IIE Transactions (2004) **36**, 529–546

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/07408170490438519

Make-to-order, make-to-stock, or delay product differentiation? A common framework for modeling and analysis

DIWAKAR GUPTA* and SAIF BENJAAFAR

Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN 55455, USA
E-mail: guptad@me.umn.edu or saif@me.umn.edu

Received June 2002 and accepted October 2003

Delaying product differentiation is a hybrid strategy that strives to reconcile the dual needs of high variety and quick response time. A common product platform is built to stock in the first stage of production (called the Make-To-Stock (MTS) stage) which is then differentiated into different products after demand is known in the second stage (known as the Make-To-Order (MTO) stage). In this article, we develop models to compute the costs and benefits of delaying differentiation in series production systems when the order lead times are load dependent. The models are then used to gain insights through analytical and numerical comparisons. For example, we observe the following patterns in a large number of numerical experiments. The effect of congestion in the MTS and MTO stages is asymmetric with tighter capacity at the MTO stage having a greater detrimental effect on the desirability of delaying differentiation. If there is flexibility in choosing the point of differentiation, higher loading is observed to favor later differentiation. Also, if the sequence in which work is performed can be affected, then placing workstations that have a tighter capacity in the MTS stage lowers costs.

1. Introduction

In today's business environment, a manufacturing firm that has the ability to fill customer orders quickly, as well as offer custom products, enjoys a competitive advantage. However, the need to a have high product variety and quick response time places conflicting demands on the production system (Lee, 1996; Lee and Tang, 1997; Fisher *et al.*, 1999). It is for this reason that businesses that compete on response time focus on producing a limited portfolio of products. Items are produced ahead of demand and kept in stock, ready to be shipped upon receipt of orders. Producing to stock becomes costly when the number of products is large. It is also risky when demand is highly variable and/or products have short life cycles. Therefore, a significant increase in product variety normally goes hand-in-hand with a shift from a Make-To-Stock (MTS) to a Make-To-Order (MTO) mode of production. In the MTO mode, production is not initiated until a customer order is received. While this strategy eliminates finished-goods inventories and reduces a firm's exposure to financial risk, it usually spells long customer lead times and large order backlogs. An alternative to both MTO and MTS paradigms that has recently gained in popularity is *delayed differentiation* (Lee and Billington, 1994; Feitzinger and Lee, 1997; Swaminathan and Tayur, 1998). Delayed differentiation is a hybrid strategy in which a common product platform is built to stock. It is differentiated, by assigning to it certain customer-specific features, only after demand is realized. Hence, manufacturing occurs in two stages, (i) a MTS stage where one or more undifferentiated platforms are produced and stocked; and (ii) a MTO stage where product differentiation takes place in response to specific customer orders.

Delaying differentiation carries several benefits. Maintaining stocks of semi-finished goods reduces the order-fulfillment delay relative to the pure MTO system. Since many different end products have common parts, holding semi-finished goods inventory benefits from demand pooling, which is known to lower the amount of inventory needed to achieve a service-level performance equal to that of a comparable system with no pooling (Eppen, 1979). Furthermore, investment in semi-finished inventories is smaller when compared with the option to maintain a similar amount of finished-goods inventory. There is also the benefit of learning, realized from having better demand information before committing generic semi-finished products to unique end products. Additional benefits from delaying differentiation include a significant streamlining of the MTS segment of the manufacturing process and

*Corresponding author

simplification of production scheduling, sequencing and raw material purchasing. However, implementing delayed differentiation also carries costs; included here are the costs of extra materials (when common designs are made possible by having redundant or more expensive parts) and less efficient processing (when common processing leads to the use of a less specialized production equipment, extra processing steps, or greater yield losses). In this paper, we call all costs necessary to realize a common product platform as the product and process redesign costs. Therefore, in assessing the value of delaying differentiation, its costs have to be carefully weighed against its counter-balancing benefits.

Recent literature showcases several examples in which delayed differentiation has been successfully used to control inventory costs while maintaining high service levels. For example, Feitzinger and Lee (1997) describe how delayed differentiation, through operation reversal and component sharing, helped Hewlett-Packard's printer division to customize its products in a cost-effective manner. Swaminathan and Tayur (1998) describe how IBM exploited component commonalities in personal computers to design a common platform, or a vanilla box, from which end products are differentiated based on customer orders. Fisher *et al.* (1999) discuss how component sharing is used by several major automotive companies to standardize braking systems. Bruce (1987) reports on the well-known case of Benetton who, by reversing the order in which yarn is dyed and knitted, are able to successfully delay color selection until the season's fashion preferences become more established. Graman and Magazine (1998) describe how delayed packaging is used by a manufacturer of household cleaning products to reduce its finished-goods inventory and improve service levels.

Previous studies also present quantitative models to assess the costs and benefits of delaying differentiation (see, for example, Lee (1996), Garg and Tang (1997), Lee and Tang (1997), Swaminathan and Tayur (1998, 1999), and Aviv and Federgruen (1999, 2001a, 2001b), among others). These models focus on the trade-off between the benefits of inventory pooling and learning on the one hand, and product/process redesign costs, on the other. The majority use order-up-to-level inventory models in which order lead times are not affected by the order size, the number of pending orders, or the number of end products. If limited production capacity is modeled, the order lead times are assumed to be constant which ignores any congestion at the production facility. In doing so, these models ignore how the utilization of the production facility and processing time variability interact to affect order delays. Recently, Aviv and Federgruen (2001a, 2001b) have modeled a system that is closely related to the model under investigation in this article. They consider a two-stage system in which stage-1 produces undifferentiated items that are later differentiated in stage-2. The lead times in both stages are assumed constant. In Aviv and Federgruen (2001a), both stages have no capacity constraint, whereas in Aviv and Federgruen (2001b), stage-1 has limited capacity.

In this paper, we extend the existing literature on delayed differentiation by explicitly modeling the effect of queueing at both the MTS and the MTO stages of the production process. To our knowledge, our paper is the first to examine the benefits of delaying differentiation when lead times are load dependent and induced by a capacitated production system. Our work is related to a growing body of literature on MTS queues, an overview of which can be found in Buzacott and Shanthikumar (1993). Other treatments of multi-class production-inventory systems include Lee and Zipkin (1992, 1995), Wein (1992), Zipkin (1995), Ha (1997), and De Véricourt *et al*. (2000).

The models presented in this article are deliberately kept simple so that operations managers may use them to examine the key trade-offs arising from different manufacturing-system configurations. In fact, for some commonly occurring configurations, we carry out numerical and analytical comparisons and identify new insights. That is, there are three types of results presented in this article. First, we present algebraic expressions to compute the aggregate performance measures of systems that use the Delayed Differentiation (DD) strategy and systems that keep finished-goods inventories. In all cases, the performance measures of systems with DD are based on an approximation. These results are presented as propositions. Next, we use these performance measures to compute the optimal stocking levels for different system configurations in numerical experiments. Generalizations based on observing patterns that emerge from these numerical experiments are reported as observations. Such observations appear to be generally true, but we have not been able to prove them through rigorous mathematical arguments. Finally, we compare the limiting behavior of some configurations using the algebraic expressions of performance measures. These comparisons are also reported as propositions, and they are in agreement with the generalizations based on numerical experiments.

Some examples of the generalizations that are supported by a large number of numerical experiments are as follows. A higher utilization at either stage (which implies greater congestion) makes DD less desirable. Intuitively, this can be explained as follows. A tighter second-stage capacity forces the system with DD to hold more inventory in order to meet service-level requirements or avoid shortage penalties (depending on the model formulation). The inventory investment is more significant under DD because the second stage does not carry inventory, with the result that when utilization is sufficiently high, DD becomes inferior to pure MTS. The effect of a tighter capacity at the first stage is more subtle. A higher stage-1 utilization induces a higher congestion and causes the lead times experienced by consecutive orders processed by stage-1 to have a greater positive dependence. An increased positive dependence in turn diminishes the value of the inventory pooling associated with DD. We observe that having a tighter capacity in the second

stage has a greater negative effect on the desirability of DD. A corollary of this phenomenon is that if there exists some flexibility in choosing the point of differentiation, a higher loading tends to favor later differentiation. Also, when there is flexibility in ordering the workstations that constitute the production line, placing workstations that have a tighter capacity in the MTS stage is found to be more cost effective.

The remainder of this article is organized as follows. In Section 2, we discuss our target application and modeling assumptions. In Section 3, we evaluate the advantage of DD relative to a pure MTS system when commonalities among products already exist. In Section 4, we identify the optimal point of differentiation when varying degrees of DD can be achieved at the cost of redesigning the products and/or the production processes. The effect of using several partially differentiated products, instead of a single undifferentiated platform common to all products, is explored in Section 5. Finally, in Section 6, we summarize key results. The proofs of selected propositions are provided in the Appendix.

2. Target application and modeling assumptions

We focus attention on assemblers that initially choose a strategy for competing on price, quality, size of product menu and delivery-time dimensions. These choices and the competitive response, in turn determine their market share and demand rate. Assemblers then design their production systems to achieve the desired performance on the chosen competitive dimensions. This paper is concerned with the selection of an appropriate manufacturing-system configuration with the goal of minimizing the overall system costs. The choices considered are the pure MTO, the pure MTS and a hybrid design which utilizes DD.

The primary application of our models is to industries, such as computer assembly, that make different products utilizing a component assembly process with product-invariant assembly time. For example, all laptop models in a particular product line undergo the same assembly process with nearly uniform assembly times. In these systems, DD is enabled either by taking advantage of existing commonalities among products, or by standardizing components across products or by reorganizing operations so that those that are common to all products are performed first (Lee and Tang, 1997).

A multistage production system has three types of inventory: (i) a Work-In-Process (WIP) inventory; (ii) a semi-finished-goods inventory; and (iii) a finished-goods inventory. The WIP inventory consists of raw materials or semi-finished products that have been released to a workstation, but have not yet completed processing. These items are worked upon as soon as the workstation becomes available. In contrast, the semi-finished goods inventory resides between production stages. It is inventory that is staged ahead of demand and has not yet been released to the downstream stage. Release of this inventory is triggered only by the arrival of a customer order, at which time a unit of semi-finished-goods inventory (if one is available) is transfered to the immediate downstream station to become WIP for that station. The location of the semi-finished-goods inventory defines the boundary between MTS and MTO stages, also known as the *push-pull* boundary. The removal of semi-finished and finished goods from their respective stores is driven by demand from the immediate downstream stage, or exogenous customers, as appropriate. In contrast, WIP is found in the input buffers of workstations, regardless of whether the workstation belongs to the MTS or MTO stage.

For each fixed service level, the standardization of components and semi-finished products and inventory pooling that goes hand-in-hand with DD can lower requirements of all three types of inventory. In this article, we focus on the reduction in cost attributed to semi-finished-goods inventory because such benefits accrue only to the assembler. In contrast, the benefits of component standardization are mostly realized by the supplier since the applications we have in mind are those in which components are delivered just-in-time and little or no inventory is held at the assembly site (Magretta, 1998).

The reduction in WIP inventory is due to the fact that the input process to the MTO stage is affected by the size of the undifferentiated inventory buffer separating the MTS and the MTO stages (see Section 3). Having an inventory buffer between the two stages can lower the inter-arrival time variability to the second stage, thereby reducing WIP in that stage. However, the overall impact of pooling on WIP inventory is relatively small since both the total demand and the assembly times are assumed not to be affected by DD. In Section 3, we present further evidence that the net change in WIP inventory is negligible in all practical applications and consequently argue that it may be ignored.

Manufacturing managers often set and strive to achieve explicit delivery time goals, which they may measure either as the average order-fulfillment time, or as the proportion of orders that exceed a critical delivery-time target (e.g., a quoted lead time). The models we present treat either of these points of view. Our choice of order delay as the measure of service stems from the observation that most applications of DD arise in situations where quick response to customer orders is central to the competitiveness of the firm. We also treat the case where instead of striving to meet a service-level constraint, the assembler incurs a backorder cost per unit short per unit time. We assume that there is no setup involved in switching from assembling one product to another. This is true in many industries where our models will be applicable, e.g., computer assembly. We assume in all our models that customer orders are met on a First-Come First-Served (FCFS) basis. Since all customers carry the same priority and are quoted the same lead time, the use of the FCFS policy is reasonable.

We consider three situations in this article. In the first situation, there are some pre-existing commonalities among products, e.g., when the first few operations/components

are naturally common among all products. The question then is "should we take advantage of the existing commonalities and delay differentiation until demand is realized or should we continue to produce each product in the MTS fashion?" In the second situation, we consider systems that have flexibility in choosing the point of differentiation. The question there is "what is the optimal point of differentiation and how is it affected by system parameters?" In the third situation, we consider systems where instead of building a single and expensive common platform across all products, it is cheaper to build several semi-differentiated platforms; each of which may correspond to a different product family. The question then is "what is the relative advantage of full DD (a single platform) as compared to partial differentiation (multiple platforms)? Put differently, "are there cases for which partial DD is nearly as good as full DD?"

3. MTS versus DD

In this section, we identify conditions under which a firm should take advantage of naturally occurring commonalities and delay differentiation instead of producing in the MTS fashion. A graphical depiction of the two systems being compared is shown in Fig. 1.

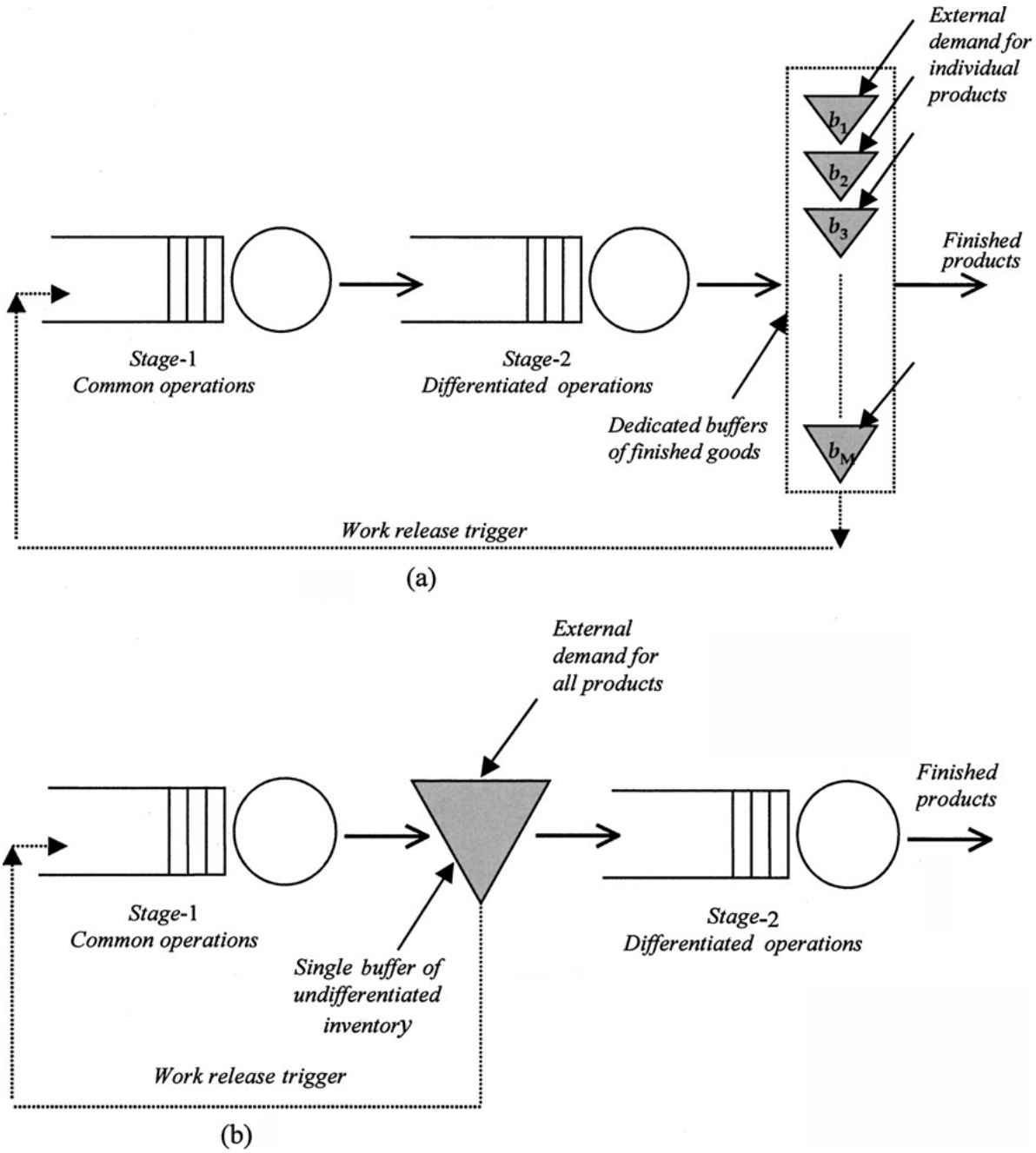

Fig. 1. The two systems under consideration: (a) pure MTS; and (b) system with DD.

3.1. *The model*

There are M finished products, indexed by j; $j = 1, 2, \cdots, M$. Demand for product-j occurs according to a Poisson process of rate λ_j; Λ denotes the total demand rate. For the pure MTS system, demand is satisfied from stock unless the corresponding inventory store is empty. All shortages are backordered. For the system with DD, each demand arrival releases an undifferentiated item from the intermediate inventory store, which then joins the queue of jobs, if any, that are waiting to be processed at stage-2 where differentiation takes place. However, if there are no semi-finished items in store, then the demand is backlogged for processing at stage-1. Our DD model can also be viewed as an instance of a generalized queueing network with signals where a demand arrival signals a transfer from the buffer to stage-2 (see Chao and Zheng (1998) for details).

For both systems, inventories are managed according to a base-stock policy, that is, each demand triggers the immediate release of a new raw-material kit (which is assumed to be always available) to the queue of kits at stage-1. The base-stock level for semi-finished items is denoted as b_{d}, whereas the base-stock level for type-j finished items is called $b_{\mathrm{f}}(j)$. The base-stock levels are also referred to as the buffer sizes since replenishments under a base-stock control policy are triggered to keep the inventory buffers full. The processing rate at stage-i is denoted by μ_i (regardless of product type). For stability, we require that $\Lambda/\mu_i = \rho_i < 1$, where ρ_i is the stage-i utilization. We assume that the unit processing times at each stage are exponentially distributed. This helps simplify analysis and represents the practical worst case for benchmarking production system performance (Hopp and Spearman, 2000). Finally, we let $\lambda_j = \lambda = \Lambda/M$ for all j and assume that per-unit holding and backorder costs are independent of product type. These assumptions of symmetry imply that optimal buffer sizes for finished products are also equal. That is, $b_{\mathrm{f}}(j) = b_{\mathrm{f}}$ for all j. The above-mentioned simplifications are necessary in order to carry out fair comparisons between systems with different levels of product variety. It is, however, mathematically straightforward to extend the analysis to systems with unequal demand and product-specific holding and backorder costs.

Before the DD and the MTS configurations can be compared, managers need to resolve the related operational problem of choosing the best values of b_{d} and b_{f}, which we shall denote with the superscript "$*$." Our modeling approach allows several different formulations of the operations managers' objective function for the purpose of finding b_{d}^* and b_{f}^*. For example, a plausible objective is to minimize the relevant inventory costs subject to a service-level constraint. Alternatively, managers may want to minimize the sum of inventory and backorder costs. An acceptable service can be specified either as an upper bound on the average order-fulfillment time or as an upper bound on the probability of not filling customer orders within a quoted lead time. (Note that the order-fulfillment time is the total time elapsed from the moment a demand arrives to the moment the finished product is supplied to the customer.) Fill rate or stock-out probability are not meaningful measures of service in the setting we describe since no order can be filled immediately upon receipt under DD. Indeed, we are implicitly assuming that the customers are willing to tolerate some delay, although they measure performance by a function of the length of this delay. Since we impose the same service-level requirement on both systems, the two systems can be compared in terms of the cost of inventories held either as finished goods or as undifferentiated items (for cost-based systems, we must also consider backorder costs). We shall observe later in this section that WIP inventory level is approximately the same for the two systems. Carrying costs of WIP inventory are therefore irrelevant for the task of finding the optimal buffer sizes.

We shall focus next on developing a detailed formulation of the problem for finding b_{d}^* and b_{f}^*. To this end, we shall derive expressions for the relevant costs and service performance measures in each case. The inventory and backorder costs for the pure MTS system are given respectively by: $Mh_{\mathrm{f}}\bar{I}_{\mathrm{f}}(b_{\mathrm{f}})$ and $M\beta\bar{B}_{\mathrm{f}}(b_{\mathrm{f}})$, where h_{f} is the holding-cost rate of finished goods, β is the backorder cost per item short per unit time and $\bar{I}_{\mathrm{f}}(y)$ and $\bar{B}_{\mathrm{f}}(y)$ are respectively the average finished-goods inventory and backorder level for each end product when the base-stock level is y. Similarly, inventory cost and backorder costs for the system with DD are given respectively by $h_{\mathrm{d}}\bar{I}_{\mathrm{d}}(b_{\mathrm{d}})$ and $\beta\bar{B}_{\mathrm{d}}(b_{\mathrm{d}})$ where h_{d} is the holding-cost rate for undifferentiated inventory and $\bar{I}_{\mathrm{d}}(y)$ is the average semi-finished-goods inventory given a base-stock level of y. The term $\bar{B}_{\mathrm{d}}(y)$ stands for the average total backorder level for all finished items. We use the notation $\bar{F}_{\mathrm{f}}(y)$ and $\bar{F}_{\mathrm{d}}(y)$ to refer, respectively, to the expected order-fulfillment time for the pure MTS and the DD systems. Similarly, we use the notation $F_{\mathrm{x}}(y)$ to refer to the probability that the order-fulfillment time exceeds a quoted lead time x, when the base-stock level is y.

Upon treating each stage as a single-server queueing system, performance measures of interest for the MTS system can be derived as shown in the following proposition.

Proposition 1. *In the MTS system, the expected inventory, backorder level and the order-fulfillment time are*:

$$\bar{I}_{\mathrm{f}}(b_{\mathrm{f}}) = \begin{cases} M\left(b_{\mathrm{f}}(1+\hat{\rho}^{b_{\mathrm{f}}+1}) - \dfrac{2\hat{\rho}(1-\hat{\rho}^{b_{\mathrm{f}}})}{1-\hat{\rho}}\right) & \text{if } \rho_1 = \rho_2 = \rho, \\ \dfrac{M^2(1-\rho_1)(1-\rho_2)}{(\rho_2-\rho_1)}\left(\dfrac{b_{\mathrm{f}}\hat{\rho}_2(1-\hat{\rho}_2) - \hat{\rho}_2^2\left(1-\hat{\rho}_2^{b_{\mathrm{f}}}\right)}{(1-\hat{\rho}_2)^2}\right. & \\ \quad \left. - \dfrac{b_{\mathrm{f}}\hat{\rho}_1(1-\hat{\rho}_1) - \hat{\rho}_1^2\left(1-\hat{\rho}_1^{b_{\mathrm{f}}}\right)}{(1-\hat{\rho}_1)^2}\right) & \text{otherwise,} \end{cases} \tag{1}$$

$$\bar{F}_{\mathrm{f}}(b_{\mathrm{f}}) = \begin{cases} \dfrac{M}{\Lambda}\left(b_{\mathrm{f}}\hat{\rho}^{b_{\mathrm{f}}+1} + \dfrac{2\hat{\rho}^{b_{\mathrm{f}}+1}}{1-\hat{\rho}}\right) & \text{if } \rho_1 = \rho_2 = \rho, \quad \text{and} \\ \dfrac{M^2(1-\rho_1)(1-\rho_2)}{\Lambda(\rho_2-\rho_1)}\left(\dfrac{\hat{\rho}_2^{b_{\mathrm{f}}+2}}{(1-\hat{\rho}_2)^2} - \dfrac{\hat{\rho}_1^{b_{\mathrm{f}}+2}}{(1-\hat{\rho}_1)^2}\right) & \text{otherwise,} \end{cases} \tag{2}$$

$$\bar{B}_{\mathrm{f}}(b_{\mathrm{f}}) = \bar{I}_{\mathrm{f}}(b_{\mathrm{f}}) + M\left(\frac{\hat{\rho}_1}{1-\hat{\rho}_1} + \frac{\hat{\rho}_2}{1-\hat{\rho}_2} - b_{\mathrm{f}}\right), \tag{3}$$

where $\hat{\rho} = \rho/[M(1-\rho)+\rho]$ *and* $\hat{\rho}_i = \rho_i/[M(1-\rho_i)+\rho_i]$. *Furthermore, the probability that the order-fulfillment time exceeds a quoted lead time* x ($x \geq 0$) *is given by*:

$$F_x(b_{\mathrm{f}}) = \Pr(\bar{F}_{\mathrm{f}}(b_{\mathrm{f}}) \geq x) = \hat{\rho}_2^{b_{\mathrm{f}}} e^{-[\lambda(1-\hat{\rho}_2)/\hat{\rho}_2]x} + \hat{\rho}_1 a(x, b_{\mathrm{f}}), \tag{4}$$

where, for any $y \geq 0$:

$$a(x,y) = \begin{cases} (1-\hat{\rho}_2)\left(\dfrac{\hat{\rho}_2^y e^{-[\lambda(1-\hat{\rho}_2)/\hat{\rho}_2]x} - \hat{\rho}_1^y e^{-[\lambda(1-\hat{\rho}_1)/\hat{\rho}_1]x}}{\hat{\rho}_2 - \hat{\rho}_1}\right), & \text{if } \hat{\rho}_1 \neq \hat{\rho}_2, \\ e^{-[\lambda(1-\hat{\rho})/\hat{\rho}]x}\hat{\rho}^{y-2}(1-\hat{\rho})(\hat{\rho}y + \lambda x), & \text{if } \hat{\rho}_1 = \hat{\rho}_2 = \hat{\rho}. \end{cases} \tag{5}$$

Proof. A proof is provided in Appendix A. ■

Evaluating performance measures for the model with DD is more complicated since the stage-2 input process is Poisson only in two special cases: $b_{\mathrm{d}} = 0$ and $b_{\mathrm{d}} = \infty$. The former instance results in two $M/M/1$ queues in tandem whose steady-state probabilities are known to have a product-form structure (Jackson, 1957). Similarly, when buffer size is very large, the two stages are completely decoupled and behave like two independent $M/M/1$ queues. However, when $0 < b_{\mathrm{d}} < \infty$, there is a positive dependence between the arrival of input units from stage-1 to stage-2 (see Appendix A for details), which makes the determination of the steady-state probabilities difficult.

Lee and Zipkin (1992) develop an approximation scheme for evaluating performance measures of multistage production systems by proposing that each stage be treated as an exogenous, sequential supply system. The term exogenous is used to underscore their assumption that the workload at each stage is independent of the other stages in the system. Each supply system is then modeled as having an exponentially distributed delay with parameter $v_i = \mu_i(1-\rho_i)$. In our model, this is tantamount to assuming that each stage operates like a $M/M/1$ queue. Lee and Zipkin's approximation procedure can be adapted for systems for which the processing delay at each stage can be modeled as a continuous phase-type distribution after ignoring the interdependence among different stages.

Buzacott *et al.* (1992) characterize the distribution of inter-arrival times to the second stage in a two-stage system of the type shown in Fig. 1(b). They observe that the coefficient of variation of the inter-arrival times at stage-2 is between 0.8 and one. Therefore, they recommend using the $M/M/1$ approximation for stage-2 queueing systems. In effect, both approximations replace a system with $b_{\mathrm{d}} > 0$ by a system with $b_{\mathrm{d}} = 0$ for the purpose of finding joint queue occupancy levels. Both articles report that the approximation works well after comparing approximate performance measures to their estimates obtained via simulation. Lee and Zipkin also point out that their approximate procedure is sufficiently accurate to be used in applications in which one wishes to find optimal base-stock levels. In this article, we use Lee and Zipkin's approximation to estimate performance measures of all systems that use DD. Given our goal of providing a rapid modeling tool useful for generating insights, this represents an appropriate compromise between accuracy and ease of modeling.

Proposition 2 below provides expressions for performance measures of interest. The proof is omitted as it is similar to the way in which Equations (1)–(4) are derived.

Proposition 2. *In a system with DD, expected inventory, the backorder level and the order-fulfillment time are given by*:

$$\bar{I}_{\mathrm{d}}(b_{\mathrm{d}}) = b_{\mathrm{d}} - \left(\frac{\rho_1(1-\rho_1^{b_{\mathrm{d}}})}{1-\rho_1}\right). \tag{6}$$

$$\bar{B}(b_{\mathrm{d}}) \approx \frac{\rho_2}{1-\rho_2} + \frac{\rho_1^{b_{\mathrm{d}}+1}}{1-\rho_1}, \tag{7}$$

$$\bar{F}_{\mathrm{d}}(b_{\mathrm{d}}) \approx \frac{\rho_1^{b_{\mathrm{d}}+1}}{\Lambda(1-\rho_1)} + \frac{\rho_2}{\Lambda(1-\rho_2)}. \tag{8}$$

Furthermore, the probability that the order-fulfillment time exceeds a quoted lead time x *is*

$$F_x(b_{\mathrm{d}}) = \Pr(\bar{F}_{\mathrm{d}}(b_{\mathrm{d}}) \geq x),$$
$$\approx \begin{cases} \left(1+\rho^{b_{\mathrm{d}}}(1-\rho)\mu x\right)e^{-\mu(1-\rho)x} & \text{if } \rho_1 = \rho_2 = \rho, \\ e^{-\mu_2(1-\rho_2)x} + \left(\dfrac{(1-\rho_2)\rho_1^{b_{\mathrm{d}}+1}}{\rho_2-\rho_1}\right) & \\ \quad \times\left(e^{-\mu_2(1-\rho_2)x} - e^{-\mu_1(1-\rho_1)x}\right) & \text{otherwise.} \end{cases} \tag{9}$$

Consider now the WIP inventory in the two types of production systems. In the pure MTS system, each production stage is a $M/M/1$ queue with a utilization factor of ρ_i and therefore the total WIP is $\sum_i \rho_i/(1-\rho_i)$ (see, for example, Kleinrock (1975)). Since the MTS stage of the system with DD is indeed a $M/M/1$ queue, its average WIP is $\rho_1/(1-\rho_1)$. The MTO stage is not $M/M/1$, however, it can be approximated quite accurately by a $M/M/1$ queue (Lee and Zipkin, 1992). Thus, a close approximation for the average WIP in the DD system is the same as the average

WIP in the pure MTS system. This is the reason why the relevant inventory cost terms do not include the WIP inventory costs.

Let $z_f(b_f)$ and $z_d(b_d)$ denote expected costs for the MTS and DD systems respectively. Then, for systems with service-level constraints, $z_f(b_f) = h_f M \bar{I}_f(b_f)$ and $z_d(b_d) = h_d \bar{I}_d(b_d)$. For cost-based systems, the objective is to minimize the sum of expected inventory holding and backordering costs, i.e., $z_f(b_f) = M(h_f \bar{I}_f(b_f) + \beta \bar{B}(b_f))$ and $z_d(b_d) = h_d \bar{I}_d(b_d) + \beta \bar{B}(b_d)$. From the fact that for both pure MTS and DD systems, the average inventory is increasing in the base-stock level, b_f or b_d, and that the average order delay and the probability of the delay not exceeding a quoted lead time are both decreasing in the same, it follows that the optimal values of b_f and b_d are always the smallest values that satisfy the service-level constraints. For cost-based systems, the functions $z_f(b_f)$ and $z_d(b_d)$ are convex in b_f and b_d respectively. Therefore, the optimal base-stock level is the smallest integer b_f that satisfies $z_f(b_f + 1) - z_f(b_f) \geq 0$ for the MTS system and $z_d(b_d + 1) - z_d(b_d) \geq 0$ for the DD system.

Obtaining a closed-form expression for the optimal base-stock level for the MTS system is difficult. However, it can be easily computed numerically. For the approximating system with DD, whose performance measures are listed in Proposition 2, a closed-form expression for the optimal base-stock level can be derived for each problem formulation. These are presented in the proposition below (the proof is omitted for brevity).

Proposition 3. *For a DD system, if the objective is to minimize the expected inventory cost subject to the expected order-fulfillment time not exceeding* α, *then*:

$$b_d^*(\alpha) = \left\lceil \frac{\ln[\Lambda(1-\rho_1)(\alpha - \rho_2/\Lambda(1-\rho_2))]}{\ln(\rho_1)} - 1 \right\rceil, \quad (10)$$

where $\lceil t \rceil$ *represents the integer ceiling of* t. *If the service performance is defined by an upper bound on the probability of the order-replenishment time exceeding a quoted lead time* x, *then*:

$$b_d^*(x) = \begin{cases} \left\lceil \dfrac{\ln((\alpha - \rho^x)/(\rho^x(1-\rho)))}{\ln(\rho)} \right\rceil & \text{if } \rho_1 = \rho_2 = \rho, \\ \left\lceil \dfrac{\ln\left((\alpha - \rho_2^x)(\rho_2 - \rho_1)/(\rho_1(1-\rho_2)(\rho_2^x - \rho_1^x))\right)}{\ln(\rho_1)} \right\rceil & \text{otherwise,} \end{cases} \quad (11)$$

and, if the objective is to minimize the sum of the expected holding and backorder costs, then:

$$b_d^* = \left\lfloor \frac{\ln[h_d/(h_d + \beta)]}{\ln(\rho_1)} \right\rfloor. \quad (12)$$

Notice that the optimal base-stock for a cost-based system is independent of ρ_2 since with DD all items are backordered.

3.2. *Analysis and comparisons*

Next, we compare the optimal costs incurred in the MTS and the DD system in a large number of numerical experiments. Patterns that emerge from these experiments are reported below as observations. We also comment on the importance of these observations to operations managers. Later, in Section 5, we shall see that the limiting behavior of the optimal cost functions observed in the numerical experiments indeed matches with the limiting behavior of the algebraic functions representing the costs in closely related systems. However, we have not been able to prove the results reported in the remainder of this section through rigorous mathematical arguments.

Observations are reported under three headings, where we focus, respectively, on: (i) the effect of loading; (ii) the number of products; and (iii) the desired service level. The system design objective is to minimize the expected inventory holding cost while keeping the average order delay below a maximum permissible level. However, the qualitative insights remain largely valid when a constraint on the probability of the order delay exceeding a quoted lead time is used, or when a cost-only formulation is used. Differences are noted when appropriate.

Since $\bar{F}_d(b_d) \to \rho_2/\Lambda(1-\rho_2)$ when $b_d \to \infty$ and $\bar{F}_d(b_d) = \bar{F}_f(b_f) \to \rho_1/\Lambda(1-\rho_1) + \rho_2/\Lambda(1-\rho_2)$ when both $b_d \to 0$ and $b_f \to 0$, comparing DD and MTS is meaningful only if $\rho_2/\Lambda(1-\rho_2) \leq \alpha < \rho_1/\Lambda(1-\rho_1) + \rho_2/\Lambda(1-\rho_2)$. If $\alpha < \rho_2/\Lambda(1-\rho_2)$, then DD is not a feasible option and a pure MTS system must be adopted. On the other hand, if $\rho_1/\Lambda(1-\rho_1) + \rho_2/\Lambda(1-\rho_2) < \alpha$, there is no need to hold inventory and a pure MTO system is optimal.

3.2.1. *The effect of loading*

Figure 2 shows the results of our first experiment in which ρ_1 is varied and all other parameters are kept fixed. The ratio of the optimal inventory costs under pure MTS and the DD systems, denoted as z_f^*/z_d^*, is plotted against ρ_1. Recall that $z_f^* = M h_f \bar{I}_f(b_f^*)$ and $z_d^* = h_d \bar{I}_d(b_d^*)$. Both systems meet the same service performance constraint. Therefore, larger values of this ratio indicate a greater relative cost advantage of using DD over MTS.

The effect of ρ_1 is non-monotonic. For $\rho_1 \leq \rho_1^{\min}$, where:

$$\rho_1^{\min} \equiv \frac{\alpha\Lambda(1-\rho_2) - \rho_2}{(1-\rho_2)(1+\alpha\Lambda) - \rho_2},$$
$$z_f^* = z_d^* = 0$$

since both systems operate in the MTO fashion. For ρ_1 slightly larger than $\rho_1^{\min}$, a larger ρ_1 tends to increase the relative cost advantage of DD. This initial increase is an artifact of the integrality of the base-stock levels since the MTS system is required to hold at least one unit of stock

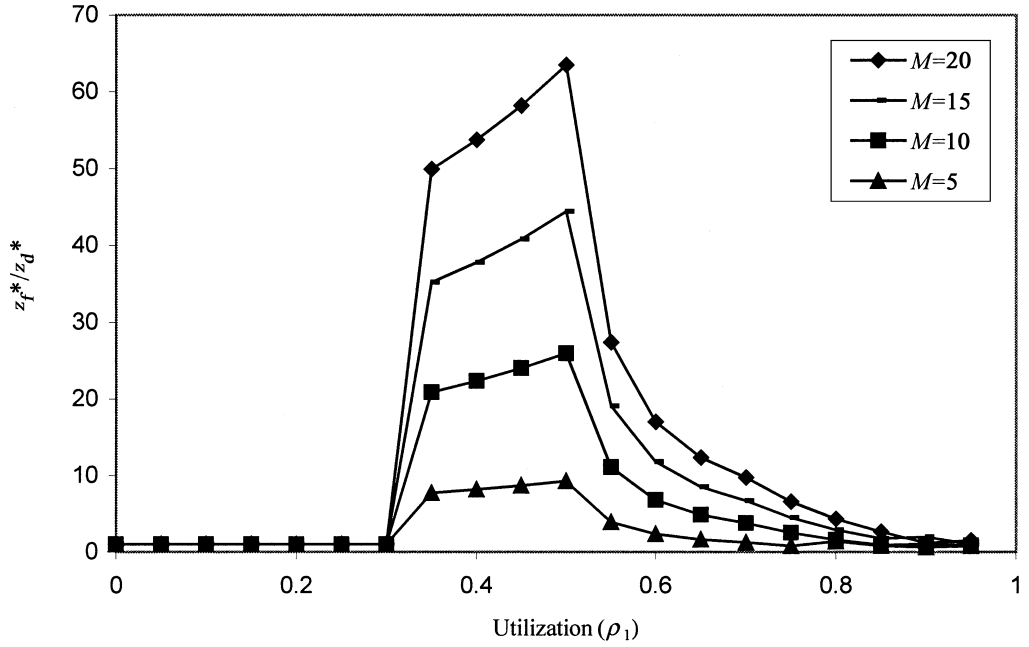

Fig. 2. The effect of utilization of the first stage on the relative benefit of DD ($h_f = 1$, $h_d = 0.5$, $\alpha = 4.5$ and $\rho_2 = 0.8$.

for each product even though a small fractional amount is optimal. The effect diminishes for a higher ρ_1 and the ratio z_f^*/z_d^* decreases in ρ_1. In the limit as $\rho_1 \to 1$, z_f^*/z_d^* approaches h_f/h_d, or equivalently $\bar{I}_f(b_f^*)/\bar{I}_d(b_d^*)$ approaches one. Since in most practical situations, stage-1 utilization is likely to be high, we expect the relative advantage of DD to decrease as stage-1 becomes more capacity constrained. Note that DD is cheaper so long as $h_d \leq h_f$.

Observation 1. *When ρ_1 is sufficiently large, the relative advantage of DD over the pure MTS strategy is diminishing in ρ_1, with $\lim_{\rho_1 \to 1} z_f^*/z_d^* = h_f/h_d$.*

In order to understand on an intuitive level why the ratio z_f^*/z_d^* decreases in ρ_1, note that when ρ_1 is high, stage-1 begins to dominate the replenishment delay. Item-for-item, the semi-finished inventory is less effective in reducing the replenishment time as compared to the finished-goods inventory. At the same time, a high ρ_1 implies that the inter-departure times from stage-1 become more dependent. Since inventory pooling becomes less valuable when the lead time demands of the items being pooled are more positively correlated, DD also becomes less valuable. When ρ_1 is high both the DD and the MTS systems require significant levels of inventory to meet the service-level requirement. Eventually, both b_d^* and b_f^* explode causing the ratio z_f^*/z_d^* to approach the ratio of their holding cost rates (i.e., $\bar{I}_f/\bar{I}_d \to 1$ as $\bar{I}_f, \bar{I}_d \to \infty$).

The ratio z_f^*/z_d^* is affected in a different way by changes in ρ_2. First, note that, similar to ρ_1, when:

$$\rho_2 \leq \rho_2^{\min} \equiv \frac{\alpha\Lambda(1-\rho_1)-\rho_1}{(1-\rho_1)(1+\alpha\Lambda)-\rho_1},$$

no inventory is held under both MTS and DD and the two systems are equivalent. However, in contrast to the effect of ρ_1, the ratio z_f^*/z_d^* is monotonically decreasing in ρ_2 when $\rho_2 > \rho_2^{\min}$. For a sufficiently large ρ_2, DD becomes more expensive than MTS and eventually infeasible. These effects are illustrated in Table 1 for an example system. The main reason why DD is progressively less desirable can be explained by the fact that in the DD system, the only way to compensate for a longer delay at stage-2 is to reduce the delay at stage-1 by holding more semi-finished inventory. The effectiveness of additional semi-finished inventory is increasingly smaller as more inventory is added (see Equation (8)). That forces the DD system to hold a

Table 1. The effect of utilization of the second stage on the ratio z_f^*/z_d^* ($h_f = 1$, $h_d = 0.5$, $\alpha = 5$, and $\rho_1 = 0.8$)

ρ_2	$M = 5$	$M = 10$	$M = 15$	$M = 20$
0.3	1	1	1	1
0.35	25.08	67.78	114.32	162.3
0.4	24.51	66.96	113.38	161.29
0.45	23.87	66.03	112.3	160.12
0.5	23.15	64.94	111.02	158.73
0.55	22.32	63.65	109.5	157.07
0.6	7.63	22.18	38.45	55.37
0.65	7.23	21.51	37.63	54.47
0.7	3.61	11.05	19.56	28.48
0.75	1.50	4.76	8.54	12.54
0.80	1.39	1.59	2.91	4.32
0.81	0.94	1.09	2.00	2.98
0.815	0.66	0.76	1.40	2.09
0.818	0.33	0.38	0.70	1.05

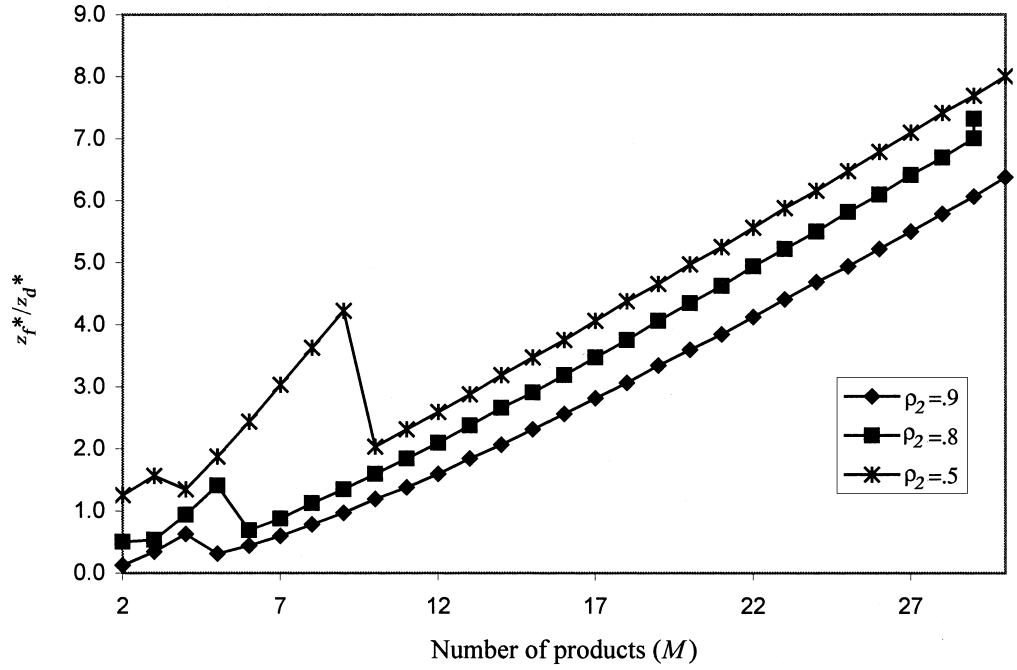

Fig. 3. The effect of the number of products on the relative benefit of DD ($h_f = 1$, $h_d = 0.5$, $\alpha = 9.5$ and $\rho_i = 0.8$).

proportionally greater amount of inventory to meet the target order-fulfillment time.

Observation 2. *DD becomes less desirable with increases in ρ_2. When ρ_2 is sufficiently high, DD becomes an inferior strategy to MTS, and eventually infeasible.*

3.2.2. *The effect of the number of products*

The effect of the number of products on the ratio z_f^*/z_d^* is illustrated in Fig. 3. In this figure, we vary M for a fixed Λ. A larger M therefore corresponds to greater fragmentation of the finished-goods inventory for the MTS system or, equivalently, to a bigger relative advantage for the DD system owing to inventory pooling. In other words, the ratio z_f^*/z_d^* is generally increasing in M.

The MTS system needs to keep at least one unit of inventory in order to satisfy the constraint on the order-fulfillment time. Since individual-item demand, and therefore the need for inventory decreases with M, for sufficiently large M, the MTS system keeps exactly one unit of inventory for each item, making its cost linear in M. In contrast, from Equations (6) and (8), the DD system's order delay and inventory carrying cost remain unaffected. This explains why z_f^*/z_d^* is linear in M for large M. However, since b_f and b_d need to be integer quantities, the effect of increasing M is not always monotone. Specifically, the requirement of an integer b_f can sometimes lead to choosing a higher inventory level than the precise amount needed to meet the order-delay requirement. Consequently, within a limited range, small increases in M can result in a significant decrease in the average inventory (see Fig. 3).

When observing the effect of M on z_f^*/z_d^*, the results observed for a cost-only formulation are different. In a cost-based formulation, when M is sufficiently large, no stock is held in the finished-goods inventory and the MTS system reduces to a pure MTO system. In that region, z_f^*/z_d^* is constant.

Observation 3. *For systems with a service-level constraint, the ratio z_f^*/z_d^* is linear in M when M is large; for cost-based systems, the ratio is unaffected by M for large M.*

3.2.3. *The effect of the service level*

The effect of varying the service level is illustrated in Fig. 4. Here, in order to carry out a meaningful comparison we let $\alpha = \rho_2/\Lambda(1-\rho_2) + \epsilon$, and vary ϵ, where $0 < \epsilon < \rho_1/\Lambda(1-\rho_1)$. For $\epsilon \geq \rho_1/\Lambda(1-\rho_1)$, no inventory is held and production occurs in a MTO fashion. Thus, increasing α means increasing the order-delay slack available to the system with DD. As we can see, a smaller order-delay slack tends to diminish the relative benefit of DD since it requires the system to maintain more inventory. In fact, when ϵ is sufficiently small, DD becomes a more expensive strategy than MTS. This effect is particularly pronounced when the utilization at stage-2 is high. The overall effect of reducing service-level slack is similar to the effect of increasing utilization. Therefore, the behavior of z_f^*/z_d^* can be explained using arguments similar to the case when ρ_1 and ρ_2 are increased. In practice, this means that when customers have little tolerance for delay, DD is not an economic strategy.

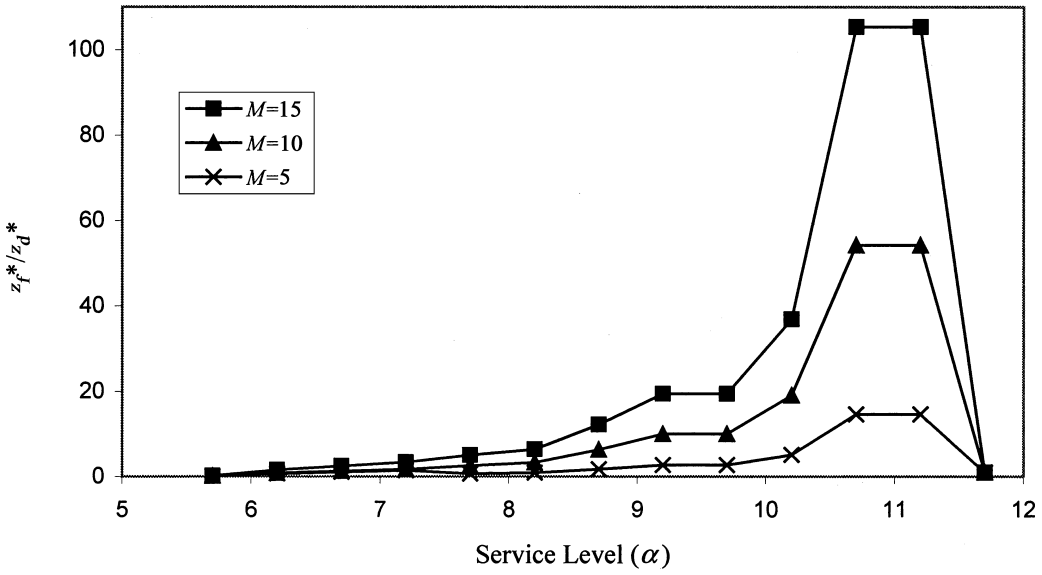

Fig. 4. The effect of the service level on the relative benefit of DD ($h_f = 1$, $h_d = 0.5$ and $\rho_1 = \rho_2 = 0.85$).

Observation 4. *The relative value of DD decreases with tighter service-level requirements. The decrease is particularly significant when the utilization at stage-2 is high.*

4. The optimal point of differentiation

Here we explore the economics of affecting the Point of Differentiation (PoD), when the production process consists of an arbitrary number of stations in series. In this context, choosing a PoD corresponds to locating the pull-push boundary, or equivalently, to the point where we place the buffer of undifferentiated inventory, dividing the production process into MTS and MTO stages. There are many documented examples of companies successfully changing the pull-push boundary. For example, HP used a universal power supply to advance the inventory staging point in their printer supply chain (Feitzinger and Lee, 1997). Benetton developed a dyeing process that allows it to dye knitted sweaters into desired colors (Bruce, 1987). This advanced the semi-finished inventory's stocking point from wool to knitted sweaters. However, such changes do involve additional costs. In what follows, we assume that later differentiation is associated with a higher production system operating cost and a higher per-unit inventory holding cost. The former is the amortized cost of altering the product design and/or the production process to make delayed differentiation a reality; it is also called the product/process redesign cost.

4.1. *The model*

We consider a series production system comprising of K distinct stations/indivisible tasks. Each station is characterized by an exponential processing time with mean $E(T_i) = 1/\mu_i$ and a utilization of $\rho_i = \Lambda E(T_i)$. The choice of the PoD is analogous to choosing the number of stations, k, after which the buffer of undifferentiated inventory is placed (see Fig. 5 for a schematic). The holding cost per unit per unit time of undifferentiated inventory is $h(i)$, if the PoD occurs after station-i. For the moment, we assume that

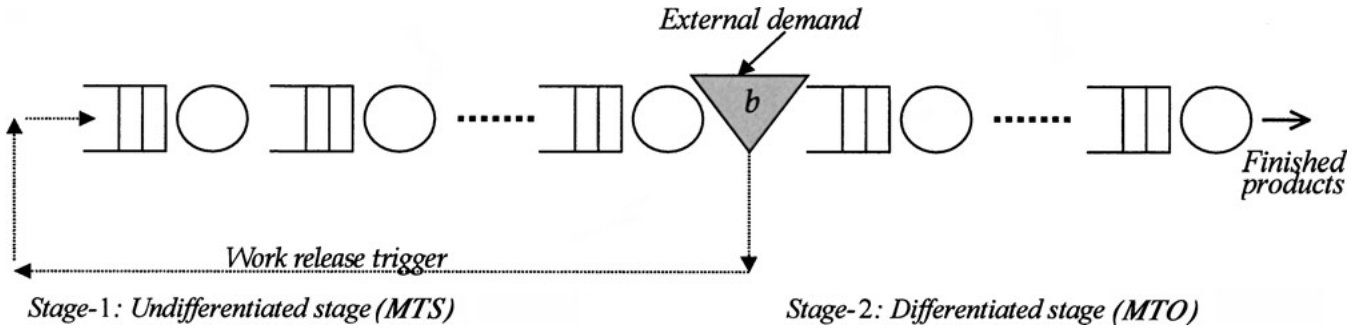

Fig. 5. The serial production system.

there is complete flexibility in choosing k, i.e., $0 \le k \le K$, with $k = 0$ corresponding to a pure MTO system and $k = K$ representing the MTS system with a fully flexible product that can satisfy all types of customer demands. In practice, the latter situation corresponds to the case when product differentiation occurs at the point of sale, for example, in the form of dealer installed options on automobiles.

The production system design problem can be formulated in a variety of ways. For example, we may choose to optimize the inventory holding cost and product/process redesign cost subject to a service-level constraint, as formulated below:

$$\min z(k, b) = h(k)\bar{I}(k, b) + c(k) \tag{13}$$

subject to

$$\bar{F}(k, b) \le \alpha \text{ (or } F_x(k, b) \le \alpha), \tag{14}$$

$$k \le K, \tag{15}$$

$$k, b \ge 0. \tag{16}$$

In the formulation above, we have used b to denote the target base-stock level, the function $c(k)$ to denote the amortized product/process redesign cost when PoD occurs after station k and $F(k, b)$ to denote the associated order-delay distribution. Both $c(k)$ and $h(k)$ are assumed to be non-decreasing in k. In an alternative formulation, we could minimize the sum of inventory holding, backorder and process redesign costs and eliminate the service-level constraint. If the PoD cannot be moved beyond the station indexed $k_{\max}$, then constraint (15) is replaced by $k \le k_{\max}$.

Notice that at any station $j \le k$, the WIP inventory is not affected by the position of the inventory buffer of undifferentiated items since these stations continue to operate as $M/M/1$ queues. This happens because a raw-material kit is released into the system for each order received and the station processing times are assumed invariant as the PoD is moved. For stations indexed $j > k$, the WIP levels are affected by the size of the inventory buffer since it influences the inter-arrival times of the input units to station $k + 1$, which in turn affects station $k + 2$, and so on. Lee and Zipkin (1992) present evidence that even for multistage systems of this kind, modeling each stage as an exogenous supply system with exponentially distributed sojourn times is a good approximation. Since we have only one staging area for planned inventories, their method is equivalent to treating each station-j, $j > k$, as a $M/M/1$ queue. Therefore, the total WIP inventory is not affected significantly by changing the PoD.

Key performance measures of interest are $\bar{I}(k, b)$ and $\bar{F}(k, b)$ (or $F_x(k, b)$). These can be computed using the matrix-geometric approach. Similar arguments have been used to obtain Equation (4) in Proposition 1 (for additional details, see Lee and Zipkin (1992)). In what follows, we sketch out the significant steps of this analysis.

Let $\mathbf{C}_j$ and $\mathbf{P}_j$ be $j \times j$ matrices such that:

$$\mathbf{C}_j = \begin{bmatrix} -\nu_1 & \nu_1 & & & \\ & -\nu_2 & \nu_2 & & \\ & & \ddots & & \\ & & & -\nu_{j-1} & \nu_{j-1} \\ & & & & -\nu_j \end{bmatrix},$$

$\nu_j = \mu_j(1 - \rho_j)$ and $\mathbf{P}_j = \Lambda(\Lambda\mathbf{I} - \mathbf{C}_j)^{-1}$. (Note that $\mathbf{I}$ is the identity matrix of appropriate dimensions.) We also define row vectors $\gamma_j = [1,0,\ldots,0]$ for $j = 1, \ldots, k-1$, $\gamma_k = [\gamma_{k-1}\mathbf{P}^b_{k-1}, (1 - \gamma_{k-1}\mathbf{P}^b_{k-1}\mathbf{e})]$ and $\gamma_{j+1} = [\gamma_j, (1 - \gamma_j\mathbf{e})]$, for $j = k, \ldots, K-1$, where $\mathbf{e}$ is a column vector of ones with similar dimensions as γ_j. Setting $\pi_j = \gamma_j\mathbf{P}_j$ and using Lee and Zipkin's approximation, the backorder level at station j, B_j, has a discrete-phase-type distribution with parameters $(\pi_j, \mathbf{P}_j)$ for $j \ne k$ and parameters $(\pi_k\mathbf{P}^b_k, \mathbf{P}_k)$ for $j = k$. This also means that order delay F_j has a continuous-phase-type distribution with parameters $(\gamma_j, \mathbf{C}_j)$ for $j \ne k$ and $(\gamma_k\mathbf{P}^b_k, \mathbf{C}_k)$ for $j = k$. Furthermore, if we let N_k denote the number of units on order at station k ($N_k = b - I_k + B_k$), then N_k has a discrete-phase-type distribution with parameters $(\pi_k, \mathbf{P}_k)$. Performance measures of interest are now obtained as summarized below.

Proposition 4. *In a series system with the PoD at station-k and buffer size b, the expected inventory, backorder level and average order-fulfillment time at station-j are given as follows.*

$$\bar{I}_j = \begin{cases} 0 & j \ne k, \\ b - \bar{N}_k + \bar{B}_k & \text{where } \bar{N}_k = \pi_k(\mathbf{I} - \mathbf{P}_k)^{-1}\mathbf{e}. \\ & \text{otherwise,} \end{cases} \tag{17}$$

$$\bar{B}_j \approx \begin{cases} \pi_j(\mathbf{I} - \mathbf{P}_j)^{-1}\mathbf{e} & j \ne k, \\ \pi_k P^b_k(\mathbf{I} - \mathbf{P}_k)^{-1}\mathbf{e} & \text{otherwise.} \end{cases} \tag{18}$$

$$\bar{F}_j \approx \bar{B}_j/\Lambda. \tag{19}$$

Furthermore, the probability that the order delay exceeds a quoted lead time x at station j is:

$$\Pr(F_j \ge x) \approx \begin{cases} \gamma_j e^{\mathbf{C}_j x}\mathbf{e} & \text{for } j \ne k, \\ \gamma_j\mathbf{P}^b_j e^{\mathbf{C}_j x}\mathbf{e} & \text{otherwise.} \end{cases} \tag{20}$$

System-level performance measures are $\bar{I}(k, b) = \bar{I}_k$, $\bar{B}(k, b) = \bar{B}_K$, $\bar{F}(k, b) = \bar{F}_K$, *and* $F_x(k, b) = \Pr(F_K \ge x)$.

The approximations in the above expressions come from the fact that for stations $j > k$, we are approximating the queue-length and order-delay distributions by those of independent $M/M/1$ queues.

The expected inventory, backorder level and the order-fulfillment delay can be further simplified after recognizing that the distribution of number of items in the MTS segment

follows a generalized negative binomial distribution. These relationships are given below:

$$\bar{I}(k,b) = \sum_{r=0}^{b} (b-r)\pi_k(r), \tag{21}$$

where

$$\pi_k(r) = \Pr(N_k = r) = \begin{cases} \binom{r+k-1}{k}(1-\rho)^k \rho & \text{if } \rho_j = \rho, \\ \sum_{j=1}^{k} \dfrac{\rho_j^{r+k-1}\prod_{i=1}^{k}(1-\rho_i)}{\prod_{i\neq j}(\rho_j-\rho_i)} & \text{otherwise.} \end{cases} \tag{22}$$

$$\bar{B}(k,b) \approx \bar{I}(k,b) + \sum_{j=1}^{K} \frac{\rho_j}{1-\rho_j} - b. \tag{23}$$

$$\bar{F}(k,b) \approx \bar{B}(k,b)/\Lambda. \tag{24}$$

It is easy to see that for each fixed k, $\bar{I}(k,b)$ is monotonically increasing and $\bar{F}(k,b)$ and $\Pr(F_k \geq x)$ are monotonically decreasing in b. Therefore, given k the value of b that minimizes the holding cost subject to a service-level constraint is the smallest feasible b. In light of this property, it is easy to devise a procedure for finding the optimal k and b. Starting with the smallest feasible k, denoted $k_{\min}$, we calculate contingent optimal $b(k)$ and $z(k, b(k))$, until $k = \min(k_{\max}, K)$. The best (k, b) pair is the one that results in the minimum overall cost.

4.2. *Examples and insights*

Next, we carry out numerical experiments in which we study how the optimal values of k and b and the optimal cost change with respect to changes in the capacity, service level and holding and redesign costs. Results of selected experiments, as well as generalizations supported by these experiments are reported in the remainder of this section. (For brevity, we only consider the case where the objective is to minimize the sum of the inventory holding and process redesign costs, subject to a constraint on the expected order delay.) Note that even without redesign costs, choosing the optimal k is non-trivial. For the same service performance level, b is decreasing in k in a non-linear fashion. Since a larger b results in a greater average inventory, there is a non-trivial trade-off between the unit holding costs (which are increasing in k) and the average inventory (which is decreasing in k). (In this article, increasing means non-decreasing and decreasing means non-increasing. The qualifier "strict" is used to denote a strict inequality.) Some generalizations based on numerical experiments that capture these trade-offs are summarized in Observations 5 and 6 below. We do not have rigorous proofs for the statements made in these observations.

Observation 5. *A series system with balanced stations has the following properties*:

1. *k^* and b^* are decreasing in α (lower $\alpha \Rightarrow$ tighter service requirement) and increasing in Λ (higher utilization). However, due to the requirement that k^* and b^* be integer quantities, this pattern may not hold for small changes in α and Λ.*
2. *k^* is increasing and b^* is decreasing in the unit inventory cost $h(i)$, when $h(i)$ is assumed linear in i.*
3. *k^* is decreasing and b^* is increasing in the redesign cost $c(k)$.*

The above properties are illustrated in Figs. 6–9. Figures 6 and 7 show, respectively, the overall pattern and the non-monotone relationship between k^* and b^* and α for an example system. Both figures use the same basic data: $\rho = 0.85$, $K = 10$ and linear holding and production system costs. In Fig. 6, α is increased in steps of one over the range [1, 57] whereas in Fig. 7, we choose a finer granularity by way of changing α in steps of 0.5 from 35 to 55. An example illustrating the effect of b^* and $c(k)$ on k^* is shown in Fig. 8 and the effect of $c(k)$ on b^* is shown in Fig. 9.

Observation 6. *If there is flexibility in arranging the sequence in which different workstations are visited, it is desirable to place stations with a tighter capacity in the MTS stage. Furthermore, it is desirable to balance stations within each stage but not necessarily across stages.*

We give concrete examples of these trends in Tables 2 and 3. Table 2 data pertains to a system with $K = 4$ stations, exactly two of which are in the MTS stage, i.e., $k = 2$. This production system is comprised of two stations with utilization 0.7 and two stations with utilization 0.8. We assume that any two of these four stations can be placed in the MTS stage. Table 2 shows the optimal b and corresponding cost

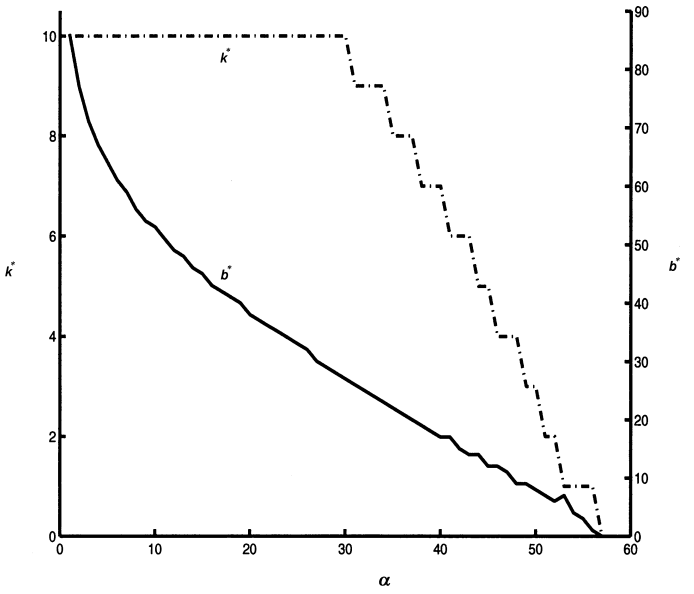

Fig. 6. The effect of the service level on the optimal PoD and buffer size ($K = 10$, $h(i) = 0.5i$, $c(i) = i$ and $\rho = 0.85$).

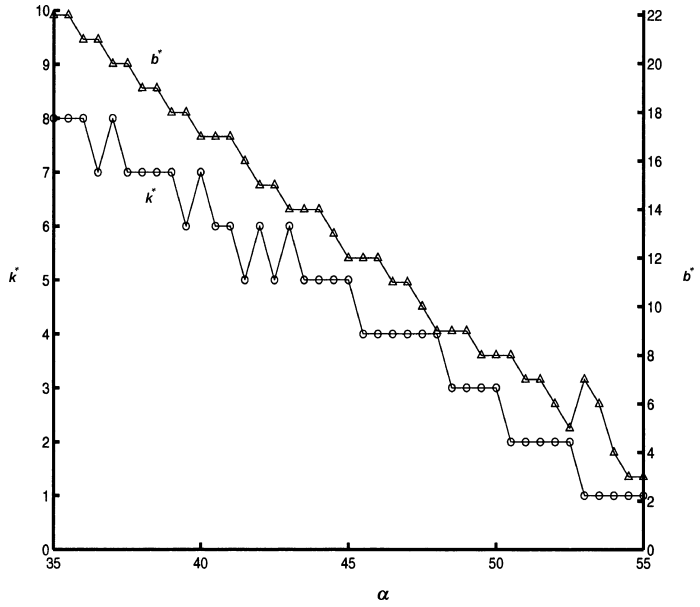

Fig. 7. The effect of the service level on the optimal PoD and buffer size ($K = 10$, $h(i) = 0.5i$, $c(i) = i$ and $\rho = 0.85$).

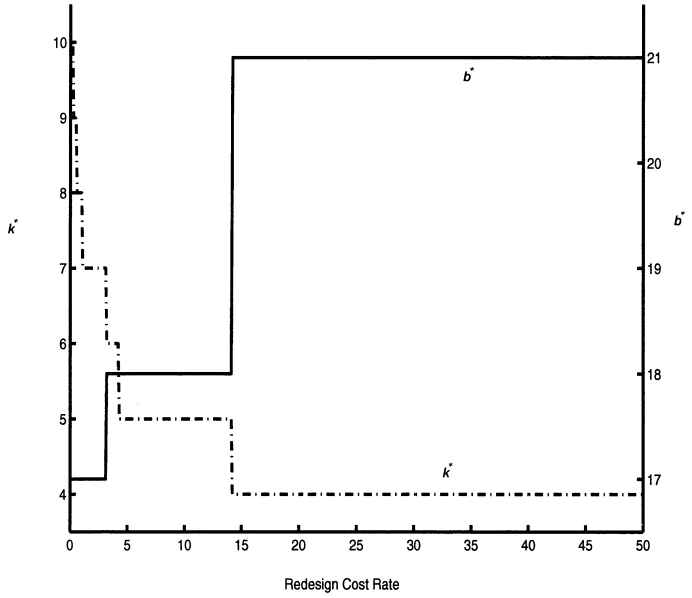

Fig. 9. The effect of the redesign cost rate on the optimal PoD and buffer size ($K = 10, \alpha = 40(\Rightarrow k_{\min} = 3), h(i) = 1.5i, c(i) = n \times i$, where n is varied and $\rho = 0.85$).

under different workstation orderings at the two production stages. Note that the configuration in which the two stations with the heaviest load are placed in the MTS stage minimizes the overall cost. This makes intuitive sense since undifferentiated inventory only reduces the effective delay in the MTS stage.

Table 3 presents evidence that balancing the workload within a production stage is desirable. Its data are: $K = 4$, $k = 3$, $\alpha = 12$ and linear holding and process design costs. The base case consists of four stations with different loads ordered in decreasing utilization. Suppose that the workload could be redistributed among the three stations in the MTS stage. How should the workload be assigned? This example shows that redistributing the load such that workstations within the MTS stage are balanced minimizes the total relevant costs. Redistributing the workload among stations in the MTS stage increases the effective production rate of the bottleneck station in that stage. This is what drives the improvement we observe upon balancing the workload.

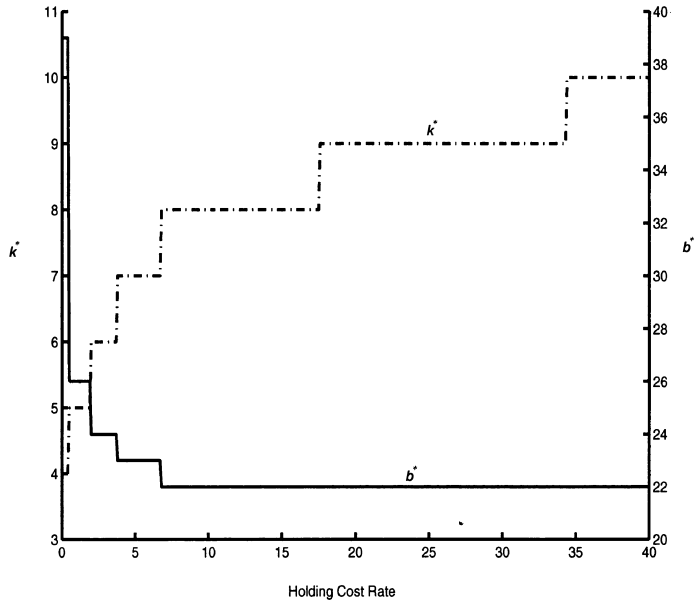

Fig. 8. The effect of the holding cost rate on the optimal PoD and buffer size ($K = 10$, $\alpha = 35(\Rightarrow k_{\min} = 4)$, $h(i) = m \times i$, where m is varied, $c(i) = i$ and $\rho = 0.85$).

5. The effect of partial DD

In the previous two sections, we assumed that a platform common to all products can be built in the undifferentiated segment of the production process. In many industries, this is either not possible or too expensive. Instead, products

Table 2. The optimal ordering of work stations ($K = 4$, $h(i) = 1.25i$, $c(i) = i$, $\alpha = 9$ and $k = 2$)

$(\rho_1, \rho_2, \rho_3, \rho_4)$	b^*	*Cost*
(0.8, 0.8, 0.7, 0.7)	5	4.3304
(0.8, 0.7, 0.8, 0.7)	5	5.1584
(0.7, 0.8, 0.8, 0.7)	5	5.1584
(0.8, 0.7, 0.7, 0.8)	5	5.1584
(0.7, 0.8, 0.7, 0.8)	5	5.1584
(0.7, 0.7, 0.8, 0.8)	7	9.8030

Table 3. The optimal redistribution of workload in the MTS stage ($K = 4$, $h(i) = i$, $c(i) = i$, $\alpha = 12$ and $k = 3$)

$(\rho_1, \rho_2, \rho_3, \rho_4)$	b^*	*Cost*
(0.9, 0.8, 0.7, 0.6)	6	4.2316
(0.9, 0.85, 0.65, 0.6)	7	4.6155
(0.9, 0.75, 0.75, 0.6)	5	3.7999
(0.95, 0.8, 0.65, 0.6)	19	15.5391
(0.85, 0.8, 0.75, 0.6)	3	3.2620
(0.95, 0.75, 0.7, 0.6)	18	14.6133
(0.85, 0.85, 0.7, 0.6)	4	3.4869
(0.8, 0.8, 0.8, 0.6)	2	3.1056
(0.8, 0.9, 0.7, 0.6)	6	4.2316
(0.8, 0.7, 0.9, 0.6)	6	4.2316
(0.9, 0.7, 0.8, 0.6)	6	4.2316
(0.7, 0.9, 0.8, 0.6)	6	4.2316
(0.7, 0.8, 0.9, 0.6)	6	4.2316

are often grouped into families and a standardized platform is designed for each family (Garg and Tang, 1997). Thus, instead of building a single undifferentiated product in stage-1, multiple partially differentiated items are produced and stocked in separate intermediate buffers. These partially differentiated items are then fully differentiated in stage-2 once demand is realized. In this section, we compare systems with full versus partial DD. We are particularly interested in identifying conditions when Partial Differentiation (PD) does not result in a significant deterioration in performance. Note that if we ignore redesign costs, a system with a single undifferentiated product is always superior. However, since the costs of designing a common platform that is shared by all products can be significant (especially when the number of these products is large), there is a need to trade-off the additional benefits of full DD against the cost savings realized with PD.

5.1. *The model*

We use a single station to model each of the MTS and MTO stages of the production process (extension to a system with multiple stations of the type described in Section 4 is possible). For a system with PD, we have M buffers of partially differentiated products. In order to isolate the effect of the number of PD products, we let the demand associated with each PD product be $\lambda = \Lambda/M$. Proposition 4 provides performance measures for systems with PD. The proof is similar to that of Proposition 2 and is omitted for brevity.

Proposition 5. *For a system with M PD products, the average inventory level, the average number of backorders, the average order-fulfillment time and the probability of the fulfillment time exceeding a quoted lead time x, can be obtained respectively as*

$$\bar{I}_{\text{pd}}(b_{\text{pd}}) = M\left(b_{\text{pd}} - \frac{\hat{\rho}_1}{1-\hat{\rho}_1}\left(1-\hat{\rho}_1^{b_{\text{pd}}}\right)\right), \tag{25}$$

$$\bar{B}_{\text{pd}}(b_{\text{pd}}) \approx M\frac{\hat{\rho}_1^{b_{\text{pd}}+1}}{1-\hat{\rho}_1} + \frac{\rho_2}{1-\rho_2}, \tag{26}$$

$$\bar{F}_{\text{pd}}(b_{\text{pd}}) \approx \frac{M\hat{\rho}_1^{b_{\text{pd}}+1}}{\Lambda(1-\hat{\rho}_1)} + \frac{\rho_2}{\Lambda(1-\rho_2)}, \quad \text{and} \tag{27}$$

$$F_{j,x}(b_{\text{pd}}) = F_x(b_{\text{pd}}) = \Pr(F_{\text{pd}}(b_{\text{pd}}) \geq x),$$

$$\approx \begin{cases} \left(1 + \dfrac{\lambda(1-\hat{\rho})\hat{\rho}^{b_{\text{pd}}}x}{\hat{\rho}}\right)e^{-[\lambda(1-\hat{\rho})/\hat{\rho}]x} \\ \qquad\qquad \text{if } \hat{\rho}_1 = \rho_2 = \hat{\rho}, \\ e^{-[\lambda(1-\hat{\rho}_2)/\hat{\rho}_2]x} + \left(\dfrac{(1-\hat{\rho}_2)\hat{\rho}_1^{b_{\text{pd}}+1}}{\hat{\rho}_2 - \hat{\rho}_1}\right) \\ \quad \times (e^{-[\lambda(1-\hat{\rho}_2)/\hat{\rho}_2]x} - e^{-[\lambda(1-\hat{\rho}_1)/\hat{\rho}_1]x}) \\ \qquad\qquad \text{otherwise.} \end{cases} \tag{28}$$

where, as before, $\hat{\rho}_i = \rho_i/(M(1-\rho_i)+\rho_i)$. Note that $F_{j,x}(b_{\text{pd}})$ refers to the waiting-time distribution for items that are made from the type-j undifferentiated platform.

The expressions in Equations (26)–(28) are approximations since the second term in each case is derived after assuming that the stage-2 production system operates approximately like a $M/M/1$ queue. This assumption allows us to simplify the constraint on the average order-fulfillment time as follows:

$$\frac{M\hat{\rho}_1^{b_{\text{pd}}+1}}{\Lambda(1-\hat{\rho}_1)} \leq \hat{\alpha}, \tag{29}$$

where $\hat{\alpha}$ is the allowable delay in stage-1, after netting out the delay in the MTO production stage. According to our approximation, $\hat{\alpha}$ can be estimated as:

$$\hat{\alpha} \approx \alpha - \frac{\rho_2}{\Lambda(1-\rho_2)}.$$

Similarly in a cost-based system we can ignore the component of backorders due to stage-2 in determining the optimal base-stock level.

Optimal base-stock levels can be obtained using the approach presented in Section 3. For systems with a constraint on the average order-fulfillment time, the optimal base-stock level is given by:

$$b^*_{\text{pd}}(\hat{\alpha}) = \left\lceil \frac{\ln[\hat{\alpha}(\mu_1 - \Lambda)]}{\ln(\rho_1/[M(1-\rho_1)+\rho_1])} \right\rceil. \tag{30}$$

Note that we assume $\hat{\alpha} < \rho_1/(\Lambda(1-\rho_1))$. If this condition is not satisfied, then a pure MTO system is optimal. For a cost-based system:

$$b^*_{\text{pd}} = \left\lfloor \frac{\ln[h/(h+\beta)]}{\ln \hat{\rho}_1} \right\rfloor. \tag{31}$$

For a system with a constraint on the probability of the order-fulfillment time exceeding a quoted lead time x, a closed-form for the base-stock level is difficult to obtain. However, it can be easily computed numerically.

The models presented in this section, with appropriate reinterpretation, may also be useful in studying systems where the second stage takes significantly less time than the first one (i.e., differentiation can be rapidly carried out). In that case, both the MTS and DD systems have a single stage, with the difference being in the number of end items stocked.

5.2. *Analysis and insights*

As we did in earlier sections, we compute the ratio of the optimal inventory costs, subject to a constraint on the average order-fulfillment time, under partial DD and full DD. We begin by examining the effect of utilization; since the MTO stage of the production process is not affected by the degree of PD, we focus attention on the effect of changing ρ_1. However, unlike the earlier sections, we are able here to analytically characterize the limiting behavior of optimal cost functions with respect to capacity utilization, the number of partially differentiated items and the order-delay slack. These results agree with the numerical insights obtained in Section 3 (proofs can be found in Appendix B).

Proposition 6. *The ratio* $z^*_{pd}/z^*_d \to h_{pd}/h_d$ *as* $\rho_1 \to 1$, *where* $z^*_{pd} = h_{pd}M\bar{I}_{pd}(b^*_{pd})$ *and* $z^*_d = h_d\bar{I}_d(b^*_d)$, *which means that* $\bar{I}_{pd}(b^*_{pd})/\bar{I}_d(b^*_d) \to 1$ *as* $\rho_1 \to 1$. *Furthermore,* $z^*_{pd} = z^*_d = 0$ *when* $\rho_1 \le \rho_{min} = \hat{\alpha}\Lambda/(1+\hat{\alpha}\Lambda)$.

Proposition 6 shows that the relative advantage of using a single common platform diminishes when the utilization is high, an effect related to the high correlation of the number of items of different platforms that are on order in the PD system. The result also shows that when the utilization is sufficiently low, PD and DD become equivalent since producing to order becomes feasible. On the other hand, when the utilization is in the midrange, the benefits of DD can be significant.

The effect of M on the ratio z^*_{pd}/z^*_d can be shown to be non-monotonic in general. The non-monotonicity is, however, limited to cases when M is relatively small. For large M, cost for the PD system increases linearly in M and, since the cost z^*_d is independent of M, the ratio z^*_{pd}/z^*_d also increases linearly in M.

Proposition 7. *If* $M \ge M_{max} = [\rho_1/(1-\rho_1)][\rho_1/\hat{\alpha}\Lambda(1-\rho_1) - 1]$, *then* $z^*_{pd} = h_{pd}M(1-\hat{\rho}_1)$.

Proposition 8 considers the effect of the order-delay requirement on the ratio z^*_{pd}/z^*_d.

Proposition 8. $z^*_{pd}/z^*_d \to (h_{pd}/h_d)M\ln(\rho_1)/\ln(\hat{\rho}_1)$, *as* $\hat{\alpha} \to 0$. *Furthermore,* $z^*_{pd}/z^*_d = (h_{pd}/h_d)M(1-\hat{\rho}_1)/(1-\rho_1)$ *when* $\rho_1^2/[\Lambda(1-\rho_1)] \le \hat{\alpha} \le \rho_1/\Lambda(1-\rho_1)$ *and* $z^*_{pd} = z^*_d = 0$ *when* $\hat{\alpha} \ge \rho_1/\Lambda(1-\rho_1)$.

In the range $\hat{\alpha} \le \rho_1^2/[\Lambda(1-\rho_1)]$, z^*_{pd}/z^*_d is observed to be mostly increasing (although not monotone) in $\hat{\alpha}$. DD is less valuable when either $\hat{\alpha}$ is small or large enough that MTO production becomes feasible. The value of DD can, however, be significant when $\hat{\alpha}$ is in the mid-range. For example, when $\rho_1^2/[\Lambda(1-\rho_1)] \le \hat{\alpha} \le \rho_1/\Lambda(1-\rho_1)$, the expected inventory under PD is at least M times the expected inventory under complete DD.

6. Conclusions

In this article, we presented mathematical models, analysis and numerical examples to assess the benefits of DD in settings where the lead times are load dependent. Our models capture important interactions due to congestion that are absent in traditional inventory models. The following generalizations are supported by a large number of numerical experiments.

1. The desirability of DD depends on the amount of slack capacity available. A tighter capacity diminishes the value of DD and favors MTS production. This effect has two underlying causes; (i) a higher utilization, at either stage of production, induces the higher inventory levels that are needed to meet service level requirements or to mitigate backorder penalties; and (ii) a higher utilization in the MTS stage increases the positive correlation in the lead times of consecutive orders, thus reducing the value of inventory pooling due to DD.
2. The effects of slack capacity in the MTS and MTO segments of the production process for systems with DD are not symmetric. A tighter capacity in the MTO segment is more detrimental to the desirability of DD since there is no inventory to buffer the longer lead times in the MTO segment.
3. If there is flexibility in choosing the PoD, a higher loading tends to favor later differentiation. Also, whenever there is flexibility in ordering the workstations that constitute the production process, placing workstations that have a tighter capacity in the MTS stage is more effective.

The simplicity and ease of implementation of the models make them useful for strategic decision-making and for building intuition regarding the complex interactions between capacity, congestion, inventory levels, quality of service and cost.

Acknowledgements

The authors are grateful to two anonymous referees for their help in improving an earlier version of this manuscript. This material is based in part upon work supported by the National Science Foundation under grant no. DMII 9988721. Additional funding was provided by the Graduate School of the University of Minnesota through a grant to DG and by the Honeywell Coroporation through a grant to SB.

References

Aviv, Y. and Federgruen, A. (1999) The benefits of design for postponement, in *Quantitative Models for Supply Chain Management*, Tayur, S., Ganeshan, R. and Magazine, M. (eds.), Kluwer, Boston, MA, pp. 553–584.

Aviv, Y. and Federgruen, A. (2001a) Design for postponement: a comprehensive characterization of its benefits under unknown demand distributions. *Operations Research*, **49**, 578–598.

Aviv, Y. and Federgruen, A. (2001b) Capacitated multi-item inventory systems with random and seasonally fluctuating demands: implications for postponement strategies. *Management Science*, **47**, 512–531.

Bruce, L. (1987) The bright new worlds of Benetton. *International Management*, **42**, 24–35.

Buzacott, J.A., Price, S.M. and Shanthikumar, J. G. (1992) Service level in multistage MRP and base stock controlled production systems, in *New Directions for Operations Research in Manufacturing,* Fandel, G., Gulledge, T. and Jones, A. (eds.), Springer-Verlag, Berlin, 445–463.

Buzacott, J.A. and Shanthikumar, J.G. (1993) *Stochastic Models of Manufacturing Systems*, Prentice Hall, Englewood Cliffs, NJ.

Chao, X. and Zheng, S. (1998) A result on networks of queues with customer coalescence and state-dependent signaling. *Journal of Applied Probability*, **35**, 151–164.

De Vericourt, F., Karaesmen, F. and Dallery, Y. (2000) Dynamic scheduling in a make-to-stock system: a partial characterization of optimal policies. *Operations Research*, **48**, 811–819.

Eppen, G. (1979) Effects of centralization on expected costs in multilocation newsboy problems. *Management Science*, **25**, 498–501.

Feitzinger, E. and Lee, H. (1997) Mass customization at Hewlett Packard: the power of postponement. *Harvard Business Review*, **75**, 116–121.

Fisher, M., Ramdas, K. and Ulrich, K. (1999) Component sharing in management of product variety. *Management Science*, **45**, 297–315.

Garg, A. and Tang, C.S. (1997) On postponement strategies for product families with multiple points of differentiation. *IIE Transactions*, **29**, 641–650.

Graman, G.A. and Magazine, M.J. (1998) An analysis of packaging postponement. in *Proceedings of the 1998 MSOM Conference*, Seattle, WA, pp. 67–72.

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

Hopp, W. and Spearman, M.L. (2000) *Factory Physics*, 2nd edn., Irwin/McGraw-Hill, New York, NY.

Jackson, J.R. (1957) Networks of waiting lines. *Operations Research*, **5**, 518–521.

Kleinrock, L. (1975) *Queueing Systems*, Vol. I: Theory, Wiley, New York, NY.

Lee, H.L. (1996) Effective inventory and service management through product and process redesign. *Operations Research*, **44**, 151–159.

Lee, H.L. and Billington, C. (1994) Designing products and processes for postponement, in *Management of Design: Engineering and Management Perspectives*, Dasu, S. and Eastman, C. (eds.), Kluwer, Boston, MA, pp. 105–122.

Lee, H.L. and Tang, C. S. (1997) Modeling the costs and benefits of delayed product differentiation. *Management Science*, **43**, 40–53.

Lee, Y. and Zipkin, P. (1992) Tandem queues with planned inventories. *Operations Research*. **40**, 936–947.

Lee, Y. and Zipkin, P. (1995) Processing networks with inventories: sequential refinement systems. *Operations Research*. **43**, 1025–1036.

Magretta, J. (1998) The power of virtual integration: An interview with Dell Computer's Michael Dell. *Harvard Business Review*, **76**, 72–84.

Swaminathan, J.M. and Tayur, S.R. (1998) Managing broader product lines through delayed differentiation using vanilla boxes. *Management Science*, **44**, S161–S172.

Swaminathan, J.M. and Tayur, S.R. (1999) Managing design of assembly sequences for product lines that delay product differentiation. *IIE Transactions*, **31**, 1015–1027.

Wein, L.M. (1992) Dynamic scheduling of a multiclass make-to-stock queue. *Operations Research*, **40**, 724–735.

Zipkin, P.H. (1995) Performance analysis of a multi-item production-inventory system under alternative policies. *Management Science*, **41**, 690–703.

Appendices

Appendix A: Performance metrics for a MTS system

Since demand is Poisson and processing times are exponentially distributed the two stations behave like $M/M/1$ queues in tandem. Therefore:

$$\pi(r) = \sum_{r_1=0}^{r} \pi_1(r_1)\pi_2(r-r_1) = \begin{cases} (r+1)(1-\rho)^2\rho^r, & \text{if } \rho_1 = \rho_2 = \rho, \\ \dfrac{(1-\rho_1)(1-\rho_2)\left[\rho_2^{r+1} - \rho_1^{r+1}\right]}{(\rho_2 - \rho_1)}, & \text{otherwise.} \end{cases} \tag{A1}$$

Given r jobs in the system, the number of type-j jobs is binomially distributed with parameters r and $\lambda_j/\Lambda = 1/M$. Thus, we find probability $p(k_j)$ of k_j type-j jobs at an arbitrary moment of observation as:

$$p(k_j) = \sum_{r=k_j}^{\infty} p_{k_j,r}\pi(r), \tag{A2}$$

$$p_{k_j,r} = \frac{r!}{k_j!(r-k_j)!}\left(\frac{1}{M}\right)^{k_j}\left(\frac{M-1}{M}\right)^{r-k_j}, \quad 0 \le k_j \le r. \tag{A3}$$

Equation (A2) can be simplified to give:

$$p(k_j) = \begin{cases} (1+k_j)(1-\hat{\rho})^2\hat{\rho}^{k_j} & \text{if } \rho_1 = \rho_2 = \rho, \\ \dfrac{(1-\hat{\rho}_1)(1-\hat{\rho}_2)[\hat{\rho}_2{}^{k_j+1} - \hat{\rho}_1{}^{k_j+1}]}{\hat{\rho}_2 - \hat{\rho}_1} & \text{otherwise,} \end{cases} \tag{A4}$$

where $\hat{\rho} = \rho/[M(1-\rho)+\rho]$, and $\hat{\rho}_i = \rho_i/[M(1-\rho_i)+\rho_i]$. Comparing Equations (A1) and (A4), we notice that the distribution of number of type-j jobs in the system is identical to the distribution of any r jobs in the system, with the difference that ρ_i is replaced by $\hat{\rho}_i$. In fact, it can be shown that the queue-occupancy level by type-j jobs in each of the two queues has a geometric distribution with parameter $\hat{\rho}_i$. That is, we can obtain performance metrics for each job type by considering an independent queueing system comprising only of these job types. In the equivalent system, jobs have an arrival rate of $\lambda = \Lambda/M$, the service rate in queue-i is $\hat{\mu}_i = \mu_i - (M-1)\lambda$ and the utilization is $\hat{\rho}_i$.

Consider the average inventory of type-j finished goods denoted as $\bar{I}_{f,j}$. The inventory buffer is not empty only if

the number of type-j jobs in process is less than b_f. Thus, by definition

$$\bar{I}_{f,j}(b_f) = \sum_{k_j=0}^{b_f} (b_f - k_j)p(k_j), \tag{A5}$$

and

$$\bar{I}_f(b_f) = \sum_{j=1}^{M} \bar{I}_{f,j}(b_f). \tag{A6}$$

Substituting from Equation (A4) and simplifying leads to Equation (1). In order to obtain the average order delay, we first compute the average number of backorders of type-j jobs (denoted by $\bar{B}_{f,j}$) as: $\bar{B}_{f,j}(b_f) = \sum_{k_j=b_f}^{\infty}(k_j - b_f)p(k_j)$. Upon substituting from Equation (A4) and simplifying, we obtain:

$$\bar{B}_{f,j}(b_f) = \begin{cases} b_f\hat{\rho}^{b_f+1} + \dfrac{2\hat{\rho}^{b_f+1}}{(1-\hat{\rho})} & \\ & \text{if } \rho_1 = \rho_2 = \rho, \\ M\dfrac{(1-\rho_1)(1-\rho_2)}{\rho_2-\rho_1}\left[\dfrac{\hat{\rho}_2{}^{b_f+2}}{(1-\hat{\rho}_2)^2} - \dfrac{\hat{\rho}_1{}^{b_f+2}}{(1-\hat{\rho}_1)^2}\right] & \\ & \text{otherwise.} \end{cases} \tag{A7}$$

Equation (3) is obtained from the fact that $\bar{B}_f(b_f) = M\bar{B}_{f,j}(b_f)$ and after substituting from Equations (A7) and (1). Equation (2) is obtained from Equation (A7) by noting that $\bar{F}_f(b_f) = \bar{F}_{f,j}(b_f) = M\bar{B}_{f,j}(b_f)/\Lambda$.

Since the distribution of order delay is the same for all job types, we focus in the remainder of this Appendix on an arbitrary job type. Let N_i denote the number of items that still need to be produced at stage-i in order to bring the finished-goods inventory level back to b_f. Using the fact that each stage is a $M/M/1$ queue, it is easy to see that the equilibrium N_i, $i = 1, 2$ has a discrete-phase-type distribution with parameters (p_i, P_i), where $p_1 = \hat{\rho}_1 = P_1, p_2 = [\hat{\rho}_1, \hat{\rho}_2(1-\hat{\rho}_1)]$ and

$$P_2 = \begin{bmatrix} \hat{\rho}_1 & \hat{\rho}_2(1-\hat{\rho}_1) \\ 0 & \hat{\rho}_2 \end{bmatrix},$$

which means that $\Pr[N_i > m] = p_i\mathbf{P}_i^m\mathbf{e}$, $\mathbf{e}$ being the column vector of ones. By the spectral decomposition of the matrix $\mathbf{P}_2$, we can compute the powers of the matrix easily as

$$\mathbf{P}_2^m = \begin{bmatrix} 1 & \dfrac{\hat{\nu}_1}{\hat{\nu}_1 - \hat{\nu}_2} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\rho}_1^m & 0 \\ 0 & \hat{\rho}_2^m \end{bmatrix} \begin{bmatrix} 1 & \dfrac{-\hat{\nu}_1}{\hat{\nu}_1 - \hat{\nu}_2} \\ 0 & 1 \end{bmatrix},$$

where $\hat{\nu}_i = \hat{\mu}_i(1-\hat{\rho}_i)$, $i = 1, 2$. The order delay is distributed as a continuous-phase-type distribution with parameters $(\gamma_2\mathbf{P}_2^{b_f}, \mathbf{C}_2)$, where $\gamma_2 = [\hat{\rho}_1, 1-\hat{\rho}_1]$ and

$$\mathbf{C}_2 = \begin{bmatrix} -\hat{\nu}_1 & \hat{\nu}_1 \\ 0 & -\hat{\nu}_2 \end{bmatrix},$$

which implies that $\Pr[F_f(b_f) \geq x] = \gamma_2\mathbf{P}_2^{b_f}e^{\mathbf{C}_2 x}\mathbf{e}$. The simplified expression for customer-delay distribution is shown in Equation (4).

Appendix B: Properties for a system with partial DD

Proof of Proposition 6. Recognizing that $\hat{\rho}_1/(1-\hat{\rho}_1) = \rho_1/(M(1-\rho_1))$ the ratio z^*_{pd}/z^*_d can be rewritten as:

$$z^*_{pd}/z^*_d = (h_{pd}/h_d) \times \frac{M(1-\rho_1)\lceil \ln[\hat{\alpha}\mu_1(1-\rho_1)]/\ln[\hat{\rho}_1]\rceil - \rho_1\left(1-\hat{\rho}_1^{\lceil\hat{\rho}_1\rceil}\right)}{(1-\rho_1)\lceil \ln[\hat{\alpha}\mu_1(1-\rho_1)]/\ln[\rho_1]\rceil - \rho_1\left(1-\rho_1^{\lceil \ln[\hat{\alpha}\mu_1(1-\rho_1)]/\ln[\rho_1]\rceil}\right)}, \tag{A8}$$

where $\lceil t \rceil$ is the integer ceiling of t. Noting that:

$$\left\lceil \frac{\ln[\hat{\alpha}\mu_1(1-\rho_1)]}{\ln[\hat{\rho}_1]} \right\rceil \quad \text{and} \quad \left\lceil \frac{\ln[\hat{\alpha}\mu_1(1-\rho_1)]}{\ln[\rho_1]} \right\rceil,$$

are very large numbers when $\rho_1 \to 1$, we can ignore integrality (i.e. $\lim_{x\to\infty}\lceil x\rceil/x = 1$ for $x \in \mathbb{R}$). Then, taking advantage of the fact that $a^{f(x)} = e^{\ln(a)f(x)}$, it is straightforward to show that:

$$\lim_{\rho_1\to 1} z^*_{pd}/z^*_d = (h_{pd}/h_d) \times \lim_{\rho_1\to 1} \frac{M(1-\rho_1)(\ln[\hat{\alpha}\mu_1(1-\rho_1)]/\ln[\hat{\rho}_1]) - \rho_1(1-\hat{\alpha}\mu_1(1-\rho_1))}{(1-\rho_1)(\ln[\hat{\alpha}\mu_1(1-\rho_1)]/\ln[\rho_1]) - \rho_1(1-\hat{\alpha}\mu_1(1-\rho_1))} \tag{A9}$$

Let

$$f_1(\rho_1) = M(1-\rho_1)\frac{\ln[\hat{\alpha}\mu_1(1-\rho_1)]}{\ln[\hat{\rho}_1]},$$
$$f_2(\rho_1) = (1-\rho_1)\frac{\ln[\hat{\alpha}\mu_1(1-\rho_1)]}{\ln[\rho_1]},$$

and $f_0(\rho_1) = \rho_1(1-\hat{\alpha}\mu_1(1-\rho_1))$. Then, we can rewrite Equation (A9) as:

$$\lim_{\rho_1\to 1} z^*_{pd}/z^*_d = (h_{pd}/h_d)\lim_{\rho_1\to 1}\left(\frac{f_1(\rho_1)-f_0(\rho_1)}{f_1(\rho_1)}\right) \times \left(\frac{f_2(\rho_1)}{f_2(\rho_1)-f_0(\rho_1)}\right)\left(\frac{f_1(\rho_1)}{f_2(\rho_1)}\right). \tag{A10}$$

In order to show that $\lim_{\rho_1\to 1} z^*_{pd}/z^*_d = h_{pd}/h_d$, it is sufficient to show that:

$$\lim_{\rho_1\to 1}\left(\frac{f_1(\rho_1)-f_0(\rho_1)}{f_1(\rho_1)}\right) = \lim_{\rho_1\to 1}\left(\frac{f_2(\rho_1)}{f_2(\rho_1)-f_0(\rho_1)}\right) = \lim_{\rho_1\to 1}\left(\frac{f_1(\rho_1)}{f_2(\rho_1)}\right) = 1. \tag{A11}$$

First, note that $\lim_{\rho_1 \to 1} f_0(\rho_1) = 1$ and

$$\lim_{\rho_1 \to 1} f_1(\rho_1) = \lim_{\rho_1 \to 1} \frac{M(1-\rho_1)\ln[\hat{\alpha}\mu_1(1-\rho_1)]}{\ln[\rho_1] - \ln[M(1-\rho_1)+\rho_1]},$$

which, upon applying L'Hopital's rule, reduces to:

$$\lim_{\rho_1 \to 1} f_1(\rho_1) = \lim_{\rho_1 \to 1} \frac{-M\ln[\hat{\alpha}\mu_1(1-\rho_1)] - M}{1/\rho_1 - (1-M)/[M(1-\rho_1)+\rho_1]} = \infty. \quad \text{(A12)}$$

Similarly, we can show that $\lim_{\rho_1 \to 1} f_2(\rho_1) = \infty$. Consequently, we have:

$$\lim_{\rho_1 \to 1} \frac{f_1(\rho_1) - f_0(\rho_1)}{f_1(\rho_1)} = \lim_{\rho_1 \to 1} \left(1 - \frac{f_0(\rho_1)}{f_1(\rho_1)}\right) = 1,$$

and

$$\lim_{\rho_1 \to 1} \frac{f_2(\rho_1)}{f_2(\rho_1) - f_0(\rho_1)} = \lim_{\rho_1 \to 1} \frac{1}{1 - f_0(\rho_1)/f_2(\rho_1)} = 1.$$

Finally, we have:

$$\lim_{\rho_1 \to 1} \frac{f_1(\rho_1)}{f_2(\rho_1)} = \lim_{\rho_1 \to 1} \frac{M\ln(\rho_1)}{\ln(\hat{\rho}_1)}.$$

Since both the numerator and denominator are zero as $\rho_1 \to 1$, we apply L'Hopital's rule to obtain:

$$\lim_{\rho_1 \to 1} \frac{f_1(\rho_1)}{f_2(\rho_1)} = \lim_{\rho_1 \to 1} \frac{M/\rho_1}{1/\rho_1 - (1-M)/(M(1-\rho_1)+\rho_1)} = 1.$$

Hence, $\lim_{\rho_1 \to 1} z^*_{\mathrm{pd}}/z^*_{\mathrm{d}} = h_{\mathrm{pd}}/h_{\mathrm{d}}$. ■

Proof of Proposition 7. Since the optimal buffer size is given by:

$$b^*_{\mathrm{pd}} = \left\lceil \frac{\ln[\hat{\alpha}(\mu_1 - \Lambda)]}{\ln[\hat{\rho}_1]} \right\rceil,$$

$b^*_{\mathrm{pd}} = 1$ if $\ln[\hat{\alpha}(\mu_1 - \Lambda)]/\ln[\hat{\rho}_1] \le 1$, or equivalently:

$$M \ge \frac{\rho_1}{(1-\rho_1)}\left[\frac{\rho_1}{\hat{\alpha}\Lambda(1-\rho_1)} - 1\right].$$

Substituting $b^*_{\mathrm{pd}} = 1$ in the objective function leads to $z^*_{\mathrm{pd}} = h_{\mathrm{pd}}M(1-\hat{\rho}_1)$. ■

Proof of Proposition 8. Noting that:

$$\left\lceil \frac{\ln[\hat{\alpha}\mu_1(1-\rho_1)]}{\ln[\hat{\rho}_1]} \right\rceil \text{ and } \left\lceil \frac{\ln[\hat{\alpha}\mu_1(1-\rho_1)]}{\ln[\rho_1]} \right\rceil,$$

are very large numbers when $\hat{\alpha} \to 0$, we ignore integrality. Then, the limit can be written as

$$\lim_{\hat{\alpha} \to 0} z^*_{\mathrm{pd}}/z^*_{\mathrm{d}} = \lim_{\hat{\alpha} \to 0}(h_{\mathrm{pd}}/h_{\mathrm{d}}) \times \frac{M(\ln[\hat{\alpha}\mu_1(1-\rho_1)]/\ln[\hat{\rho}_1]) - \rho_1/(1-\rho_1)(1-\hat{\alpha}\mu_1(1-\rho_1))}{(\ln[\hat{\alpha}\mu_1(1-\rho_1)]/\ln[\rho_1]) - \rho_1/(1-\rho_1)(1-\hat{\alpha}\mu_1(1-\rho_1))}. \quad \text{(A13)}$$

Since the limit of both numerator and denominator is infinity, we apply L'Hopital's rule to obtain:

$$\lim_{\hat{\alpha} \to 0} z^*_{\mathrm{pd}}/z^*_{\mathrm{d}} = \lim_{\hat{\alpha} \to 0}(h_{\mathrm{pd}}/h_{\mathrm{d}}) \frac{M}{\ln[\hat{\rho}_1]} \Big/ \frac{1}{\ln[\rho_1]} = (h_{\mathrm{pd}}/h_{\mathrm{d}}) \frac{M\ln\rho_1}{\ln\hat{\rho}_1}. \quad \text{(A14)}$$

When $\hat{\alpha} > 1/(\mu_1 - \Lambda)$, $b^*_{\mathrm{d}} = b^*_{\mathrm{pd}} = 0$ and, consequently, $z^*_{\mathrm{d}} = z^*_{\mathrm{pd}} = 0$. When $\rho_1^2/(\Lambda(1-\rho_1) \le \hat{\alpha} \le 1/(\mu_1 - \Lambda)$, $b^*_{\mathrm{d}} = 1$. Since in this region, $b^*_{\mathrm{pd}} \ge 1$ and since $b^*_{\mathrm{pd}} \le b^*_{\mathrm{d}}$, it follows that $b^*_{\mathrm{pd}} = 1$. Substituting $b^*_{\mathrm{d}} = b^*_{\mathrm{pd}} = 1$ in z^*_{d} and z^*_{pd} respectively, we obtain $z^*_{\mathrm{pd}}/z^*_{\mathrm{d}} = (h_{\mathrm{pd}}/h_{\mathrm{d}})M(1-\hat{\rho}_1)/(1-\rho_1)$. ■

Biographies

Diwakar Gupta is a Professor of Mechanical Engineering at the University of Minnesota where he teaches in the Graduate Program in Industrial Engineering. He received his Ph.D. in Management Sciences from the University of Waterloo (Canada). Before joining the University of Minnesota, he held Faculty appointments at the Technical University of Nova Scotia (now part of Dalhousie University) and at McMaster University (both in Canada). His primary research interest lies in the area of stochastic modeling with applications to manufacturing systems, inventory management and healthcare delivery systems. He is a Departmental Editor of *IIE Transactions—Scheduling and Logistics*, and an Editorial Board Member of the *M&SOM Journal*. In addition, he is an Associate Editor with the *International Journal of Flexible Manufacturing Systems*.

Saif Benjaafar is a Professor in the Department of Mechanical Engineering at the University of Minnesota where he also serves as Director for the Center for Supply Chain Research and Director for the Graduate Program in Industrial Engineering. He was a Distinguished Senior Visiting Scientist with Honeywell Laboratories. He has also been a Visiting Professor at Ecole Centrale Paris and Hong Kong University of Science and Technology. His research is in the area of supply chain management, production and inventory systems, and manufacturing and service operations. He serves as an Associate Editor for *IEEE Transactions on Automation Science and Engineering, IIE Transactions, and International Journal of Flexible Manufacturing Systems*. He holds Ph.D. and MS degrees in Industrial Engineering from Purdue University and a BS degree in Electrical Engineering from the University of Texas at Austin.

Contributed by the Engineering Statistics and Applied Probability Department