2. In cases 3 and 4, the optimal base stock level S^* was obtained by evaluating the cost for different values of S using simulation and picking the value that yielded the lowest cost. The results are: $S^*=4$ and $C(S^*)=48.07865$, for case 3, and $S^*=34$ and $C(S^*)=42.83576$, for case 4.

We should point out that when we refer to simulation-based results, we use the word "optimal" for terminological simplicity but with caution, because we did not actually perform a significance test on the sign of the cost difference between two systems having different values of S. Using the wording "near-optimal" instead of "optimal" to describe the simulation-based results would be more accurate, but might make the text heavier. Before discussing the simulation-based results, let us first say a few words about the simulation experiments.

For this and for all the other examples that follow in the rest of the paper we ran a total of roughly 8000 simulation experiments using the simulation software Arena. In each experiment, we used a simulation run length of 60 million time units. This yielded 95% confidence intervals on the estimated values of E[WIP], E[FG], and E[BD] with half-width values of less than 0.5% of their respective estimated values, in the cases of E[WIP] and E[FG], and less than 4%, in the case of E[BD]. The simulation-based results are discussed next.

Table 1
Parameter values for cases 1–4 of the single-stage base stock policy with ADI

Case	$1/\lambda$	Service time distribution	$1/\mu$	$\rho=\lambda/\mu$	h	b
1	1.25	Exponential	1.0	0.8	5	1
2	1.1	Exponential	1.0	0.90909…	1	9
3	1.25	Erlang-2	1.0	0.8	5	1
4	1.1	Erlang-2	1.0	0.90909…	1	9

In all cases, the optimal planned supply lead time L^* can be determined from (5). In cases 1 and 2, it is well-known that the distribution of the order replenishment time W is Erlang with M phases and mean $M/(\mu-\lambda)$, so it can be computed analytically. This is because W is the sum of M i.i.d. $M/M/1$-system waiting times, each time having an exponential distribution with mean $1/(\mu-\lambda)$. Specifically, the cumulative distribution of W is given by

$$P(W \leqslant w) = 1 - \sum_{k=0}^{M-1} \frac{[(\mu-\lambda)w]^k}{k!} e^{-(\mu-\lambda)w}.$$

Substituting the above expression into (5) yields $L^*=10.6396$ and 73.4886, for cases 1 and 2, respectively. Incidentally, a quick computation of $cE[W]$ yields $\ln[(5+1)/5]/[4/(1-0.8)]=3.6464$ in case 1 and $\ln[(1+9)/1]/[4/(1-0.90909)]=101.3137$ in case 2. Therefore, in either case, $L^* \neq cE[W]$ (recall that in the case of a single-server exponential station, $L^*=cE[W]$). In fact, in case 1, L^* $(=10.6396) > cE[W]$ $(=3.6464)$, whereas in case 2, L^* $(=73.4886) < cE[W]$ $(=101.3137)$. Therefore, the fact that for the single-server exponential server case, $L^*=cE[W]$, does not hold in general. In cases 3 and 4, L^* was obtained by evaluating the cost for different integer values of L using simulation and picking the value that yielded the lowest cost. The result is $L^*=6$ and 35 for the two cases, respectively.

For values of T in the interval (0, L^*), we used simulation to evaluate the cost of the system for the four sets of parameter values. In each case, we optimized the control parameter S for different values of T, using exhaustive search. In this paper, we only present the optimal results due to space considerations. For all four sets of parameter values shown in Table 1, the optimization yielded the following general results.

As T increases from zero, the optimal base stock level S^* appears to decrease linearly with T and reaches zero just below $T = L^*$, as in the case of the single-server station. The insight behind this behavior is that there appears to be a linear tradeoff between S^* and T and that L^* is just above the smallest value of T for which S^* is equal to zero. The results are shown in Table 2 for the four cases. Plots of S^* versus T are shown in Fig. 2 for the four cases.

From Fig. 2, it can be seen that the smallest values of T for which $S^* = 0$ are approximately equal to 10, 73, 6, and 35, for cases 1–4, respectively. We use the word "approximately" because we only examined integer values of T, whereas T really is a continuous parameter. Recall, that for cases 1 and 2, the analytically obtained optimal planned supply lead times L^* are 10.6396 and 73.4886, respectively. As in the case of the single-server station, the optimal planned supply lead times, L^*, are independent of T.

From Table 2, it can be seen that the minimum long-run expected average cost $C(S^*, L^*)$ decreases very little with T and attains its minimum value at $T = L^*$. The drop in $C(S^*, L^*)$ between the situations where $T = 0$ and $T = L^*$ is only

Table 2
S^* and $C(S^*, L^*)$ versus T, for $L = L^*$, for the single-stage base stock policy with ADI

Case 1			Case 2			Case 3			Case 4		
T	S^*	$C(S^*, L^*)$	T	S^*	$C(S^*, L^*)$	T	S^*	$C(S^*, L^*)$	T	S^*	$C(S^*, L^*)$
0	8	90.8954	0	68	83.6966	0	4	48.0787	0	34	42.8358
2	6	90.5209	10	59	83.3837	2	3	47.6811	10	24	42.3458
4	5	90.3351	20	50	83.0256	4	1	47.2363	20	16	41.6722
6	3	90.0959	30	40	82.7439	6	0	46.8434	30	5	41.3582
8	2	89.9054	40	31	82.3999	∞	0	46.8434	35	0	41.1147
10	0	89.6463	50	22	82.1246				∞	0	41.1147
∞	0	89.6463	60	12	81.8376						
			73	0	81.6226						
			∞	0	81.6226						

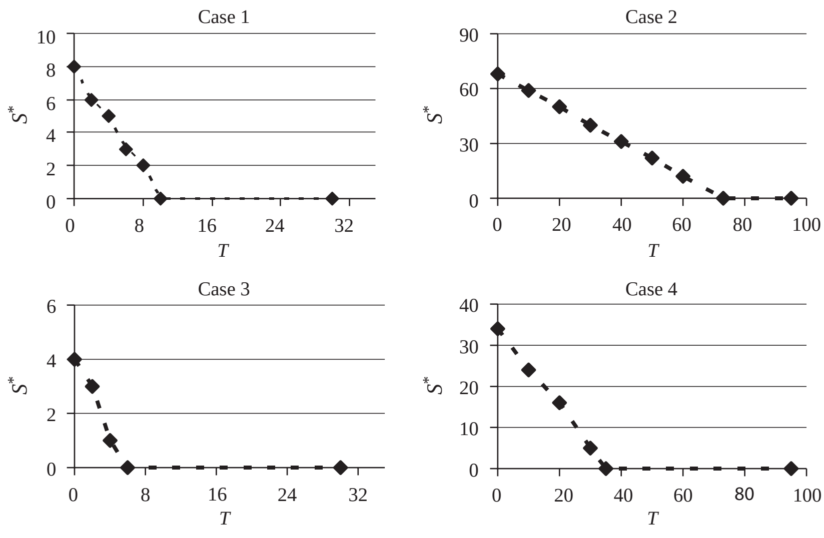

Fig. 2. S^* versus T, for $L = L^*$, for the single-stage base stock policy with ADI.

1.37%, 2.48%, 2.29%, and 4.50%, for cases 1–4, respectively. This insensitivity of the long-run expected average cost (1) with respect to the demand lead time T is to a certain extent due to the fact that a significant part of this cost is due to the term hE[WIP], which is independent of T. If we omit this term from the long-run expected average cost, then the drop in $C(S^*, L^*)$ between the situations where $T = 0$ and $T = L^*$ is 10.06%, 4.38%, 20.91%, and 7.09%, for cases 1–4, respectively. To summarize, the basic insights behind the results are the following.

In a production/inventory system operating under the single-stage base stock policy: (a) There appears to be a linear tradeoff between the demand lead time and the optimal base stock level. (b) The optimal planned supply lead time appears to be the smallest demand lead time for which the optimal base stock level is zero. This means that if the demand lead time is smaller than the optimal planned supply lead time, the optimal base stock level is positive and a production replenishment order is placed immediately after the arrival of the customer demand that triggered it. If the demand lead time is greater than the optimal planned supply lead time, the optimal base stock level is zero and a production replenishment order is placed after the arrival of the customer demand that triggered it with a delay that is equal to the difference between the demand lead time and the planned supply lead time.

3. Single-stage hybrid base stock/kanban policy with ADI

The single-stage hybrid base stock/kanban policy with ADI behaves exactly like the single-stage base stock policy with ADI as far as the placement of replenishment production orders is concerned. The difference is that in the single-stage hybrid base stock/kanban policy, when a replenishment production order is placed, it is not immediately authorized to be released into the facility (as is the case in the base stock policy) unless the inventory in the facility, i.e. WIP, or in the entire system, i.e. WIP + FG, is below a given *inventory-cap* level.

Setting an inventory-cap in any section of a production/inventory system makes sense if this section and/or the section downstream of it have limited processing capacity. This is because releasing a part in an already congested section of the system with limited processing capacity, or in a section without limited processing capacity (e.g., a buffer) but which is followed by a section with limited processing capacity, will increase the inventory in that section with little or no decrease in the part's completion time. In the kind of multi-stage serial systems that we study in this paper, where each stage consists of a facility containing WIP and an output store containing FG inventory, all facilities have limited processing capacity, and all output buffers, except the output buffer of the last stage, are followed by facilities which have limited processing capacity. In such systems, therefore, it makes sense to set inventory-caps on the (WIP + FG) inventory of all stages except the last one and to set an inventory-cap on the WIP of the last stage. In the case of a single-stage system considered in this section, the one and only stage is the last stage; therefore, for a single-stage system we will only consider a base stock/kanban policy where a *WIP-cap* is set on the WIP of the stage.

With the above discussion in mind, in the single-stage hybrid base stock/kanban policy, when a replenishment production order is placed, it is not immediately authorized to be released in the facility unless the WIP in the system is below a given WIP-cap of K parts. If the WIP in the system is at or above K, the order is put on hold until the WIP drops below K (the inventory drops as parts exit the facility). Once the order is authorized to go through, a new part is immediately released into the facility. This policy can be implemented by requiring that every part entering the facility be granted a production authorization card or *kanban*, where the total number of kanbans is equal to the WIP-cap level. Once a part leaves the facility, the kanban that was granted (and attached) to it is detached and is used to authorize the release of a new part into the facility. Notice that the single-stage hybrid base stock/kanban policy with no ADI is equivalent to the single-stage generalized kanban policy (Buzacott, 1989; Zipkin, 1989).

In a single-stage hybrid base stock/kanban policy with ADI, the system starts with a base stock of S end-items in FG inventory and K free kanbans. These kanbans are available to authorize an equal number of replenishment production orders. The number of free kanbans represents the number of parts that can be released into the facility before the WIP in the system reaches the WIP-cap level K. A queuing network model of the hybrid base stock/kanban policy with ADI is shown in Fig. 3, where queue FK contains *free kanbans*.

From Fig. 3, it can be seen that kanbans trace a loop within a closed network linking FK and WIP. The constant population of this closed network is K, i.e. at all times, $\text{FK} + \text{WIP} = K$. The throughput of this closed network, denoted by TH_K, depends on K and determines the processing capacity of the system, i.e. the maximum demand rate λ that the system can meet in the long run. Under some fairly general conditions (that essentially require that the facility exhibits "max-plus" behavior in the sense that the timings of events in the system can be expressed as functions of the timings of other events involving the operators "max" and "+" only), TH_K is an increasing concave function of K, such that $\text{TH}_0 = 0$ and $\text{TH}_\infty < \infty$. For every feasible demand rate λ, such that $\lambda < \text{TH}_\infty$, there is a finite minimum value of K, say $K_{\min}$, such that for any $K \geqslant K_{\min}$, $\text{TH}_K > \lambda$, which means that the system has enough capacity to meet demand in the long run.

The single-stage hybrid base stock/kanban policy includes the single-stage base stock and kanban policies as special cases. Namely, the single-stage hybrid base stock/kanban policy with $K = \infty$ and $S < \infty$ is equivalent to the single-stage base stock policy with base stock level S. The single-stage hybrid base stock/kanban policy with $K = S < \infty$ is equivalent to the single-stage kanban policy or equivalently to a make-to-stock CONWIP policy with K (or equivalently, S) kanbans (Di Mascolo et al., 1996).

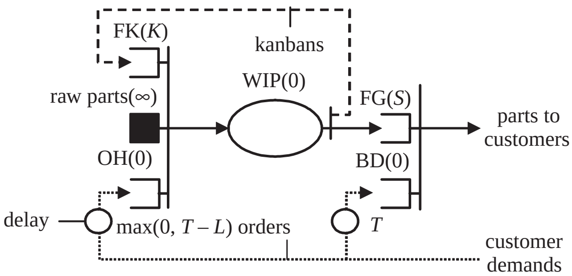

Fig. 3. Single-stage hybrid base stock/kanban policy with ADI.

We consider an optimization problem similar to that considered in Section 2, whose objective is to find the values of K, S, and L that minimize the long-run expected average cost of holding and backordering inventory,

$$C(K,S,L) = hE[\text{WIP}_K + \text{FG}_K(S,L)] + bE[\text{BD}_K(S,L)], \quad (6)$$

where h and b are defined as in Section 2. It is not difficult to see that control parameters S and L affect only the expected average FG and BD and not the expected average WIP or OH, whereas parameter K affects the expected average FG, BD as well as WIP and OH. We explicitly expressed these dependencies in the cost function (6).

3.1. The case where there is no ADI

If there is no ADI, i.e. if $T = 0$, the planned supply lead time L is irrelevant, and the optimal base stock level for any given value of K, such that $K \geqslant K_{\min}$, S_K^*, is the smallest integer that satisfies (see Liberopoulos and Dallery, 2002)

$$P(\text{OH}_K + \text{WIP}_K \leqslant S_K^*) \geqslant b/(b+h). \quad (7)$$

If the facility consists of a Jackson network of servers, there is no analytical expression (not even in non-closed form) to determine the steady-state distribution of OH_K and WIP_K and therefore S^*, and only approximation methods exist (e.g., see Frein et al., 1995). To shed some light into this case, we numerically investigated the same instance of the system that we investigated in Section 2.2, i.e. an instance in which the facility consists of a Jackson network of four identical single-server stations in series, for the same four sets of parameter values shown in Table 1. For cases 1 and 2, TH_K can be calculated analytically as $\text{TH}_K = \mu/[1 + (M-1)/K]$, where $M = 4$ (Frein et al., 1995). Since $K_{\min}$ is the smallest integer for which $\text{TH}_{K_{\min}} > \lambda$, it follows that $K_{\min}$ is the smallest integer that satisfies $\mu/[1 + (M -$

$1)/K_{\min}] > \lambda$, i.e. $K_{\min} > (M-1)\rho/(1-\rho)$, where $\rho = \lambda/\mu$. This implies that $K_{\min}$ is equal to 13 and 30, for cases 1 and 2, respectively.

We used simulation to evaluate the long-run expected average cost of the system for the four sets of parameter values. In each case, we found the optimal base stock levels for different values of K, S^*_K, using exhaustive search. The results are shown in Table 3. Plots of S^*_K versus K are shown in Fig. 4 for the four cases. The values for $K = \infty$ are taken from Table 2 for the case where $T = 0$.

From Table 3 and Fig. 4, it can be seen that the optimal base stock level S^*_K is non-increasing in K, i.e. $S^*_{K+1} \leqslant S^*_K$, for $K \geqslant K_{\min}$. Moreover, there exists a finite critical value of K, K_c, such that $S^*_K = S^*_\infty$, for $K \geqslant K_c$, where S^*_∞ is the optimal base stock level for the same system operating under the pure base stock policy, i.e. a hybrid base stock/kanban policy with $K = \infty$. This means that there is a tradeoff between K and S^*_K and that this tradeoff holds for up to a finite critical value of K, K_c. This critical value is equal to 17, 69, 7, and 23, for cases 1–4, respectively. The same result is proven analytically in Liberopoulos and Dallery (2002) for a similar system with a slightly different objective, namely, minimize the long-run expected

Table 3
S^*_K and $C(S^*_K, L^*)$ versus K for the single-stage hybrid base stock/kanban policy with no ADI

Case 1			Case 2			Case 3			Case 4		
K	S^*_K	$C(K, S^*_K)$	K	S^*_K	$C(K, S^*_K)$	K	S^*_K	$C(K, S^*_K)$	K	S^*_K	$C(K, S^*_K)$
13	15	104.7284	33	227	246.6715	6	5	40.4173	13	54	57.4847
14	10	88.4036	40	96	109.8939	7	4	38.3957	15	43	47.4977
15	11	84.0617	45	81	94.3872	8	4	39.3074	20	36	41.9549
16	9	82.9463	50	75	88.3950	9	4	40.6576	22	35	41.4670
17	8	82.7963	60	71	84.4438	10	4	41.9764	23	34	41.1375
18	8	83.1151	68	69	83.6047	11	4	43.0968	24	34	41.1463
∞	8	90.8954	69	68	82.8599	∞	4	48.0787	25	34	41.3125
			70	68	82.9287				∞	34	42.8358
			∞	68	83.6966						

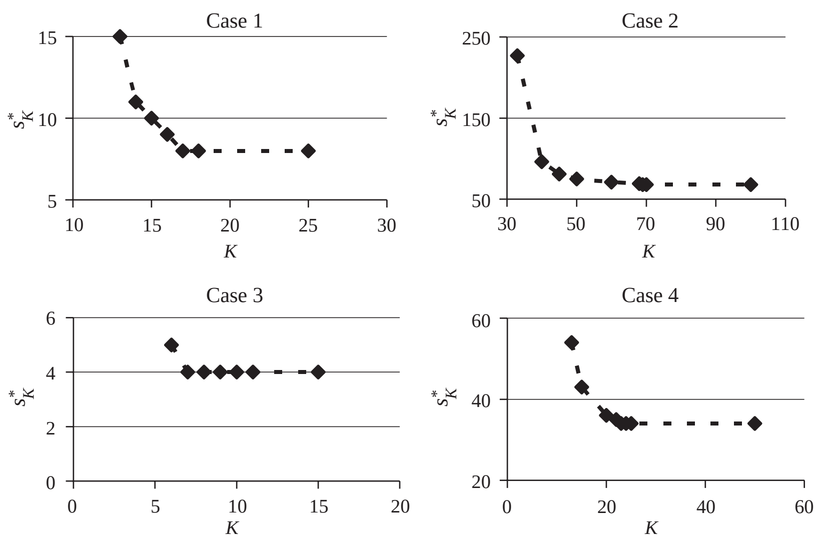

Fig. 4. S^*_K and $C(S^*_K, L^*)$ versus K for the single-stage hybrid base stock/kanban policy with no ADI.

average cost of holding inventory subject to a specified fill rate constraint. The insight behind this result is the following.

As K increases, parts are released in the facility earlier and depart from it earlier, causing the average FG inventory to increase too. At the same time, the congestion in the system also increases, and therefore parts stay in the facility longer. This implies that the rate of increase of the average FG inventory is diminishing in K. At $K = K_c$, the facility reaches a critical congestion level. That is, for values of K below K_c, the system is *under-congested* in the sense that increasing K causes an increase in the average FG inventory that is enough to cause a further decrease in S_K^*. For values of K at or above K_c, however, the system is *over-congested* in the sense that increasing K does not cause an increase in the average FG inventory that is enough to cause a further decrease in S_K^*. An important question that remains to be answered is what is the overall optimal number of kanbans K^* and the resulting optimal base stock level $S_{K^*}^*$?

The most striking result of the optimization is that in all four cases, the overall optimal number of kanbans, K^*, is equal to K_c, and therefore the overall optimal base stock level, S^*, is equal to S_∞^*. More specifically, K^* is equal to 17, 69, 7, and 23, for cases 1–4, respectively, and S^* is equal to 8, 68, 4, and 34, for cases 1–4, respectively. This is not an obvious result. The insight behind it is that the optimal base stock level of the hybrid base stock/kanban policy, S^*, appears to be equal to the optimal base stock level of the pure base stock policy, S_∞^*, i.e. the smallest possible value of S_K^*. Moreover, the optimal number of kanbans, K^*, is the smallest value of K for which $S_K^* = S_\infty^*$. In other words, it is optimal to set K to a value that is just big enough so that the corresponding optimal base stock level is equal to the optimal base stock level of the pure base stock policy, S_∞^*. This means that the pure base stock policy is not optimal but the optimal base stock level of the pure base stock policy is also optimal for the hybrid base stock/kanban policy. Computational experience reported in Duri et al. (2000), Karaesmen and Dallery (2000), and Zipkin (2000, Section 8.8.2) for simpler single-machine systems also confirms this result. The difficulty in proving it stems from the fact that no analytical expression for the steady-state distribution of OH_K and WIP_K exists, except for a trivial system where the facility consists of a single-server station with exponential service rate μ, in which case $K_{min} = K_c = 1$. Nevertheless, an indication of the validity of this result is given in Liberopoulos and Dallery (2002).

The simulation-based results also suggest that the long-run expected average cost increases more steeply to the left of K^* than to the right of K^*. This means that it is more costly to underestimate K relatively to the optimal value K^* than to overestimate it. Naturally, as $K \to \infty$, the long-run expected average cost approaches $C(\infty, S_\infty^*)$, i.e. the minimum cost of the pure base stock policy. In our numerical examples, the minimum long-run expected average cost for the optimal base stock policy is 90.8954, 83.6966, 48.0787, and 42.8358, for cases 1–4, respectively, as is seen in Table 2, whereas, the minimum long-run expected average cost for the optimal hybrid base stock/kanban policy is 82.7963, 82.8599, 38.3957, and 41.1375, for cases 1–4, respectively, as is seen in Table 3. This means that the minimum long-run expected average cost is 8.91%, 1%, 20.14%, and 3.96% smaller under the optimal hybrid base stock/kanban policy than it is under the optimal base stock policy, for cases 1–4, respectively. The fact that the reduction in cost is more dramatic in case 1 than in case 2 (and similarly in case 3 than in case 4) is due to two reasons. The first reason is that the cost ratio h/b is higher in case 1 than in case 2; therefore, reducing the average WIP with a WIP-cap mechanism is more effective in case 1 than in case 2, since in case 1, every part in WIP costs relatively more. The second reason is that the utilization coefficient, ρ, is higher in case 2 than in case 1. This implies that in case 2, the distribution of the inter-departure times from the facility is less sensitive to the distribution of the inter-arrival times to the facility than it is in case 1. It further implies that in case 2, S_K^* and $C(K, S_K^*)$ are less sensitive to K than they are in case 1. Finally, the fact that the reduction in cost is more dramatic in case 3 than in case 1 (and similarly in case 4 than in case 2) implies that imposing a WIP-cap mechanism is more effective in a system with lower flow time variability.

3.2. The case where there is ADI

If there is ADI, i.e. if $T>0$, and the facility consists of a Jackson network of servers, there are no analytical results for the optimal parameter values. Intuitively, we would expect that as T increases, the optimal base stock level of the hybrid base stock/kanban policy should decrease. The question is how exactly does it decrease, in particular with respect to the optimal base stock level of the pure base stock policy? Also, does the optimal number of kanbans decrease too?

To shed some light into this case, we numerically investigated the same instance of the system that we investigated in Sections 2.2 and 3.1, i.e. an instance in which the facility consists of a Jackson network of four identical single-server stations in series, for the same four sets of parameter values shown in Table 1. We used simulation to evaluate the long-run expected average cost of the system for the four cases, and in each case we optimized the control parameters, K, S, and L for different values of T using exhaustive search. The results are shown in Table 4 for selected values of K around the optimal values and $L=L^*$.

From the results in Table 4, it appears that in all cases, the optimal number of kanbans, K^*, is equal to K_c for all values of T, i.e. K^* appears to be independent of T. Namely, K^* is equal to 17, 69, 7, and 23, for cases 1–4, respectively. Moreover, L^* and S^* have the same values as in the single-stage base stock policy with ADI discussed in Section 2.2. Namely, L^* is approximately equal to 10, 73, 6, and 35, for cases 1–4, respectively, and S^* has the same values as those shown in Table 2. This is not an obvious result. The insight behind it is the following.

When $T=0$, $S^*>0$. When $S^*>0$, it appears that it is optimal to place a replenishment production order immediately upon the arrival of a customer demand to the system, irrespectively of the value of T (as long as T is small enough so that $S^*>0$). Whenever a replenishment production order is placed immediately upon the arrival of a customer demand to the system, T does not affect what goes on in the facility but only affects FG and BD. More specifically, T is a tradeoff for S^*, where S^* also affects only FG and BD. Therefore, the value of K that determines the optimal processing capacity and congestion level in the facility when $T=0$, namely K_c, is also optimal when $T>0$.

The results in Section 3.1 showed that for $T=0$, S^* is equal to the optimal base stock level of the pure base stock policy, S^*_∞. For $T>0$, it appears that the tradeoff that exists between T and S^* in the hybrid base stock/kanban policy is exactly the same as the tradeoff between T and S^*_∞ in the pure base stock policy presented in Section 2.2. In other words, $S^*=S^*_\infty$, for $T>0$. This means that as T increases starting from zero, S^* decreases and

Table 4
S^*_K and $C(S^*_K, L^*)$ versus T and K, for $L=L^*$, for the single-stage hybrid base stock/kanban policy with ADI

Case 1				Case 2				Case 3				Case 4			
T	K	S^*_K	$C(K,S^*_K)$	T	K	S^*_K	$C(K,S^*_K)$	T	K	S^*_K	$C(K,S^*_K)$	T	K	S^*_K	$C(K,S^*_K)$
4	16	6	82.6148	40	67	34	82.3769	2	6	4	40.1270	10	21	26	41.2114
	17	5	82.4321		69	31	81.7778		7	3	38.0275		23	24	40.7101
	18	5	82.7398		71	31	82.0570		8	3	38.9747		25	24	40.8552
8	16	3	82.2708	60	67	15	81.7949	4	6	2	39.8828	20	21	17	40.7926
	17	2	82.0512		69	12	81.2404		7	1	37.6979		23	16	40.2179
	18	2	82.3473		71	12	81.5260		8	1	38.5364		25	16	40.4074
10	16	1	82.0972	73	67	3	81.4527	6	6	1	39.5846	35	21	2	40.0908
	17	0	81.9289		69	0	80.8782		7	0	37.2553		23	0	39.5676
	18	0	82.1906		71	0	81.1567		8	0	38.1188		25	0	39.6699

reaches zero at T just below L^*, exactly as in the pure base stock policy. To summarize, the basic insights behind the results are the following.

In a production/inventory system operating under the single-stage hybrid base stock/kanban policy, the following is true. When there is no ADI, i.e. when the demand lead time is zero, then: (a) There is a tradeoff between the optimal base stock level and the number of kanbans. (b) This tradeoff holds for up to a finite critical number of kanbans. If the number of kanbans is above this critical number, the optimal base stock level is at its minimum value. This value is equal to the optimal base stock level of the same system operating under the single-stage base stock policy with no ADI. (c) The critical number of kanbans and the corresponding minimum base stock level appear to be the optimal control parameters of the hybrid policy. When there is ADI, i.e. when the demand lead time is zero, then: (a) The optimal number of kanbans appears to be equal to the critical number of kanbans that is optimal in the case where there in no ADI. (b) The optimal base stock level appears to be equal to the optimal base stock level of the same system operating under the single-stage base stock policy with ADI. This means that the linear tradeoff between the optimal base stock level and the demand lead time that appears to hold for the pure base stock policy also holds for the hybrid base stock/kanban policy.

4. Two-stage base stock policy with ADI

In this section, we extend the single-stage base stock policy with ADI considered in Section 2 to a system having two stages. The two-stage base stock policy with ADI is similar to the policy considered in Karaesmen et al. (2002). In the two-stage policy, customer demands arrive for one end-item at a time according to a Poisson process with rate λ, with a constant demand lead time, T, in advance of their due dates, as in the single-stage case. The arrival of every customer demand eventually triggers the consumption of an end-item from FG inventory and the placement of a replenishment production order to the facility of each of the two stages in the system. More specifically, the consumption of an end-item from FG inventory is triggered T time units after the arrival time of the demand, as in the single-stage case. If no end-items are available at that time, the demand is backordered. The control policy depends on four design parameters, namely, the base stock level of end-items in FG inventory at stage n, $n = 1, 2$, denoted by S_n, and the planned supply lead time of stage n, $n = 1, 2$, denoted by L_n. Initially, the system starts with a base stock of S_n end-items in FG inventory at stage n, $n = 1, 2$. The time of placing the replenishment order at stage 2 is determined by offsetting the demand due date by the planned supply lead time of stage 2, L_2. The time of placing the replenishment order at stage 1 is determined by offsetting the demand due date by the sum of the planned supply lead times of stages 1 and 2, $L_1 + L_2$. This means that the delay in placing an order at stage 2 is equal to max(0, $T - L_2$), as in the single-stage case. The delay in placing an order at stage 1, on the other hand, is equal to max[0, $T - (L_1 + L_2)$]. In general, in a system with N stages, the delay in placing an order at stage n is equal to $\max(0, T - L_n^e)$, where L_n^e denotes the *echelon planned supply lead time* at stage n, which is defined as $L_n^e = L_n + L_{n+1} + \cdots + L_N$, $n = 1, 2, \ldots, N$. When an order is placed at stage 1, a new part is immediately released into the facility of stage 1. When an order is placed at stage 2, a new part is also immediately released into the facility of stage 2, provided that such a part is available in the FG output store of stage 1. Otherwise, the order remains on hold until a part becomes available in the FG output store of stage 1. If there is no ADI, i.e. if $T = 0$, both the consumption of an end-item from FG inventory and the replenishment orders are triggered at the demand arrival time, and the resulting policy is the classical base stock policy. A queuing network

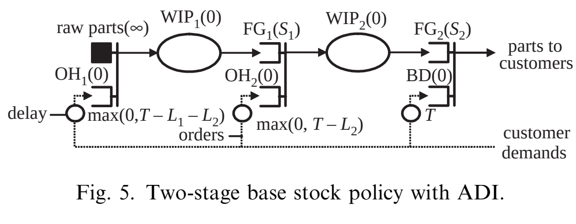

Fig. 5. Two-stage base stock policy with ADI.

model of the two-stage base stock policy with ADI is shown in Fig. 5.

We consider an optimization problem which is similar to that considered in Section 2, where the objective is to find the values of S_1, S_2, L_1, and L_2 that minimize the long-run expected average cost of holding and backordering inventory,

$$\begin{aligned} C(S_1, S_2, L_1, L_2) = {} & h_1 E[\mathrm{WIP}_1 + \mathrm{FG}_1(S_1, L_1, L_2)] \\ & + h_2 E[\mathrm{WIP}_2(S_1, L_1, L_2) \\ & + \mathrm{FG}_2(S_1, S_2, L_1, L_2)] \\ & + bE[\mathrm{BD}(S_1, S_2, L_1, L_2)], \qquad (8) \end{aligned}$$

where h_n is the unit cost of holding WIP+FG inventory per unit time at stage n and b is the unit cost of backordering end-item inventory per unit time. In expression (8), we explicitly express the dependencies of WIP_1, FG_1, WIP_2, FG_2, and BD on parameters S_1, S_2, L_1, and L_2.

4.1. The case where there is no ADI

If there is no ADI, i.e. if $T = 0$, the planned supply lead time parameters L_1 and L_2 are irrelevant. Unfortunately, even in this case there are no analytical results for the optimal base stock levels S_1^* and S_2^*, even when each facility consists of a Jackson network of servers. Some approximation methods for the performance evaluation of the system have been developed in Buzacott and Shanthikumar (1993, Section 10.7), Duri et al. (2000), and Bonvik et al. (1997, Section 8.3.4.3). The only analytically tractable case is the case where $S_1^* = 0$. In this case, the two-stage base stock policy is equivalent to a single-stage base stock policy, where the facilities of stages 1 and 2 are merged into a single facility. The case where $S_1^* = 0$ clearly arises when $h_1 \geqslant h_2$, because then, holding FG inventory at stage 1 not only has a smaller positive impact on customer service than holding FG inventory at stage 2 but is also at least as expensive as holding FG inventory at stage 2. Therefore, if $T = 0$, the only interesting case to look at is the case where $h_1 < h_2$. This case is closer to reality anyway, because, the further downstream inventory is held, the more value has been added to it, and therefore, the more its holding cost. In what follows, we will therefore assume that $h_1 < h_2$.

4.2. The case where there is ADI

If there is ADI, i.e. if $T > 0$, there are no analytical results for the optimal parameter values. As in the single-stage base stock policy with ADI considered in Section 2, intuitively, we would expect that as T increases, the optimal base stock levels of both stages should decrease. The question is how exactly do they decrease in T?

To shed some light into this issue, we numerically investigated a particular instance of the system, in which each facility consists of a Jackson network of two identical, single-server stations in series, and each server has an exponential service rate μ. For this instance, we considered the set of parameter values shown in Table 5. We only looked at one case because there are four parameters to optimize and optimization via simulation is computationally very demanding. The inventory holding cost rates are such that $h_1 < h_2$, so that $S_1^* > 0$ (recall that if $S_1^* = 0$, the two-stage policy is equivalent to a single-stage policy where the facilities of stages 1 and 2 are merged into a single facility).

For this set of parameter values, we used simulation to evaluate the long-run expected average cost of the system, and we optimized the control parameters, S_1, S_2, L_1, and L_2 for different values of T using exhaustive search. The optimization yielded the following results.

For $T = 0$, L_1 and L_2 are irrelevant and $S_1^* = 24$ and $S_2^* = 32$. As T increases from zero, S_1^*appears to remain constant, while S_2^* appears to decrease linearly with T and reach zero just below $T = L_2^*$, where $L_2^* = 34$. As T further increases from L_2^*, S_2^* appears to remain zero, while S_1^* appears to decrease linearly with T and reach zero just below $T = L_1^* + L_2^*$, where $L_1^* +$

Table 5
Parameter values for the case of the two-stage base stock policy with ADI

Case	$1/\lambda$	$1/\mu$	$\rho = \lambda/\mu$	h_1	h_2	b
1	1.1	1.0	0.90909…	1	3	9

Table 6
S_1^*, S_2^*, and $C(S_1^*, S_2^*)$ versus T, for $L_1 = L_1^*$ and $L_2 = L_2^*$, for the two-stage base stock policy with ADI

T	S_1^*	S_2^*	$C(S_1^*, S_2^*)$
0	24	32	158.7183
10	24	23	157.8996
20	24	13	157.0982
25	24	9	156.6986
33	24	1	156.0171
34	24	0	155.9370
40	17	0	155.7578
50	10	0	154.8409
60	1	0	155.2108
61	0	0	154.9056
95	0	0	155.0616

$L_2^* = 61$, therefore $L_1^* = 27$. The optimal values of S_1^* and S_2^* and the resulting cost $C(S_1^*, S_2^*)$ versus T are shown in Table 6 and are plotted in Fig. 6, for the optimal planned supply lead times $L_1^* = 27$ and $L_2^* = 34$. Table 7 shows the optimal values of S_1^* and S_2^* and the resulting cost $C(S_1^*, S_2^*)$ versus T and selected values of L_1 and L_2 around the optimal values.

The insight behind the results is the following. When $T = 0$, $S_1^* > 0$ and $S_2^* > 0$. When $S_1^* > 0$ and $S_2^* > 0$, it appears that it is optimal to place a replenishment production order immediately upon the arrival of a customer demand to every stage, irrespectively of the value of T (as long as T is small enough so that $S_1^* > 0$ and $S_2^* > 0$). Whenever a replenishment production order is placed immediately upon the arrival of a customer demand to every stage, T affects only the FG inventory of stage 2 and BD and not what goes on in the two facilities. In this case, T is a tradeoff for S_2^*, where S_2^* also affects only the FG inventory of stage 2 and BD. Therefore, as T increases from zero, it appears that it is optimal to reduce only S_2^* and not S_1^*. When T is just below L_2^*, S_2^* becomes zero. As T increases beyond L_2^*, it appears that S_2^* remains at zero, and orders are placed at stage 2 with a delay of $T - L_2^*$. At the same time, S_1^* starts decreasing with T, while orders are still placed at stage 1 with no delay. When T is just below $L_1^* + L_2^*$, it appears that S_1^* becomes zero too. As T increases beyond $L_1^* + L_2^*$, both S_1^* and S_2^* remain at zero, while orders are placed at stages 2 and 1 with delays of $T - L_2^*$ and $T - (L_1^* + L_2^*)$, respectively. As in the case of the single-server station, the optimal planned supply lead times are independent of T. The minimum long-run expected average cost decreases *very little* with T and attains its minimum value at $T = L_1^* + L_2^*$.

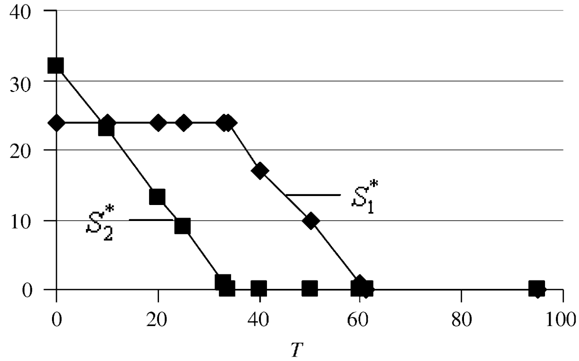

Fig. 6. S_1^*, S_2^*, and $C(S_1^*, S_2^*)$ versus T, for $L_1 = L_1^*$ and $L_2 = L_2^*$, for the two-stage base stock policy with ADI.

Table 7
S_1^*, S_2^*, and $C(S_1^*, S_2^*)$ versus T, L_1 and L_2 for the two-stage base stock policy with ADI

T	L_1	L_2	S_1^*	S_2^*	$C(S_1^*, S_2^*)$
34	27	33	25	0	156.1356
	27	34	24	0	155.9370
	27	35	24	0	155.9370
40	27	33	18	0	156.0771
	27	34	17	0	155.7578
	27	35	15	0	156.5931
61	26	34	6	0	155.9039
	27	34	0	0	154.9056
	28	34	0	0	154.9056

The results imply that as T increases and therefore more demand information becomes available in advance, the optimal base stock levels of all stages appear to drop to zero one after the other, starting from the last stage. An alternative way of looking at this is that as T increases, the optimal *echelon* base stock level of every stage appears to drop to zero, where by echelon base stock of a stage we mean the sum of the base stock levels of the stage and all its downstream stages. Moreover, replenishment production orders are placed with a delay at a stage only when T is large enough so that the optimal echelon base stock level of the stage is zero. To summarize, the basic insights behind the results are the following.

In a production/inventory system controlled by a two-stage base stock policy, at each stage: (a) there appears to be a linear tradeoff between the demand lead time and the optimal echelon base stock level, and (b) the optimal echelon planned supply lead time appears to be the smallest demand lead time for which the optimal echelon base stock level is zero.

5. Two-stage hybrid base stock/kanban policy with ADI

The two-stage hybrid base stock/kanban policy with ADI is an extension of the single-stage hybrid base stock/kanban policy with ADI, presented in Section 3, to two stages. Recall from our discussion in the second paragraph of Section 3 that in the kind of multi-stage serial systems that we study in this paper, it makes sense to set a (WIP + FG)-cap on the (WIP + FG) inventory in all but the last stage and to set a WIP-cap on the WIP of the last stage; therefore, for the two-stage system considered in this section, we will only consider the hybrid base stock/kanban policy where a (WIP + FG)-cap is set on the (WIP + FG) inventory of the first stage and WIP-cap is set on the WIP of the second stage.

With the above discussion in mind, the two-stage hybrid base stock/kanban policy with ADI behaves exactly like the two-stage base stock policy with ADI as far as the placement of replenishment production orders is concerned. The difference is that in the two-stage hybrid base stock/kanban policy, when a replenishment production order is placed to the facility of stage 1, it is not immediately authorized to be released into that facility unless the (WIP + FG) inventory at stage 1 is below a given (WIP + FG)-cap of K_1 parts. If the (WIP + FG) inventory at stage 1 is at or above K_1, the order is put on hold until the (WIP + FG) inventory drops below K_1 (the (WIP + FG) inventory drops as FG parts from stage 1 are consumed by stage 2). Once the order is authorized to be released into the facility, a new part is immediately released into the facility, because it is assumed that there is an infinite number of raw parts. This policy can be implemented by requiring that every part entering the facility be granted a kanban, where the total number of kanbans is equal to the (WIP + FG)-cap level. Once a part leaves the FG output store, the kanban that was granted (and attached) to it is detached and is used to authorize the release of a new part into the facility. A similar mechanism is in place at stage 2, except that it is the WIP rather than the (WIP + FG) inventory that is constrained, i.e. when a replenishment production order is placed to the facility of stage 2, it is not immediately authorized to be released into that facility unless the WIP at stage 2 is below a given WIP-cap of K_2 parts.

Notice that the two-stage hybrid base stock/kanban policy with no ADI is equivalent to a mixture of the extended kanban policy (Dallery and Liberopoulos, 2000) at stage 1 and the

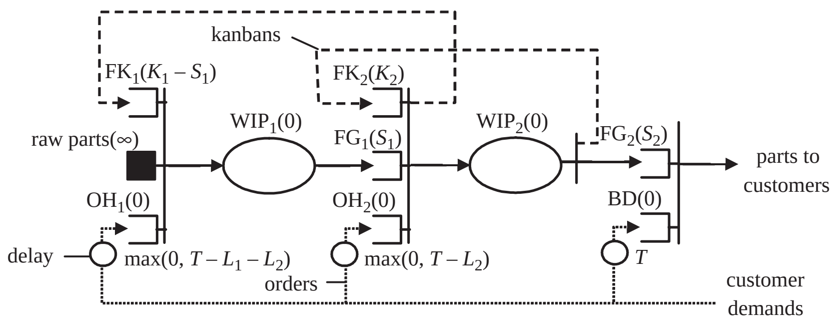

Fig. 7. Two-stage hybrid base stock/kanban policy with ADI.

generalized kanban policy (Buzacott, 1989; Zipkin, 1989) at stage 2. A queuing network model of the two-stage hybrid base stock/kanban policy with ADI is shown in Fig. 7.

If there is ADI, i.e. if $T>0$, there are no analytical results available for the optimal parameter values. To shed some light into this case, we numerically investigated the same instance of the system as that in Section 4, for the same set of parameter values shown in Table 5. For this set of parameter values, we used simulation to evaluate the long-run expected average cost of the system, and we set out to optimize the control parameters K_1, K_2, S_1, S_2, L_1, and L_2 for different values of T, using exhaustive search.

The results of the optimization indicate that the properties of the optimal parameter values are similar to those of the optimal parameter values in the single-stage hybrid base stock/kanban policy. Namely, for $T=0$, L_1 and L_2 are irrelevant, and S_1^* and S_2^* appear to be equal to the optimal base stock levels for the two-stage pure base stock policy, i.e. $S_1^*=24$ and $S_2^*=32$. Moreover, the optimal numbers of kanbans K_1^* and K_2^* appear to be equal to the smallest values of K_1 and K_2 for which the optimal base stock levels are equal to the optimal base stock levels in the two-stage pure base stock policy. These values are $K_1^*=44$ and $K_2^*=28$.

For $T>0$, K_1^* and K_2^* appear to remain constant for all values of T, whereas L_1^*, L_2^*, S_1^*, and S_2^* appear to have the exact same values as in the two-stage base stock policy with ADI discussed in Section 4. Namely, $L_1^*=27$ and $L_2^*=34$, and S_1^* and S_2^* are given in Table 6. The insight behind these results is the same as that behind the results for the single-stage hybrid base stock/kanban policy. The optimal are shown in Table 8 for selected values of K_1 and K_2 around the optimal values and $L_1=L_1^*$, $L_2=L_2^*$.

6. Conclusions

We numerically investigated the tradeoffs between optimal base stock levels, numbers of kanbans, and planned supply lead times in single-stage and two-stage production/inventory systems operating under base stock and hybrid base stock/kanban policies with ADI. The results of our investigation lead to the following conjectures.

In multi-stage make-to-stock production/inventory control policies in which a base stock level of FG inventory is set at every stage, that base stock level represents FG that have been produced

Table 8
S_1^*, S_2^*, and $C(S_1^*,S_2^*)$ versus T, K_1, and K_2, for $L_1=L_1^*$ and $L_2=L_2^*$, for the two-stage hybrid base stock/kanban policy with ADI

T	K_1	K_2	S_1^*	S_2^*	$C(S_1^*,S_2^*)$
0	42	26	31	32	155.7978
	42	28	26	32	155.0267
	42	30	26	32	155.0935
	44	26	27	32	155.1304
	44	28	24	32	154.8046
	44	30	24	32	154.9027
	46	26	26	32	154.9793
	46	28	24	32	156.6758
	46	30	24	32	155.7128
20	42	26	31	13	154.1961
	42	28	26	13	153.5168
	42	30	26	13	153.5238
	44	26	27	13	153.6286
	44	28	24	13	153.3745
	44	30	24	13	153.3911
	46	26	26	13	153.4994
	46	28	24	13	155.2802
	46	30	24	13	154.2590
40	42	26	21	0	153.3327
	42	28	19	0	153.0503
	42	30	19	0	153.2261
	44	26	20	0	153.1847
	44	28	17	0	152.9341
	44	30	17	0	154.1602
	46	26	19	0	153.0693
	46	28	17	0	153.0853
	46	30	17	0	153.1135
95	42	26	0	0	153.1015
	42	28	0	0	151.6153
	42	30	0	0	152.4563
	44	26	0	0	152.9508
	44	28	0	0	151.4917
	44	30	0	0	152.8287
	46	26	0	0	153.0745
	46	28	0	0	152.9572
	46	30	0	0	152.6148

before any demands have arrived to the system in order to satisfy the expected demand during the supply lead time and protect the system against uncertainties in production or demand that may cause costly backorders.

The results in this paper indicate that when there is ADI, the echelon planned supply lead times and the number of kanbans should be as small as possible, so that the placement and authorization for the release of replenishment production orders are delayed as much as possible, as long as this does not cause an increase in the optimal base stock level of FG inventory above its lowest possible value. The lowest possible optimal base stock level is attained when the replenishment policy adopted is such that a replenishment order is placed and released into the facility of every stage immediately after the arrival of the customer demand that triggered it. This can be achieved by setting the echelon planned supply lead time at the first stage greater than or equal to the demand lead time, and by setting the number of kanbans equal to infinity at every stage, so that no inventory-cap is imposed at any stage.

Moreover, for a fixed demand lead time, the more upstream a stage is, the less ADI is available to it and so the higher the need to keep a base stock of FG inventory of that stage. As the demand lead time increases, the amount of ADI increases from downstream to upstream, and so the need to keep a base stock of FG inventory at each stage decreases from downstream to upstream. The results in this paper indicate that it is optimal to reduce the optimal base stock levels at all stages until they drop to zero, one after the other, starting from the last stage and moving upstream the system.

Finally, the optimal number of kanbans determines the optimal production capacity of the system and appears to be independent of the amount of ADI.

References

Bonvik, A.M., Couch, C.E., Gershwin, S.B., 1997. A comparison of production-line control mechanisms. International Journal of Production Research 35 (3), 789–804.

Bourland, K.E., Powell, S.G., Pyke, D.F., 1996. Exploiting timely demand information to reduce inventories. European Journal of Operational Research 92, 239–253.

Buzacott, J.A., 1989. Queueing models of kanban and MRP controlled production systems. Engineering Costs and Production Economics 17, 3–20.

Buzacott, J.A., Shanthikumar, J.G., 1993. Stochastic Models of Manufacturing Systems. Prentice-Hall, Englewood Cliffs, NJ.

Buzacott, J.A., Shanthikumar, J.G., 1994. Safety stock versus safety time in MRP controlled production systems. Management Science 40 (12), 1678–1689.

Chen, F., 2001. Market segmentation, advanced demand information, and supply chain performance. Manufacturing & Service Operations Management 3, 53–67.

Dallery, Y., Liberopoulos, G., 2000. Extended kanban control system: Combining kanban and base stock. IIE Transactions 32 (4), 369–386.

DeCroix, G.A., Mookerjee, V.S., 1997. Purchasing demand information in a stochastic-demand inventory system. European Journal of Operational Research 102 (1), 36–57.

Di Mascolo, M., Frein, Y., Dallery, Y., 1996. An analytical method for performance evaluation of kanban controlled production systems. Operations Research 44 (1), 50–64.

Duri, C., Frein, Y., Di Mascolo, M., 2000. Comparison among three pull control policies: Kanban, base stock and generalized kanban. Annals of Operations Research 93, 41–69.

Frein, Y., Di Mascolo, M., Dallery, Y., 1995. On the design of generalized kanban control systems. International Journal of Operations and Production Management 15 (9), 158–184.

Gallego, G., Özer, A.Ö., 2001. Integrating replenishment decisions with advance demand information. Management Science 47 (10), 1344–1360.

Gilbert, S.M., Ballou, R.H., 1999. Supply chain benefits from advanced customer commitments. Journal of Operations Management 18 (1), 61–73.

Güllü, R., 1996. On the value of information in dynamic production/inventory problems under forecast evolution. Naval Research Logistics 43 (2), 289–303.

Hariharan, R., Zipkin, P., 1995. Customer-order information, leadtimes, and inventories. Management Science 41 (10), 1599–1607.

Karaesmen, F., Dallery, Y., 2000. A performance comparison of pull control mechanisms for multi-stage manufacturing systems. International Journal of Production Economics 68, 59–71.

Karaesmen, F., Buzacott, J.A., Dallery, Y., 2002. Integrating advance order information in make-to-stock production. IIE Transactions 34 (8), 649–662.

Karaesmen, F., Liberopoulos, G., Dallery, Y., 2003. Production/inventory control with advance demand information. In: Shanthikumar, J.G., Yao, D.D., Zijm, W.H.M. (Eds.), Stochastic Modeling and Optimization of Manufacturing Systems and Supply Chains. International Series in Operations Research and Management Science, Vol. 63. Kluwer Academic Publishers, Boston, MA, pp. 243–270.

Karaesmen, F., Liberopoulos, G., Dallery, Y., 2004. The value of advance demand information in make-to-stock manufacturing systems. Annals of Operations Research 126, 135–157.

Liberopoulos, G., Dallery, Y., 2000. A unified framework for pull control mechanisms in multi-stage manufacturing systems. Annals of Operations Research 93, 325–355.

Liberopoulos, G., Dallery, Y., 2002. Base stock versus WIP cap in single-stage make-to-stock production–inventory systems. IIE Transactions 34 (7), 627–636.

Liberopoulos, G., Dallery, Y., 2003. Comparative modeling of multi-stage production–inventory control policies with lot sizing. International Journal of Production Research 41 (6), 1273–1298.

Liberopoulos, G., Tsikis, I., 2003. Unified modeling framework of multi-stage production–inventory control policies with lot-sizing and advance demand Information. In: Shanthikumar, J.G., Yao, D.D., Zijm, W.H.M. (Eds.), Stochastic Modeling and Optimization of Manufacturing Systems and Supply Chains. International Series in Operations Research and Management Science, Vol. 63. Kluwer Academic Publishers, Boston, MA, pp. 271–297.

Milgrom, P., Roberts, J., 1988. Communication and inventory as substitutes in organizing production. Scandinavian Journal of Economics 90, 275–289.

Rubio, R., Wein, L.W., 1996. Setting base stock levels using product-form queueing networks. Management Science 42 (2), 259–268.

Spearman, M.L., 1992. Customer service in pull production systems. Operations Research 40 (5), 948–958.

Spearman, M.L., Woodruff, D.L., Hopp, W.J., 1990. CONWIP: A pull alternative to kanban. International Journal of Production Research 28, 879–894.

Van Donselaar, K., Kopczak, L.R., Wouters, M., 2001. The use of advance demand information in a project-based supply chain. European Journal of Operational Research 130, 519–538.

Veach, M.H., Wein, L.M., 1994. Optimal control of a two-station tandem production/inventory system. Operations Research 42 (2), 337–350.

Wijngaard, J., 2004. The effect of foreknowledge of demand in case of a restricted capacity: The single-stage, single-product case. European Journal of Operational Research 159 (1), 95–109.

Zipkin, P., 1989. A kanban-like production control system: Analysis of simple models. Research Working Paper No. 89-1, Graduate School of Business, Columbia University, New York.

Zipkin, P., 2000. Foundations of Inventory Management. Management & Organization Series. McGraw-Hill, Boston, MA.