McMASTER
•UNIVERSITY•

MICHAEL G. DeGROOTE
SCHOOL OF BUSINESS

RESEARCH AND WORKING PAPER SERIES

MAKE-TO-ORDER, MAKE-TO-STOCK, OR DELAY PRODUCT DIFFERENTIATION? – A COMMON FRAMEWORK FOR MODELING AND ANALYSIS

By

Diwakar Gupta and **Saifallah Benjaafar**

Michael G. DeGroote School of Business
McMaster University
Hamilton, Ontario

Working Paper # 434

April, 1999

MAKE-TO-ORDER, MAKE-TO-STOCK, OR DELAY PRODUCT DIFFERENTIATION? – A COMMON FRAMEWORK FOR MODELING AND ANALYSIS

By

Diwakar Gupta and **Saifallah Benjaafar**

Michael G. DeGroote School of Business
McMaster University
Hamilton, Ontario

Working Paper # 434

April, 1999

Make-to-order, Make-to-stock, or Delay Product Differentiation? - A Common Framework for Modeling and Analysis

Diwakar Gupta • Saifallah Benjaafar

McMaster University, Michael G. DeGroote School of Business, Hamilton, ON L8S 4M4
University of Minnesota, Dept. of Mechanical Engineering, Minneapolis, MN 55455

Working Paper # 434, Michael G. DeGroote School of Business

April, 1999

Abstract

Widespread increase in product variety and the simultaneous emphasis on shorter order delivery times and lower costs have increased the strategic importance of how much and where (in the production process) inventories should be maintained. In this article, we develop models that can be used to rapidly investigate a strategy in which product differentiation is delayed through product and process redesign, resulting in a two stage production process. Stage-1 produces undifferentiated items to stock to fill a buffer of size b. Stage-2 makes customized products from the stock of undifferentiated items after demand materializes. Subject to a service level constraint, we determine the optimal buffer size and the optimal workload allocation to the two production stages. This model captures both make-to-order and make-to-stock situations as special cases. The former occurs when buffer size is zero and the latter when the entire workload is allocated to stage-1. Numerical experiments reveal that the degree to which postponement of differentiation is desirable depends mainly on the relative magnitudes of the system workload, the response time limit, and whether or not capacity can be flexibly allocated. These same parameters also determine the optimal size of the buffer. Surprisingly, we find that the optimal point of differentiation and buffer size are largely insensitive to the choice of holding, warehousing, and process redesign costs. The desirability of using delayed differentiation increases with product variety, and the relative magnitude of this increase is higher when greater product variety is accompanied by lower production volume of each item produced.

---

1 Introduction

In today's business environment, a manufacturing firm that has the ability to respond quickly to market changes as well as to produce a variety of different product types enjoys a competitive advantage. A major challenge for managers of such firms is determining how much strategic inventory to keep, and at which stocking points in the production system. Whereas produce-to-order and produce-to-stock represent the two ends of a spectrum, the set of available choices also includes options to maintain semi-finished goods inventories at one or more stocking points and to delay product differentiation. Maintaining stocks of semi-finished goods may reduce order fulfillment delay without increasing inventory carrying costs to the same extent as might happen if finished goods stocks are the only stocks maintained. This is more likely to be true when cost of carrying inventories rises rapidly with increasing value-added.

Delayed differentiation (DD) strategy offers two key additional benefits. It can help reduce order fulfillment time and drive inventory carrying costs even lower by taking advantage of the inventory pooling effect (Lee and Billington, 1994). For industries characterized by high market mediation costs (sum of the cost of lost sales due to lengthy order fulfillment delays and the cost of mark-downs on account of product obsolescence), DD also provides a hedge against market uncertainty since the same undifferentiated product can be used to make several different finished products. Typically, the benefits mentioned above need to be balanced against additional costs arising from product and process redesign, use of extra materials (where common designs are made possible by having redundant parts), and less efficient processing (when common processing leads to the use of a less specialized production equipment or greater yield losses). The sum total of these effects, however, is believed to be positive. Recent literature showcases several examples in which manufacturers of household appliances, electronic goods and apparel have successfully used this approach to control inventory costs while maintaining high service levels (see, for example, Lee and Tang (1997), Lee (1996), and Swaminathan and Tayur (1996)).

The goal of this article is to provide additional insights by explicitly modeling the dynamic interaction between desired service level, expressed in terms of order fulfillment delay, and economic consequences, measured by the sum of inventory and product/process redesign costs. 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). Our models can treat either of these points of view and capture the manner in which order delay depends on the amount of slack capacity in a manufacturing system, and on the flexibilities of its work centers and workers. Our choice of order delay as the measure of service stems from our observation that most applications of DD arise in situations where quick response to

customer orders is key to competitiveness. Alternative measures of customer service are possible by including, for example, an inventory backordering cost, or placing a constraint on the probability of backorders exceeding some threshold. These alternative measures can be easily accommodated using our models.

The usefulness of our models to operations managers is as a strategic tool that can allow them to rapidly examine key tradeoffs from different process design choices and inventory keeping policies. Consistent with this spirit, we have deliberately kept the models simple and relevant for gaining managerial insights. The production system is represented by two stages. Stage-1 produces to stock and manufactures undifferentiated items. Any remaining production steps are carried out at stage-2 in a make-to-order fashion. An inventory buffer separates the two stages. It is used to store output from stage-1 until demand materializes. This model can be used to determine the economically optimal point of differentiation, by choosing the amount of work content to be allocated to each stage, and the size of buffer for undifferentiated items such that the specified service level constraint is met.

We present two variations of the two-stage production system described above. In the first model, each stage is represented by a single workstation/worker. This model arises in a situation where the processing rate at a stage cannot be altered, but we may choose to assign unequal work load to the two stages. In the second model, we allow each stage to consist of several parallel workstations/workers. Here, we assume that workers are cross-trained and equipment is flexible, so that they can be assigned to either production stage. Thus, we have the ability in the second model to alter both the processing rate and the amount of work load assigned to each production stage.

We show that DD is not always superior to make-to-stock in terms of minimizing either average inventory costs or order delays. This contradicts conclusions reached in existing literature based on models that do not account for congestion effects. We find the optimal point of differentiation and the optimal buffer size to be highly sensitive to the relative magnitudes of total work content, the order delay requirement, and to whether or not capacity can be flexibly allocated between the two stages. Remarkably, the optimal solution is quite insensitive to the choice of holding, process redesign, and warehousing costs. For the model with inflexible workforce, we show that in order to minimize order delay, at least 50% of total work should be done in a make-to-stock fashion. This means that it is always optimal to delay the differentiation of at least half of the work. We find that this fraction increases when the total work content is small. In other words, the proportion of work done in a make to stock fashion is minimum (i.e., proportional delay in differentiation is minimum) when work content is high, even though one would intuitively expect the opposite to be true. Similarly, for the case of flexible workforce, we show that order delay is minimized when the proportion of work that is undifferentiated is at least equal to the fraction of capacity that is assigned to the make-to-stock stage (provided that this stage has at least half of the total capacity). Moreover, when capacity can be flexibly allocated,

we show that adding greater capacity to a particular stage does not necessarily result in a greater load being assigned to that stage. In fact, we find instances where idling some of the workers is optimal.

The models reported here are, in part, inspired by the case of a manufacturing company, Recovery Engineering, with which one of the authors has interacted. The company, a leading manufacturer and supplier of household water filtration products, stocks over 150 finished products sold at 11 retail chains in the U. S. and Japan under their own label (PŪR) and under a different label in Europe through a second party distributor (ROWENTA). Products fall into four major categories: faucet-mounted systems, counter top systems, under-sink systems, and pitchers. The company also stocks replacement filters which are sold separately in packages of varying quantities. Within each category, products differ by filter quality, water output system (stream or spray), labeling, and packaging. In the U. S., most retailers require customized labeling and packaging (e.g. different size unit packages, number of spare filters per unit package, and different number of unit packages per master-pack). Also, they keep very limited stock of their own, and require the supplier to operate in a just-in-time delivery mode. The supplier, Recovery Engineering, has been under pressure to keep order lead times short or risk losing market share to the other major competing brand (BRITA). In an effort to reduce inventory costs without compromising response time, the company is contemplating using delayed differentiation. Delayed differentiation can occur at different stages. For example, final packaging could be delayed until actual orders are received. Alternatively, labeling could be postponed; this would eliminate the need to stock separate SKU's for PŪR and ROWENTA products. More significantly, final assembly within each category could be delayed. Since, within each category products are primarily differentiated by filter quality, the assembly of the filter cartridge could be delayed until orders are received. Additional component assembly for sub-categories could also be postponed (e.g., stream and spray units share the same components with the exception of a spray disc). Other than the carbon filter manufacturing process, all other steps are manual, making it possible to shift both capacity and workload with relatively small process redesign effort.

Some examples of previous studies that have dealt with the problem of determining the optimal buffer size and the optimal point of differentiation, subject to a service level constraint, are Lee (1996), Lee and Tang (1997), Garg and Tang (1997), Swaminathan and Tayur (1996), and Graman and Magazine (1998). These studies are primarily inventory based with a significant emphasis on capturing the benefits of inventory pooling. Majority use order-up-to-level inventory models in which order processing time is not affected by either order size or the number of pending orders. If limited capacity is modeled, order processing times are assumed constant which eliminates any congestion delays. In contrast, we model processing time uncertainty, congestion delays, and an item-by-item replenishment of inventory - all common features of manufacturing systems. Consequently, our models provide insights at the manufacturing system level. For example, we find that the relative size of inventory carrying cost savings

that come from DD strategy are greater when an increase in product variety is accompanied by lower production volumes. Another approach to managing inventory costs while meeting demand for greater product variety is through product design that uses common components. Important contributions to literature in this area can be found in Collier (1982), Gerchak and Henig (1986), Baker et al. (1986), and Gerchak et al. (1988). Use of common components produces benefits of inventory pooling that are similar in nature to those obtained from keeping a stock of undifferentiated items. In fact, having common components may facilitate postponement.

Since our models are useful for general analysis, evaluation and design of hybrid make-to-stock/make-to-order systems, regardless of whether or not delayed product differentiation is used, they are also related to a significant body of previous literature dealing with stochastic analysis of production-inventory systems. Surveys of this literature can be found in two recent books: Buzacott and Shanthikumar (1993), and Altiok (1997). The majority of this literature focuses on the analysis of pure make-to-stock system. To our knowledge our paper is the first to propose production-inventory models of hybrid systems.

The remainder of this article is organized as follows. In section 2, we present details of our models, notation, and formulation of the basic optimization problem. Sections 3 and 4 are devoted to mathematical analyses of models representing inflexible and flexible workforce respectively. Results of numerical analysis and implications for manufacturing managers are described in sections 5. Section 6 compares three discrete options to perform either full, partial, or no postponement under increasing product variety. Concluding remarks, presented in section 7, summarize managerial insights.

2 The Two-Stage Production System

A schematic diagram of our two-stage model is shown in Figures 1 and 2 corresponding respectively to the inflexible and flexible workforce assumptions. The size of the buffer following stage-1 is denoted by b. There are M products and all demand is initially backlogged, at least for stage-2 operations. External demand for product i arrives at the intermediate buffer according to a Poisson process of rate λ_i, with $\Lambda = \sum_{i=1}^{M} \lambda_i$ denoting the total arrival rate. If the buffer is empty, a customer order is backlogged for processing at stage-1. Otherwise, it takes one item from the buffer and joins the queue of jobs waiting to be processed at stage-2. Each demand arrival at the buffer automatically triggers a release of raw materials to stage-1. We assume that raw material kits are always available.

The total work content of each job is a constant T_i units, of which t units are assigned to stage-1, and the remaining $T_i - t$ to stage-2. It is assumed that processing time variability is generated at the production stages and therefore the processing time distribution at each stage is unaffected by the amount of work assigned to that stage. In our models, we assume that we have full flexibility in

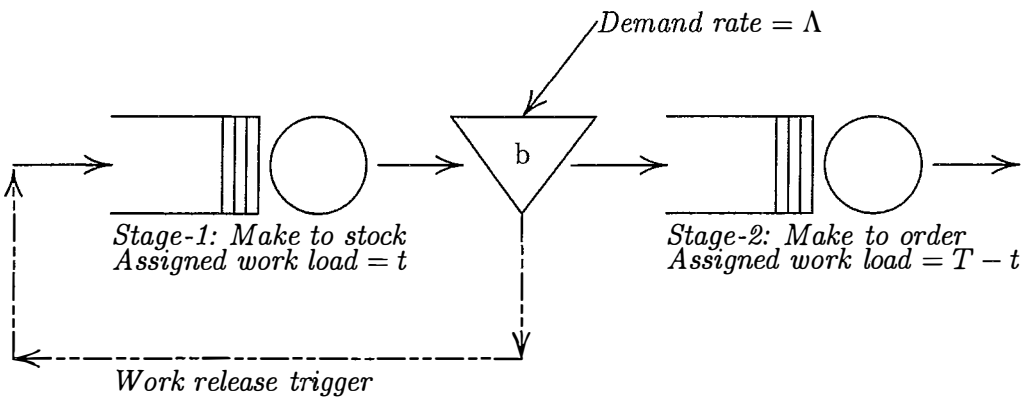

Figure 1: Schematic of a production system using delayed differentiation strategy. Items in the intermediate buffer are undifferentiated.

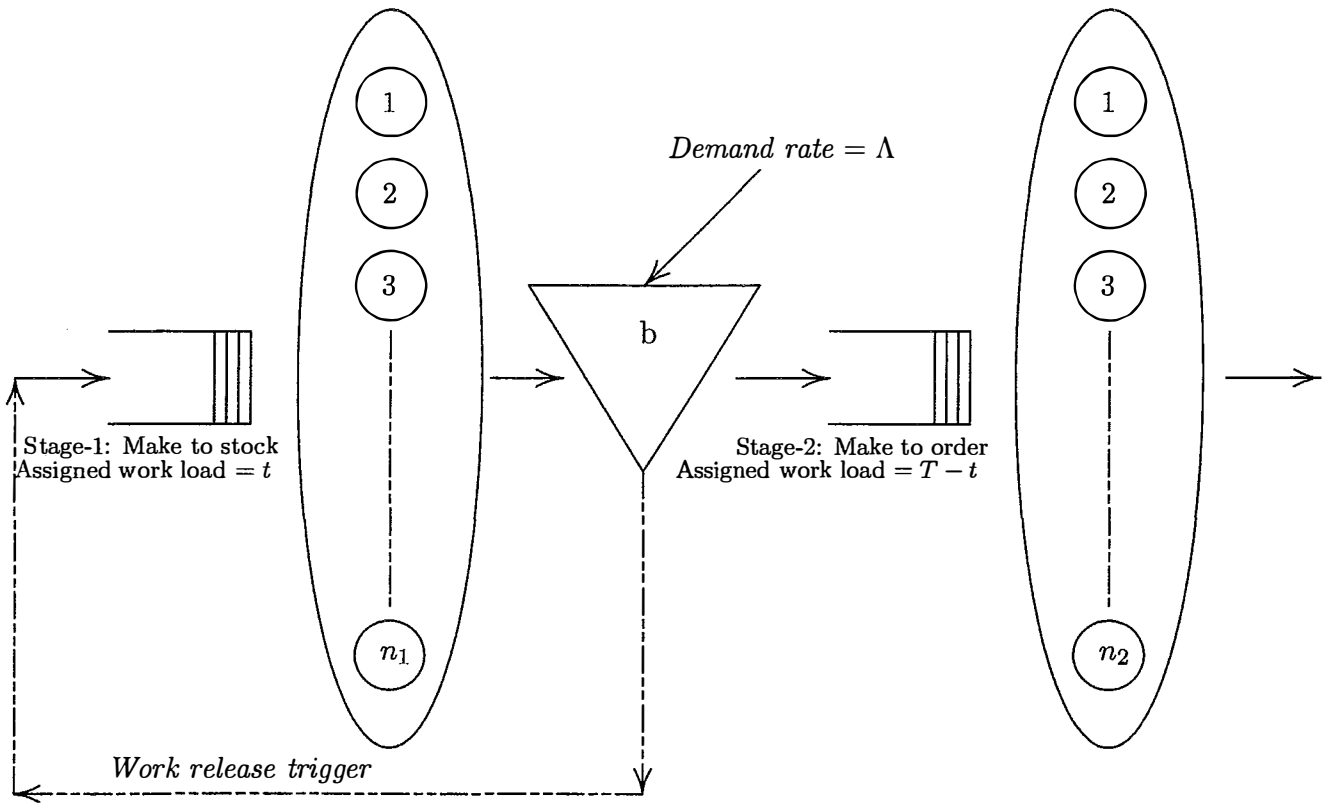

Figure 2: Schematic of a production system with flexible workforce that uses DD strategy. Items in the intermediate buffer are undifferentiated.

assigning work between the two stages. In practice, the total work content typically consists of discrete work elements. Our models remain, however, applicable when this is the case. In fact, identifying the optimal point of differentiation is simpler in those discrete case where complete enumeration is feasible.

In model 1, there are two workers and one worker is assigned to each production stage. In model 2, we have n workers of which n_1 can be assigned to stage-1. Thus, model 1 is a special instance of model 2 which is realized by setting $n = 2$ and $n_1 = 1$. However, for clarity of exposition, we present their analyses separately. Notice that the choice of point of product differentiation and b include the two extremes of entirely make-to-order and entirely make-to-stock situations. In the former, $b = 0$, and in the latter, $b > 0$ and differentiation occurs at the point of sale.

The analysis reported in sections 3 and 4 relates to the case in which $T_i = T$ for all i, that is, work content are assumed to be the same for all products. Furthermore, workers/workstations are assumed to have exponential processing time distributions. These assumptions are made for two main reasons: to focus attention on the benefits of various stocking options and on the choice of the point of differentiation; and to make analysis of the models easier.

It follows from above description and assumptions that the average processing rate in model 1 is $\mu_1 = 1/t$ at stage-1, and $\mu_2 = 1/(T - t)$ at stage-2. The corresponding quantities for model 2 are $n_1\mu_1$ and $(n - n_1)\mu_2$ respectively. For stability, it is required that $\rho_1 = \Lambda t$ and $\rho_2 = \Lambda(T - t)$ be both less than 1 for model 1; and that ρ_1 be less than n_1 and ρ_2 less than $n - n_1$ for model 2. We denote steady state order fulfillment time, number of customers backordered, and buffer inventory level, by $F(b, t)$, $S(b, t)$, and $I(b, t)$ respectively. Order fulfillment time is the total elapsed time from the moment a demand arrives to the moment that the finished product is supplied to the customer.

Let $h(t)$, $R(t)$, and $W(b)$ denote respectively the cost of holding one unit of inventory after completing t units of processing, the amortized cost per unit time of product and process redesign, and the cost per unit time of providing a warehouse of size b. As the arguments of these functions indicate h and R are functions of t, and W is a function of b. We further assume that these functions are continuous and differentiable with positive first derivatives, i.e., per unit cost of inventory and redesign are increasing in value added and cost of warehouse is increasing in its size. Note that $h(0) = R(0) = W(0) = 0$.

The optimization problem can be formulated in several different ways. Some of these are presented here to illustrate the rich class of objectives that can be handled within our framework. For example, we may wish to minimize average inventory, redesign, and warehouse costs subject to a maximum tolerable average order fulfillment time, α. Such a problem can be written as follows:

$$\underset{t,\, b \geq 0}{\text{Minimize}} \quad h(t)\bar{I} + R(t) + W(b) \tag{1}$$

Subject to:

$$\bar{F} \leq \alpha. \tag{2}$$

Alternatively, if Fx denotes the probability that order fulfillment time exceeds x days (or weeks), then constraint 2 should be replaced by the following alternative:

$$Fx \leq \alpha. \tag{3}$$

Yet another option may be to minimize the sum total of inventory and backordering costs. Let $S(b,t)$ denote the number of customers backlogged at an arbitrary observation epoch, and q denote the average cost per customer backlogged per unit time. Then, in this formulation, constraint 2 will vanish and the objective function 1 will change to

$$\underset{t,\, b \geq 0}{\text{Minimize}} \qquad h(t)\bar{I} + q\bar{S} + R(t) + W(b) \tag{4}$$

Service level constraint can also be expressed in terms of the number of customers backordered. For example, a manufacturing manager may wish to limit the number of backorders by using a constraint such as the following in place of 2.

$$\text{Prob}\{S \geq x\} \leq \alpha. \tag{5}$$

3 Inflexible Workforce Model

Although the output process of a $M/M/1$ queue is a Poisson process (Burke, 1956), the stage-2 input process of the inflexible worker model (Figure 1) is Poisson only in two special cases: $b = 0$ and $b = \infty$. 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. Using notation A_i to denote inter-arrival time at stage-i and $C^2_{A_i}$ its squared coefficient of variation, it is possible to show that $C^2_{A_2} \approx 1$ and that treating stage-2 queue as a $M/M/1$ queue is a reasonable approximation for estimating order delay at stage-2. A proof of this claim can be found in Appendix A.

Upon treating each production stage as a $M/M/1$ queue, and using standard results from queueing theory, we can obtain the following performance metrics:

$$\bar{I}(b,t) = b - \frac{\Lambda t[1-(\Lambda t)^b]}{1-\Lambda t}. \tag{6}$$

$$\bar{F}(b,t) = \frac{t(\Lambda t)^b}{1-\Lambda t} + \frac{T-t}{1-\Lambda(T-t)}. \tag{7}$$

$$
\begin{aligned}
Fx(b,t) &= \text{Prob\{ order delay} \geq x\} \\
&= \begin{cases} e^{-[1/(T-t)-\Lambda]x} + \left(\frac{(1/(T-t)-\Lambda)(\Lambda t)^b}{1/t-1/(T-t)}\right)\left(e^{-[1/(T-t)-\Lambda]x} - e^{-[1/t-\Lambda]x}\right) & \text{if } t \neq T/2, \\ \left(1 + x(\Lambda T/2)^b[2/T - \Lambda]\right) e^{-[2/T-\Lambda]x} & \text{otherwise.} \end{cases}
\end{aligned} \tag{8}
$$

$$
\bar{S}(b,t) = \frac{\rho_2}{1-\rho_2} + \frac{\rho_1^{b+1}}{1-\rho_1}. \tag{9}
$$

$$
Sx(b,t) = \text{Prob}\{S \geq x\} = \begin{cases} \rho_2^x + \left(\frac{\rho_1^{b+1}(1-\rho_2)}{\rho_2-\rho_1}\right)(\rho_2^x - \rho_1^x) & \text{if } \rho_1 \neq \rho_2, \\ \rho^x + x(1-\rho)\rho^{b+x} & \text{otherwise.} \end{cases} \tag{10}
$$

Appendix B contains arguments that can be used to arrive at the above results.

PROPOSITION 1 *The following properties of $\bar{I}$, $\bar{F}$, Fx, $\bar{S}$ and Sx hold.*

- *Average inventory $\bar{I}$ is increasing convex in b, and decreasing in t.*
- *Service level measures $\bar{F}$, Fx, $\bar{S}$, and Sx are decreasing convex in b.*
- *Average order fulfillment delay $\bar{F}$ is convex in t.*

Proof of these assertions can be found in Appendix C.

The formulation shown in 1 - 2 (section 2) can now be expanded as follows:

$$
\text{Minimize } K(b,t) = h(t)[b - \frac{\Lambda t[1-(\Lambda t)^b]}{1-\Lambda t}] + R(t) + W(b), \tag{11}
$$

Subject to:

$$
\begin{aligned}
\frac{t(\Lambda t)^b}{1-\Lambda t} + \frac{T-t}{1-\Lambda(T-t)} &\leq \alpha, && (12)\\
\Lambda t &\leq 1, && (13)\\
\Lambda(T-t) &\leq 1, && (14)\\
t &\leq T, && (15)\\
t &\geq 0, && (16)\\
b &\geq 0. && (17)
\end{aligned}
$$

If service criterion is the proportion of orders that exceed some quoted lead time, or the degree of customer backlog considered acceptable, then the LHS of 12 is replaced by 8, or 10, respectively. The optimization problem, 11 - 17, is a non-convex minimization problem. Such problems generally lack elegant solutions. However, the following result can be used to devise an effective solution algorithm.

PROPOSITION 2 *For the optimization problem described in 11 - 17, either a pure make-to-order configuration is optimal, or the service level constraint is binding.*

Proof: Notice that the pure make-to-order configuration could either have some work performed at both stages or all work performed at stage-2. Consider the Kunh-Tucker first order necessary conditions for optimality (see, for example Luenberger (1984), pp. 314-315). According to these, there exist $\gamma_i \geq 0$ such that

$$\gamma_1(\bar{F} - \alpha) = 0, \tag{18}$$

$$\gamma_2(\Lambda t - 1) = 0, \tag{19}$$

$$\gamma_3(\Lambda(T - t) - 1) = 0, \tag{20}$$

$$\gamma_4(t - T) = 0, \tag{21}$$

$$\gamma_5(-t) = 0, \tag{22}$$

$$\gamma_6(-b) = 0, \tag{23}$$

$$h_t\bar{I} + h\bar{I}_t + R_t + \gamma_1\bar{F}_t + \Lambda\gamma_2 - \Lambda\gamma_3 + \gamma_4 - \gamma_5 = 0, \text{ and,} \tag{24}$$

$$h\bar{I}_b + W_b + \gamma_1\bar{F}_b - \gamma_6 = 0. \tag{25}$$

In above equations, the service level measure $\bar{F}$ should be replaced by Fx or Sx, as appropriate, for alternate formulations. Alphabetical subscripts denote partial derivatives with respect to that variable. For example, $\bar{I}_b$ denotes the partial derivative of $\bar{I}$ with respect to b.

If $\Lambda T < 1$, and the service level constraint can be met by performing all work in a make-to-order fashion (i.e., in stage-2), then clearly $b^* = 0$ since $K(0, 0) = 0$ is the smallest possible value of K. For example, when service level is measured by $\bar{F}$, $b^* = 0$ is optimal so long as $\Lambda T < 1$ and $\alpha \geq T/(1 - \Lambda T)$. In all other instances $t = 0$ is not feasible. Hence, from complementary slackness, $\gamma_5 = 0$. We also notice that the service measures approach their limiting worst case values ($\bar{F} \rightarrow \infty$, $Fx \rightarrow 1$, and $Sx \rightarrow 1$), when either Λt or $\Lambda(T - t) \rightarrow 1$. Clearly, then constraints 2 and 3 cannot be tight for any meaningful service level constraint. From 19 and 20, this means $\gamma_2 = \gamma_3 = 0$.

Conditions 24 and 25 can now be reorganized as follows:

$$\gamma_1 = -\left(\frac{h_t\bar{I} + h\bar{I}_t + R_t + \gamma_4}{\bar{F}_t}\right), \tag{26}$$

$$\gamma_6 = h\bar{I}_b + W_b + \gamma_1\bar{F}_b \tag{27}$$

There are now only two possibilities. If $b > 0$, then $\gamma_6 = 0$ which means $\gamma_1 = -[h\bar{I}_b + W_b]/\bar{F}_b > 0$ and constraint 12 is tight. Otherwise, i.e., when $b = 0$, then the flow time constraint may or may not be tight since γ_1 is no longer strictly positive. Clearly, one of these two possibilities is the optimal solution. Similar arguments also hold when service is measured as Fx or Sx owing to fact that $(Fx)_b$ and $(Sx)_b$ are negative (Proposition 1).

Putting it all together, we see that either $b^* = 0$, i.e., a make-to-order configuration is optimal or if $b^* > 0$, then the service level constraint is binding. Hence proved. #

As a result of Proposition 2, we can write the original optimization problem as a function of a single variable (b washes out since it is either set equal to 0 or it can be expressed in terms of t from the service level constraint). When the service level constraint is active, the optimization problem can be written as follows:

$$\begin{aligned} \text{Minimize } K(t) &= h(t)\left[\ln(Y(t)(1-\Lambda t))/\ln(\Lambda t) - \frac{\Lambda t}{1-\Lambda t} + \Lambda t Y(t)\right] + R(t) \\ &+ W(\ln(Y(t)(1-\Lambda t))/\ln(\Lambda t)), \end{aligned} \tag{28}$$

where $Y(t) = [\alpha(1-\Lambda(T-t)) - (T-t)]/[(1-\Lambda(T-t))t]$. Optimal t can be obtained by searching in the range $\max\{0, T-1/\Lambda\} \leq t \leq \min\{T, 1/\Lambda\}$. Similar expressions also exist for constraints involving Fx and Sx as measures of service. Thus, we can reduce 11 -17 into a minimization problem in one variable over some finite feasible range. Appendix D describes an algorithm that we used to find optimal b and t in numerical examples reported in section 5.

4 Flexible Workforce Model

In this model we shall assume that $n_1 \leq b$. Since the cost of providing an additional storage space is typically much smaller than the cost of providing an additional worker, we expect this condition to be satisfied in practically all situations. Moreover, if we do not provide at least one space per worker, it will be possible to have jobs waiting for the completion of stage-1 operations while simultaneously having idle workers at that stage. Such a situation will be deemed undesirable by most manufacturing managers.

Just as stage-2 input process in the two worker model is not a Poisson process, here too stage-2 receives non-Poisson input. However, assuming it is Poisson leads to a reasonable approximation. The arguments necessary to justify this claim are included in Appendix A. Based on this approximation, we can write an expression in closed-form for $\pi_i(k_i)$, the probability that there are k_i jobs waiting at stage-i (including the ones being processed) in steady state, for both $i = 1$ and $i = 2$. For convenience, we use $n_2 = n - n_1$ and let $\alpha_i(k_i) = \min\{k_i, n_i\}$. Then, the average stage-1 processing rate is $\alpha_1(k_1)/t$ and stage-2 rate is $\alpha_2(k_2)/(T-t)$. Clearly, then

$$\pi_i(k_i) = \frac{\pi_i(0)\rho_i^{k_i}}{\alpha_i(k_i)!\; n_i^{k_i - \alpha_i(k_i)}} \qquad \forall\; k_i = 0, 1, \cdots, \tag{29}$$

and

$$\pi_i(0) = \left(\sum_{j=0}^{n_i - 1} \frac{\rho_i^j}{j!} + \frac{\rho_i^{n_i}}{n_i!\,(1-\rho_i/n_i)} \right)^{-1}. \tag{30}$$

Next, we compute the average buffer inventory and the average order fulfillment time for model 2.

By definition, the average buffer inventory

$$\bar{I}(b, n_1, t) = \sum_{r=0}^{b}(b-r)\pi_1(r) = b[1 - \sum_{r=b+1}^{\infty}\pi_1(r)] - [\bar{Q}_1 - \sum_{r=b+1}^{\infty} r\pi_1(r)], \tag{31}$$

where $\bar{Q}_1$ is the average number in stage-1 subsystem and has the following standard expression (see, for example, Buzacott and Shanthikumar, 1993, pp. 78-79):

$$\bar{Q}_1 = \frac{\pi_1(0)\rho_1^{n_1}(\rho_1/n_1)}{n_1!(1-\rho_1/n_1)^2} + \rho_1. \tag{32}$$

Upon simplifying each term in 31, we obtain the following

$$\bar{I}(b, n_1, t) = b - \rho_1 - \frac{B_1(\rho_1/n_1)}{(1-\rho_1/n_1)}\{1 - (\rho_1/n_1)^{b-n_1}\}. \tag{33}$$

In the above relationship, $B_1 = \frac{\pi_1(0)\rho_1^{n_1}}{n_1!(1-\rho_1/n_1)}$ is the probability that all n_1 stage-1 workers are busy.

The order fulfillment time consists of two components: F_1 which is non-zero only when stage-1 queue has at least b pending requests, and F_2, the stage-2 order delay which is experienced by all demand arrivals. Thus,

$$\bar{F}(b, n_1, t) = \sum_{r=b}^{\infty}\frac{(r-b+1)\pi_1(0)\rho_1^r(t/n_1)}{n_1! n_1^{r-n_1}} + \frac{(T-t)\rho_2^{n_2}\pi_2(0)}{(n_2-\rho_2)n_2!(1-\rho_2/n_2)} + (T-t). \tag{34}$$

Upon simplifying the RHS of the above relationship, we obtain

$$\bar{F}(b, n_1, t) = \frac{B_1 t(\rho_1/n_1)^{b-n_1}}{n_1 - \Lambda t} + \frac{B_2(T-t)}{(n-n_1) - \Lambda(T-t)} + (T-t). \tag{35}$$

Like the definition of B_1 in 33, $B_2 = \frac{\pi_2(0)\rho_2^{n_2}}{n_2!(1-\rho_2/n_2)}$ is the probability that all n_2 stage-2 workers are busy. Notice that B_i's are functions of n_i and workload allocation t, although this dependence is not explicitly shown in the notation.

The formulation of the optimization problem in 1 - 2 (section 2) can now be expanded as follows:

$$\text{Minimize } K(b, n_1, t) = h(t)[b - \rho_1 - \frac{B_1(\rho_1/n_1)}{(1-\rho_1/n_1)}\{1 - (\rho_1/n_1)^{b-n_1}\}] + R(t) + W(b), \tag{36}$$

Subject to:

$$\frac{B_1 t(\rho_1/n_1)^{b-n_1}}{n_1 - \Lambda t} + \frac{B_2(T-t)}{(n-n_1) - \Lambda(T-t)} + (T-t) \leq \alpha, \tag{37}$$

$$\Lambda t \leq n_1, \tag{38}$$

$$\Lambda(T-t) \leq n - n_1, \tag{39}$$

$$t \leq T, \tag{40}$$

$$n_1 \leq n, \tag{41}$$

$$n_1 \leq b, \tag{42}$$

$$t \geq 0, \tag{43}$$

$$b \geq 0, \tag{44}$$

$$n_1 \geq 0. \tag{45}$$

Analogous to the treatment in the previous section, here too the service level constraint can be expressed either in terms of the proportion of orders (or backorders) that exceed the quoted lead time (or backlog threshold). In the interest of brevity, we have omitted these details.

The formulation contained in 36 - 45 is a non-convex optimization problem. Solving such problems can be difficult. We simplify this problem, like before, by proving that for any fixed n_1, either the optimal b equals n_1 or the service level constraint 37 is binding. In order to prove this result, we need to show that the following properties of $\bar{I}$ and $\bar{F}$ hold.

PROPOSITION 3 *Average inventory, $\bar{I}$, and average order delay, $\bar{F}$, possess the following properties:*

- *$\bar{I}$ is increasing convex in b, decreasing in t and increasing in n_1.*
- *$\bar{F}$ is decreasing convex in b and convex in t.*

A proof of these assertions can be found in Appendix E.

For solving 36 - 45, we suggest a procedure that evaluates optimal b, t and $\bar{I}$ for each integer value of n_1 in $[0, n]$. Overall optimal n_1 and corresponding b and t are those parameter values that yield the smallest $K(b, n_1, t)$. Thus, we need to solve at most $(n+1)$ problems of the type we solved for the inflexible workforce model. A complete description of the algorithm that was used to solve the numerical examples reported in the next section can be found in Appendix F.

The two limiting cases $n_1 = 0$ and $n_1 = n$ represent the two extremes of pure make-to-stock and pure make-to-order systems using DD strategy respectively. These are treated in a special way, as explained in Appendix F. In all other instances, i.e., when $n_1 = m$, where $1 \leq m \leq n-1$, we use the following fact.

PROPOSITION 4 *For the optimization problem in 36 - 45, if $n_1 = m$ is selected, where $1 \leq m \leq n-1$, then either $b^*(m) = m$, or the service level constraint is binding.*

Proof: Consider the Kuhn-Tucker first order necessary conditions for 36 - 45 when $n_1 = m$ has been selected (see, for example Luenberger (1984), pp. 314-315). There exist $\gamma_i \geq 0$ such that

$$\gamma_1(\bar{F} - \alpha) = 0, \tag{46}$$

$$\gamma_2(\Lambda t - n_1) = 0, \tag{47}$$

$$\gamma_3(\Lambda(T-t)-n+n_1) = 0, \quad (48)$$

$$\gamma_4(t-T) = 0, \quad (49)$$

$$\gamma_5(n_1-n) = 0, \quad (50)$$

$$\gamma_6(n_1-b) = 0, \quad (51)$$

$$\gamma_7(-t) = 0, \quad (52)$$

$$\gamma_8(-b) = 0, \quad (53)$$

$$\gamma_9(-n_1) = 0, \quad (54)$$

$$h_t\bar{I} + h\bar{I}_t + R_t + \gamma_1\bar{F}_t + \Lambda\gamma_2 - \Lambda\gamma_3 + \gamma_4 - \gamma_7 = 0, \text{ and,} \quad (55)$$

$$h\bar{I}_b + W_b + \gamma_1\bar{F}_b - \gamma_6 - \gamma_8 = 0. \quad (56)$$

If $\bar{F}(m,m,0) \leq \alpha$, then the service level requirement can be met by setting $t = 0$, i.e., producing all items to order and idling m workers at stage-1. Clearly, in this case $b(m) = m$ is a candidate solution.

If, on the other hand, $\bar{F}(m,m,0) > \alpha$, then $t = 0$ is not feasible and from complementary slackness, $\gamma_7 = 0$. Notice also that $n_1 = m$, where $1 \leq m \leq n-1$. Therefore, $\gamma_5 = \gamma_9 = 0$. Similarly, since $\alpha < \infty$ and $\bar{F}$ approaches ∞ as either $\Lambda t \to n_1$ or $\Lambda(T-t) \to (n-n_1)$, constraints 2 and 3 cannot be tight. Therefore, $\gamma_2 = \gamma_3 = 0$. Finally, for $n_1 \in [1, n-1]$, $b \geq n_1$ implies $b > 0$, hence $\gamma_8 = 0$.

Constraint 56 can now be written as

$$\gamma_6 = h\bar{I}_b + \gamma_1\bar{F}_b + W_b \quad (57)$$

It is easy to confirm from 57 and complementary slackness conditions that if $b(m) > m$, γ_6 must equal zero, and $\gamma_1 = -[h\bar{I}_b + W_b]/\bar{F}_b$ is strictly positive. Therefore, constraint 37 must be tight. On the other hand, if constraint 37 is not tight, then $\gamma_1 = 0$ which implies $\gamma_6 = h\bar{I}_b + W_b > 0$ and therefore $b(m) = m$. Hence proved. #

5 Model Analysis and Insights

We begin by examining the effect of delayed differentiation on the various performance metrics discussed in section 2. We first consider order fulfillment delay in the model with fixed worker assignments. Here, the average order delay is convex in t for a fixed b, but it is not monotonic in t. We therefore observe that increased delay in differentiation may or may not improve order fulfillment time (see Figure 3). In fact, the point of differentiation $t(b)$ that minimizes order delay for a fixed level of buffer size of b can be shown to be always at least 50% of the total work content, that is, $t(b) \geq T/2$ (see Appendix G for proof). That is, it is always optimal to let the undifferentiated items be at least half finished, regardless of actual workload or buffer level. For hybrid make-to-stock, make-to-order systems, this means that it is always desirable to do at least half of the work in a make-to-stock fashion.

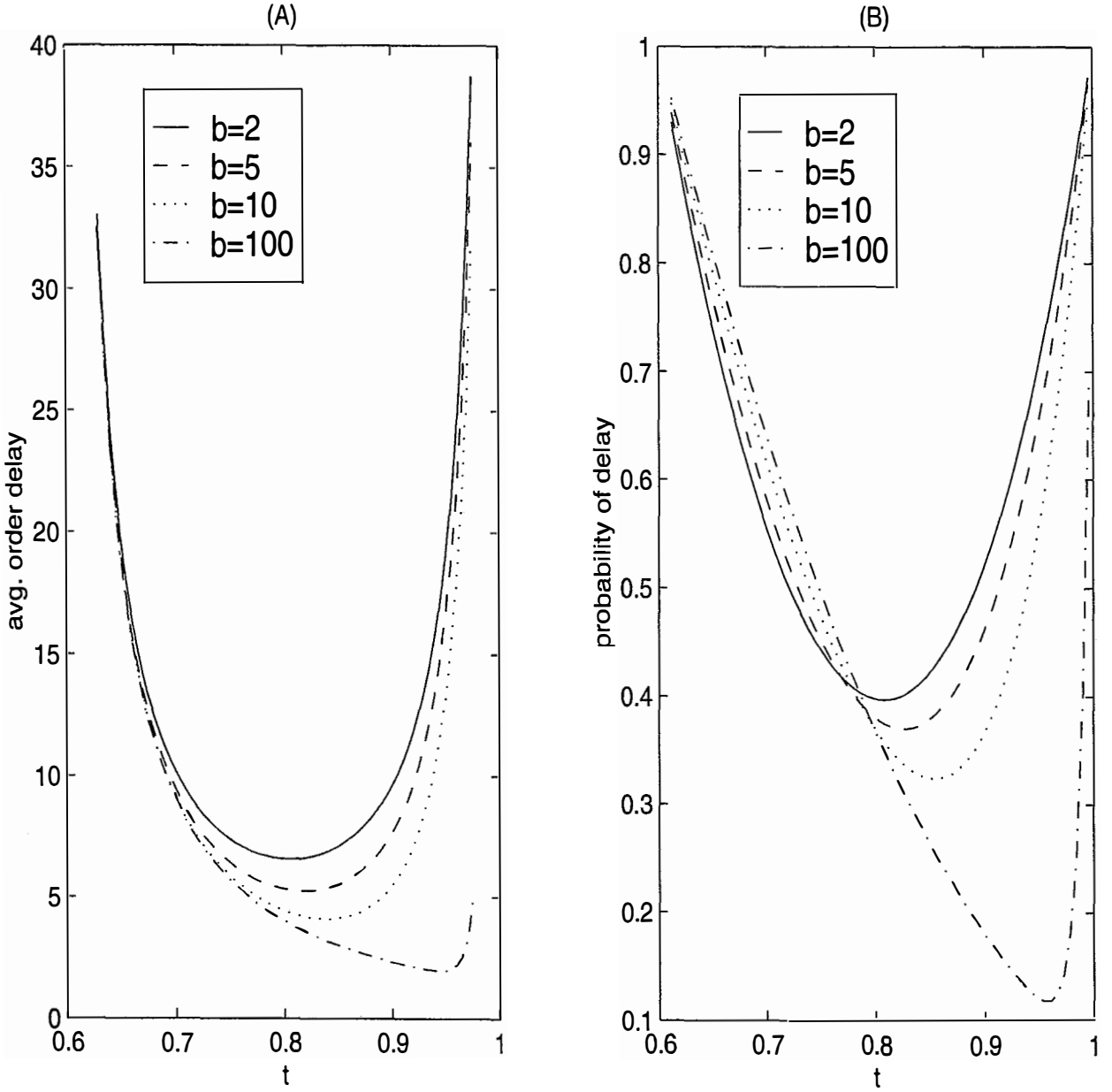

Figure 3: Plot of average order fulfillment time, $\bar{F}$, and probability of delay exceeding a quoted lead time x (denoted by Fx) as a function of the point of differentiation for the inflexible workforce model. Data are as follows: $\Lambda = 1$, $T = 1.6$, $t \in (0.6, 1)$, and $x = \bar{F}(b, T/2)$ for each b.

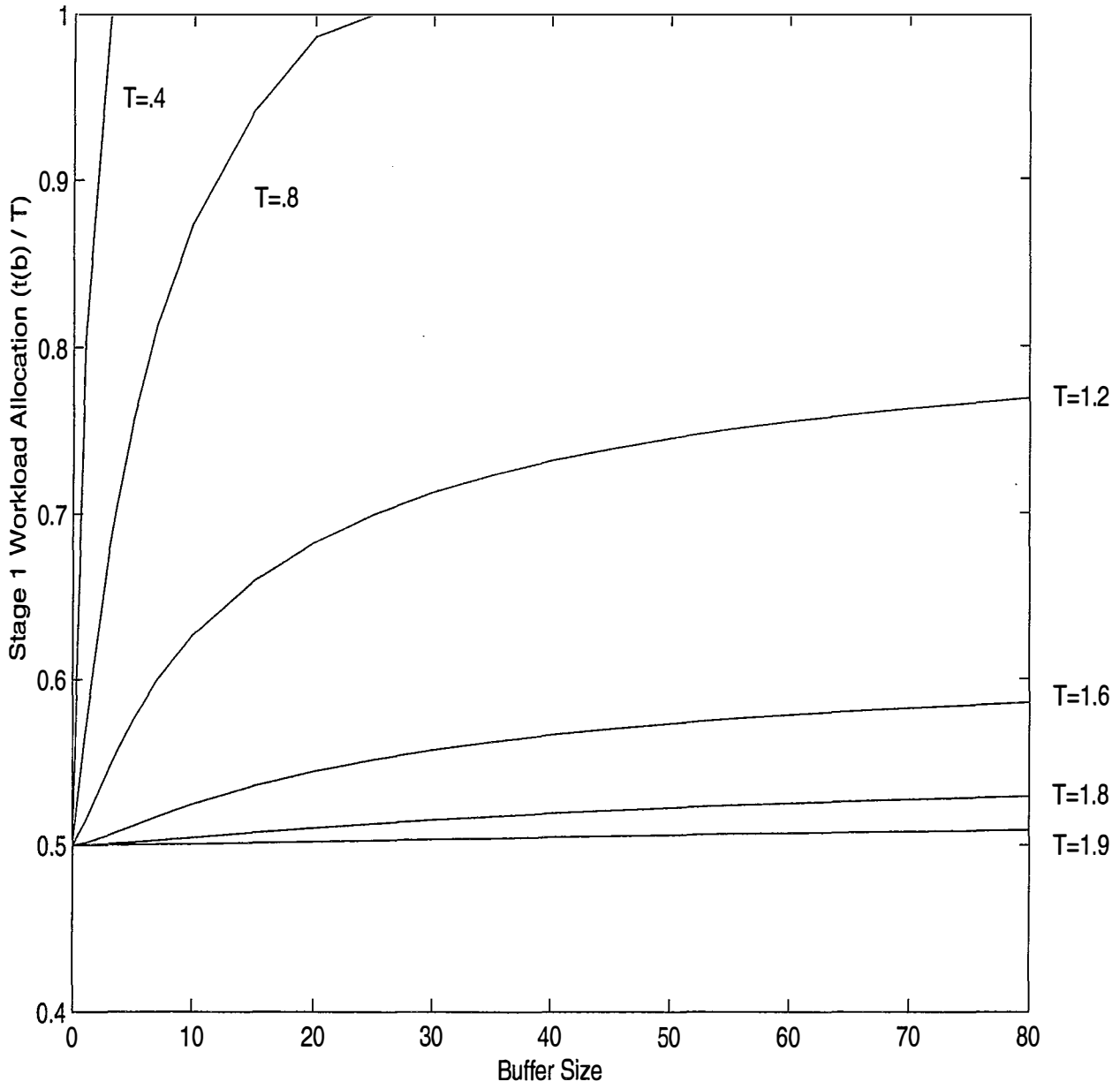

Figure 4: Impact of buffer size on the proportion of stage-1 work that minimizes average order fulfillment delay.

In Figure 4, it is observed that an increase in the available buffer size favors greater delayed differentiation, which suggests that, in order to enable greater differentiation without sacrificing customer response time, investments in larger buffers should be made. This goes somewhat counter to the prevalent arguments that delayed differentiation would always result in smaller inventories, arguments that typically ignore the impact of postponing differentiation on order delay. Also note that, with increased overall workload, delayed differentiation becomes less desirable. This means that delayed differentiation is a much more effective strategy for products with a small work content and/or for systems with excess capacity. This is also counter-intuitive since one would assume that, in a lightly loaded system, a make-to-order strategy would be optimal.

Similar observations can also be extended to the model with flexible worker assignment. In this case, the optimal point of differentiation depends on the number of workers allocated to each stage. We show (see Appendix G) that if $n_1 \geq n/2$, then the optimal point of differentiation always satisfies the following condition:

$$t^*/T \geq n_1/n \tag{58}$$

Put differently, the fraction of total work that is undifferentiated is at least as much as the fraction of capacity assigned to stage 1. If a balanced capacity allocation is carried out, then it is still optimal to keep at least half of the total work content as undifferentiated.

Next, we consider the impact of delayed differentiation on inventory level and inventory cost. As mentioned in proposition 1, for a fixed buffer size, average inventory is decreasing in t. This means that greater differentiation would always reduce inventory level (by increasing inventory replenishment lead times). On the other hand, delaying differentiation increases the value added of the held inventory. Therefore, delaying differentiation may or may not reduce inventory costs. This effect is graphically depicted in Figure 5. Note that inventory cost is small at both ends of the differentiation spectrum; on one end, due to the fact that products are made to order and, on the other, due to the long replenishment lead times.

In order to gain insights into the optimal solution of the formulations in 11-17 and 36-45 and to evaluate the impact of different holding, warehousing, and redesign costs, we solved a series of problems with different cost functions and different service level requirements. Samples of these results are reported in Table 1 for model 1 and Tables 2-3 for model 2. In Table 1, all three values of the service level α are chosen in the range $\alpha_1 < \alpha < \alpha_2$, where $\alpha_1 = \max\{0, \frac{T-1/\Lambda}{2-\Lambda T}\}$ and $\alpha_2 = \frac{2T}{2-\Lambda T}$. It is shown in Appendix D that in this range $b^* > 0$. The low, medium and high values of α are defined relative to the range $(\alpha_2 - \alpha_1)$ such that

$$\begin{aligned} \alpha_\ell &= \alpha_1 + 0.05(\alpha_2 - \alpha_1), \\ \alpha_m &= (\alpha_1 + \alpha_2)/2, \quad \text{and} \end{aligned}$$

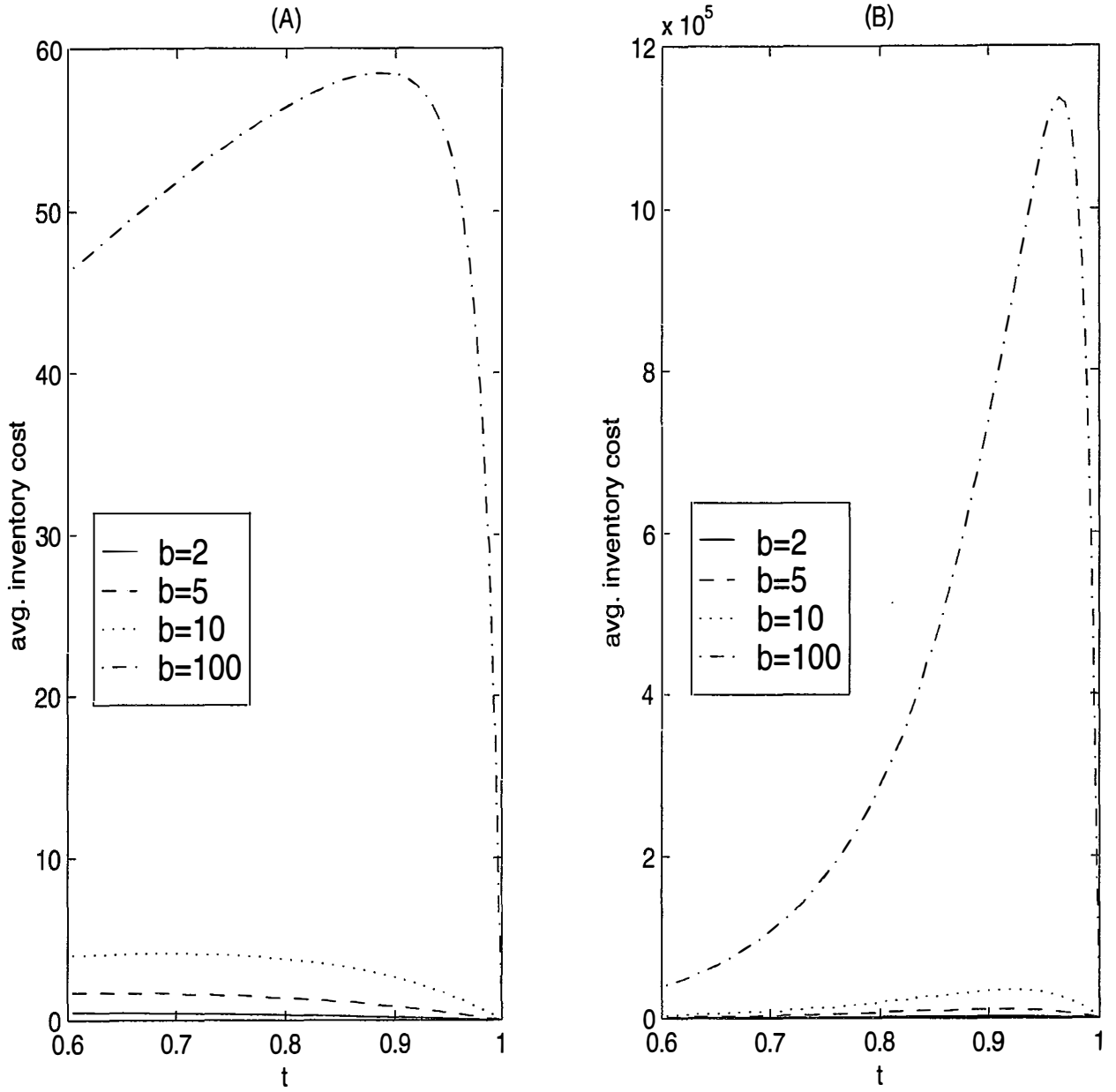

Figure 5: Plot of average inventory cost as a function of point of differentiation for the inflexible workforce model. Data are as follows: $\Lambda = 1$, $T = 1.6$, and $t \in (0.6, 1)$; (A) $h(t) = \ln(1 + t)$, and (B) $h(t) = e^{10t} - 1$.

$$\alpha_h = \alpha_1 + 0.95(\alpha_2 - \alpha_1). \quad (59)$$

Notice that these values do not remain constant when T changes.

Table 1 reveals that later differentiation is favored when α and T are small. This corresponds to a situation where the order fulfillment time constraint is severe, but we have excess capacity that allows us to complete most of the work in the make-to-stock portion of the supply chain. In contrast, when T is large, the optimal workload allocation is approximately balanced in all cases. Also, when α is large and T is small, there is no incentive to maintain an inventory of semi-processed parts (since order delay is not a concern) and stage-1 receives less than 50% of the total work content, i.e., early product differentiation becomes desirable. In these cases, b^* is also small. Note that the effect of increasing work content T on the optimal point of differentiation t is not always predictable. When order delays must be kept small (small α), an increase in T results in earlier differentiation. On the other hand, when order delays can be long (large α), an increase in T results in further delay in differentiation.

More significantly, while the optimal cost $K(b^*, t^*)$ is highly sensitive to our choice of the holding, warehousing, and redesign cost functions, the optimal values of b are remarkably insensitive. This is due to the fact that the size of the inventory buffer is solely determined by constraint 12 which is always tight when $b^* > 0$. In this case, the value of b^* is set equal to its smallest value that allows us to meet the order delay constraint. Equally significant, the ratio t^*/T remains quite stable for different cost functions, which means that the optimal differentiation point is insensitive to the choice of $h(t)$, $R(t)$, and $W(b)$. These observations suggest that managers need not be too concerned about estimating these cost functions accurately as long as $h(t)$ and $R(t)$ are increasing in t and that $W(b)$ is increasing in b.

Some results of our experiments with model 2 are summarized in Tables 2-3. We report two data sets. In one case (Table 2), the total number of workers is 5, and in the other case (Tables 3), it is 10. For both data sets we consider three different cost functions for $h(t)$ which are the same as the ones we chose for model 1. The T values are chosen so as to create an average load per worker of 0.35, 0.6 and 0.9 (consistent with model 1). Parameter α is set equal to 0.5 for the 5 worker problem and 1.0 for the 10 worker problem. The optimal (b, n_1, t) triplets have been shown in boldface. We solved other examples with different values of α and cost parameters which are not reported here. Results shown are representative of what we obtained for other data as well.

As in model 1, we observe that whenever it is possible to satisfy order fulfillment delay requirement in a make-to-order fashion, then that arrangement is also optimal. In a few instances this happens even when some of the available workers are idled. In such cases only the arrangement that results in minimum mean order fulfillment delay has been shown in bold font in Tables 2 - 3. For small α and large T, the only feasible configuration is the make-to-stock arrangement, as seen in the last column of

Tables 2 and 3. When both α and T are small, it is generally desirable to have a hybrid system with differentiation occurring much later in the production process (i.e., the value of t^*/T is close to 1). For large α it becomes increasingly more desirable to differentiate early. These results are different in some important ways from those obtained in model 1. For example, when work content is high, delaying differentiation as late as possible is desirable in model 2. This contrasts with model 1, where delayed differentiation is desirable when work content is low, while a balanced work allocation is optimal when work content is high. This difference is due to the fact that, in model 2, we have full flexibility in allocating capacity between the two stages with the possibility of assigning no capacity to one of these stages.

More significantly, the interaction between capacity allocation and buffer size decisions lead to a number of surprising results. For example, we can see from Tables 2 and 3 (column 1) that assigning more capacity to a stage does not necessarily result in more work being assigned to that stage. This also means that the optimal point of differentiation, as measured by t^*/T, is not monotonic in n_1. Thus, increasing capacity at a stage may not lengthen the optimal differentiation delay. Equally, surprising is the fact that it can be optimal to idle workers i.e., assign all the work to a stage but assign to it only a fraction of total capacity. For example, in Table 2, column 1 (second set of cost functions), it is optimal to assign all the work to stage 1 but to only assign 3 of the 5 workers to that stage. As in model 1, for a fixed n_1, we can also see from Table 3 that workload distribution is not monotonic in work content. However, the optimal workload distribution is indeed increasing in T in this case when n_1 is also adjusted accordingly. This also affects the optimal cost which, as we can see, from Table 3 can be reduced by increasing work content and adjusting n_1 accordingly. These somewhat counter-intuitive results highlight the strong linkages between capacity, workload and buffer size decisions. They also highlight the usefulness of models like ours in capturing these interactions. Because these interactions are not adequately captured by pure inventory models, such models may not be able to effectively capture the true cost and benefits of delayed differentiation.

6 Comparison of Full, Partial and No Postponement Strategies

In order to further examine the benefits of delayed differentiation, we provide comparisons with two benchmark alternatives: (1) systems with no postponement and differentiated finished goods inventories, and (2) systems with full postponement and undifferentiated finished goods inventories. Graphical depictions of these two alternatives, hereafter called system 1 and system 2, are shown in Figures 6 and 7 respectively. Notice that the system shown in Figure 7 is related to our model 1 with the difference that here both stages participate in production even though differentiation is fully postponed. In this class of systems, differentiation occurs immediately before shipping to a customer and takes

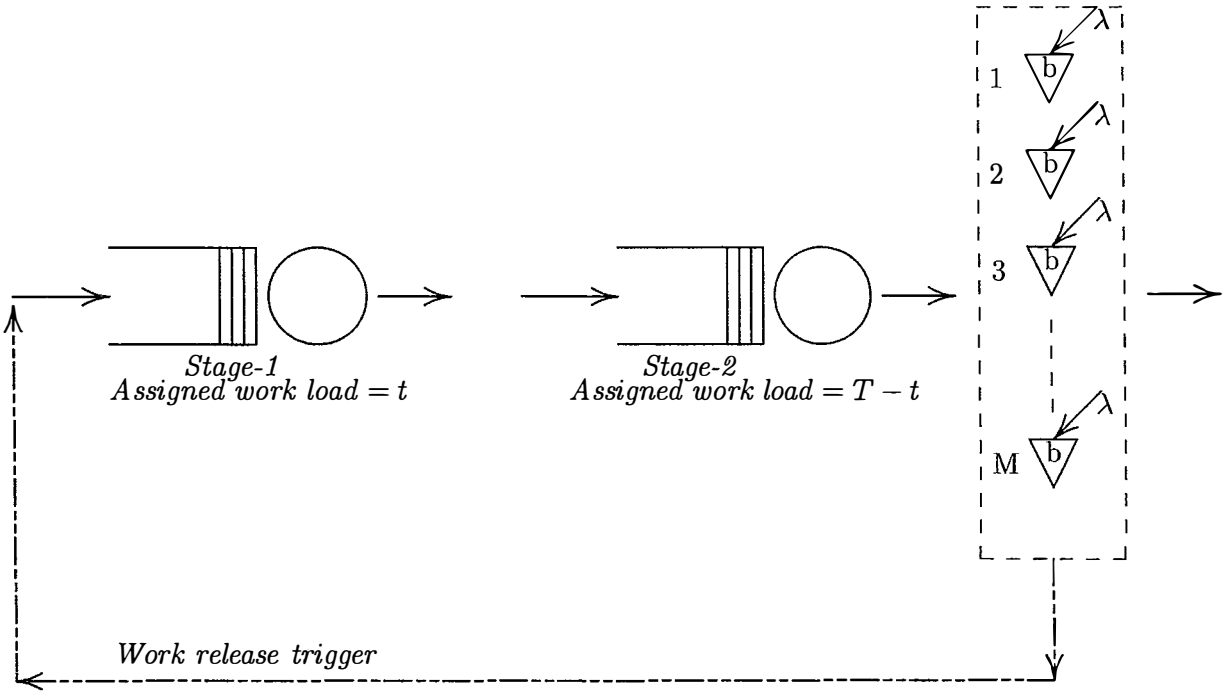

Figure 6: Schematic of a production system with differentiated finished goods inventory only.

very little time relative to the overall manufacturing work content of the product. Such a strategy is increasingly popular in the computer industry where differentiation often takes place at the point of sale or shortly prior to shipping (see, for example, Feitzinger and Lee, 1997). The first class of systems is representative of conventional strategies in make-to-stock environments.

Key performance measures for both systems are derived in Appendix H. In order to streamline comparison, here we make the assumption that $\lambda_i = \lambda$ for all i and therefore $\Lambda = M\lambda$. With this simplification, it is justifiable to have finished goods inventory buffers of equal size for all M products for the model shown in Figure 6. We shall use subscripts 1 and 2 to denote systems 1 and 2 respectively. Then, we can show that:

$$\bar{I}_1(Mb,t) = \begin{cases} (1-\rho)^2 M^3 \sum_{k=0}^{b} \frac{(b-k)(k+1)\rho^k}{(M-(M-1)\rho)^{k+2}} & \text{if } \rho_1 = \rho_2 = \rho, \\ \frac{(1-\rho_1)(1-\rho_2)M^2}{(\rho_2-\rho_1)} \sum_{k=0}^{b}(b-k)\left[\frac{\rho_2^{k+1}}{(M-(M-1)\rho_2)^{k+1}} - \frac{\rho_1^{k+1}}{(M-(M-1)\rho_1)^{k+1}}\right] & \text{otherwise.} \end{cases} \tag{60}$$

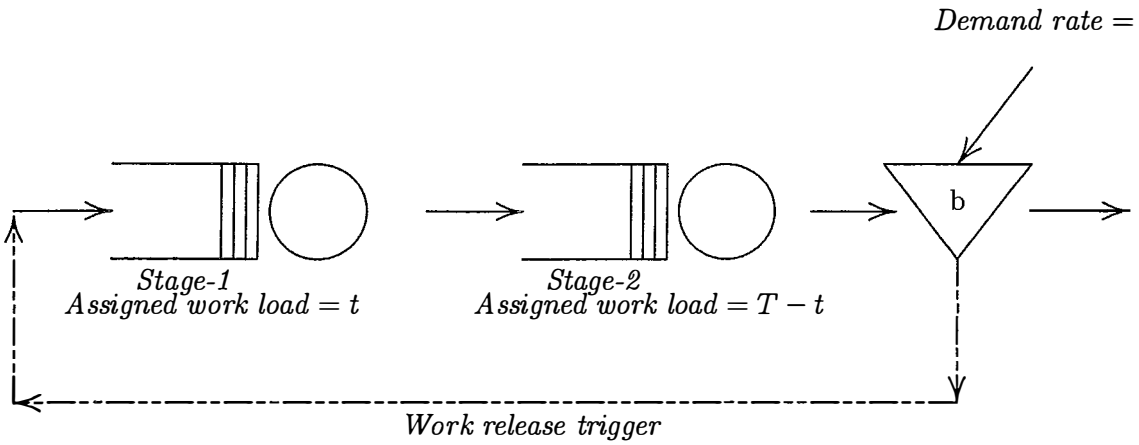

Figure 7: Schematic of a production system with full postponement and undifferentiated finished goods inventories.

$$\bar{F}_1(Mb,t) = \begin{cases} \sum_{r=b}^{\infty}\sum_{k=b}^{r}\sum_{y=0}^{r-k}\frac{(k-b+y)!(r+b-k-y-1)!}{y!(k-b)!(b-1)!(r-k-y)!}\left(\frac{1}{M}\right)^k\left(\frac{M-1}{M}\right)^{r-k} \\ \cdot(T/2)\left[\sum_{\ell=0}^{k-b+y}\frac{(r-\ell+2)}{(r-\ell+1)}\right](r+1)(1-\rho)^2\rho^r & \text{if } \rho_1=\rho_2=\rho, \\ \sum_{r=b}^{\infty}\sum_{k=b}^{r}\sum_{y=0}^{r-k}\frac{(k-b+y)!(r+b-k-y-1)!}{y!(k-b)!(b-1)!(r-k-y)!}\left(\frac{1}{M}\right)^k\left(\frac{M-1}{M}\right)^{r-k} & \\ \cdot\left[(k-b+y+1)(T-t)+t[\sum_{\ell=0}^{k-b+y}\frac{(\rho_1/\rho_2)^{r-\ell}-(\rho_1/\rho_2)^{r-\ell+1}}{1-(\rho_1/\rho_2)^{r-\ell+1}}]\right] & \\ \cdot\left(\frac{(1-\rho_1)(1-\rho_2)(1-(\rho_1/\rho_2)^{r+1})\rho_2^r}{(1-\rho_1/\rho_2)}\right) & \text{otherwise.} \end{cases} \tag{61}$$

$$\bar{I}_2(b,t) = \begin{cases} \left(\frac{(1-\rho_1)(1-\rho_2)}{\rho_2-\rho_1}\right)\sum_{i=1}^{b} i[\rho_2^{b-i+1}-\rho_1^{b-i+1}] & \text{if } \rho_1\neq\rho_2, \\ (1-\rho)^2\sum_{i=1}^{b} i(i+1)\rho^i & \text{when } \rho_1=\rho_2=\rho. \end{cases} \tag{62}$$

$$\bar{F}_2(b,t) = \begin{cases} \left(\frac{(1-\rho_1)(1-\rho_2)}{(1-\rho_1/\rho_2)}\right)\sum_{r=b}^{\infty}\left[(r-b+1)(T-t)+\sum_{k=0}^{r-b}\frac{(\rho_1/\rho_2)^{r-k}[1-\rho_1/\rho_2]t}{1-(\rho_1/\rho_2)^{r-k+1}}\right] & \\ \cdot[\rho_2^r(1-(\rho_1/\rho_2)^{r+1})] & \text{if } \rho_1\neq\rho_2, \\ (1-\rho)^2\sum_{r=b}^{\infty}\left[(r-b+1)(T-t)+\sum_{k=0}^{r-b}\frac{t}{r-k+1}\right](r+1)\rho^r & \text{otherwise.} \end{cases} \tag{63}$$

Relationships shown in 60 - 63 are difficult to evaluate numerically. Therefore, only some specific configurations are compared to the DD configuration of Figure 1. The latter is called system 3 in this section. We report in Table 4 and 5 the buffer size and average inventory needed to meet the same average order fulfillment delay target in all three systems. In these comparisons, we vary the number of products M, and keep the system load ($M\lambda T$) constant by changing either T (in Table 4) or λ (in Table 5). Table 4 corresponds to systems where increased variety is achieved by greater processing efficiency or by selection of products with lower work content. Table 5 corresponds to systems where variety is achieved by producing a larger number of products in lower volumes. In systems 1 and 2, equal work is assigned to each of the two production stages, i.e., $t = T/2$. However, in system 3, we can calculate the optimal workload allocation, which is denoted by t_3^*. Table 4 shows average inventory and inventory cost for each of the three systems under three different carrying cost functions. It also

shows the relative benefits of system 2 over system 1 and also of system 3 over system 2, denoted by Δ_1 and Δ_2 respectively.

In both tables, we notice that the relative desirability of delayed differentiation increases with product variety. This is due to the inventory pooling effect that occurs when we postpone differentiation, either partially or completely. In both tables, we notice that system 2 is always superior to system 1 in terms of inventory carrying costs. The savings realized from using system 2 thus measure the amount of additional cost that is justifiable for product and process redesign to achieve complete postponement. Depending on the holding cost function and the number of products, system 3 may or may not be superior to systems 1 or 2 in terms of inventory carrying costs. However, in cases where finished inventory is expensive, as is the case with exponentially increasing holding costs, it is superior to the other two. These results confirm that when quick response is desired and the cost of product and process redesign is not very sensitive to t, the industry practice of complete postponement is indeed economical. However, it also shows that partial postponement can be superior to either no or full postponement when holding costs increase at an increasing rate with postponement. It is interesting to note that when greater product variety goes hand-in-hand with lower production volumes (Table 5), the inventory holding cost for system 1 tends to increase with product variety, while the inventory holding costs for the other two systems are unaffected. In contrast, when product variety is achieved by reducing work content (or improving efficiency), the inventory holding costs for all three systems tend to decrease with product variety. As a result, delayed differentiation, either partial or complete, appears to be more desirable in terms of the absolute cost difference when product variety is accompanied by lower production volumes.

7 Concluding Remarks

In the classical inventory based models, the DD strategy is always found superior to no postponement strategy in terms of inventory carrying costs. The relative advantage of DD strategy increases as more end products are made from the same undifferentiated parent. These results follow from inventory pooling advantage associated with the DD strategy. We add additional considerations of system load and order fulfillment time target to the evaluation of an option to delay differentiation. We believe these are issues that directly impact manufacturing. Our models show that the relative benefits of DD, the optimal size of the buffer, and the optimal point of differentiation depend in a non-trivial fashion on the interaction between the work content relative to system capacity, product mix, and the flexibility of the capacity/workforce. Surprizingly, the size of buffer and the optimal point of differentiation, are not affected a great deal by product and process redesign and warehouse costs.

When both order lead times and capacity are tight, point of differentiation should be pushed right

out to the customer stage. This can be seen in the computer industry where cost pressures demand high utilization of production facilities and it is possible to design products that can be customized at the point of sale by removing or adding modular parts. In other situations, where such redesign is not possible or where inventory costs are not affected much by the value-added through processing, companies should either keep finished goods inventories, or invest in more capacity. Whereas DD strategy is more beneficial when more end products are made from the same parent, whether full or partial differentiation is more economical depends on the inventory carrying cost function, and individual product volumes.

References

[1] Altiok, T., *Performance Analysis of Manufacturing Systems*, Springer Series in Operations Research, Editor: P. Glynn, Springer, New York, 1997.

[2] Baker, K. R. M. J. Magazine and H. L. W. Nuttle, "The Effect of Commonality on Safety Stocks in a Simple Inventory Model," *Management Science*, **32** (1986), 982-988.

[3] Buzacott, J. A. and J. G. Shanthikumar, *Stochastic Models of Manufacturing Systems*, Prentice Hall, NJ, 1993.

[4] Burke, P. J., "The Output of a Queueing System," *Operations Research*, **4** (1956), 699-704.

[5] Collier, D. A. "Aggregate Safety Stock Levels and Component Part Commonality," *Management Science*, **28** (1982), 1296-1303.

[6] Feitzinger, E. and H. Lee, "Mass Customization at Hewlett Packard: The Power of Postponement," *Harvard Business Review,* January-February 1997, 116-121.

[7] Garg, A. and C. S. Tang, "On Postponement Strategies for Product Families with Multiple Points of Differentiation," *IIE Transactions*, **29** (1997), 641-650.

[8] Gerchak, Y. and M. Henig, "An Inventory Model with Component Commonality," *Operations Research Letters*, **5** (1986), 157-160.

[9] Gerchak, Y. M. J. Magazine and A. B. Gamble, "Component Commonality with Service Level Requirements," *Management Science*, **34** (1988), 753-760.

[10] Graman, G. A. and M. J. Magazine, "An Analysis of Packaging Postponement," Proceedings of the 1998 MSOM Conference, School of Business Administration, The University of Washington, Seattle, June 29-30, 1998, 67-72.

[11] Harel, A. and P. H. Zipkin, "Strong Convexity Results for Queueing Systems," *Operations Research*, **35** (1987), 405-418.

[12] Jackson, J. R., "Networks of Waiting Lines," *Operations Research*, **5** (1957), 518-521.

[13] Kleinrock, L., *Queueing Systems, Volume I: Theory*, John Wiley and Sons, 1975.

[14] Lee, H. L., "Effective Inventory and Service Management Through Product and Process Redesign," *Operations Research,* **44** (1996), 151-159.

[15] Lee, H. L. and C. Billington, "Designing Products and Processes for Postponement," *Management of Design: Engineering and Management Perspectives,* S. Dasu and C. Eastman (Eds.), Kluwer Academic Publishers, Boston, MA, 1994, 105-122.

[16] Lee, H. L., and C. S. Tang, "Modeling the Costs and Benefits of Delayed Product Differentiation," *Management Science,* **43** (1997), 40-53.

[17] Luenberger, D. G., *Linear and Nonlinear Programming*, Addison-Wesley Publishing Company, Reading, MA, 1984.

[18] Swaminathan, J. M. and S. R. Tayur, "Managing Broader Product Lines Through Delayed Differentiation Using Vanilla Boxes," *Management Science* **44**, (1998), S161-S172.

Appendix

A Characterizing the stage-2 input process in the two-stage production system

LEMMA **A.1** *The input process to production stage-2 can be characterized as follows.*

$$A_2^*(s) = \begin{cases} \frac{\Lambda}{\Lambda+s} & \text{if } b = 0, \\ \frac{\Lambda}{\Lambda+s} - \frac{\Lambda\rho_1^b(\mu_1-\Lambda)s^2}{(\Lambda+s)(\mu_1+s)(\Lambda+\mu_1+s)^2} & \text{if } b \geq 1,\ n_1 = 1,\ \text{and } n = 2, \\ \frac{\Lambda}{\Lambda+s} - \frac{R(\rho_1/n_1)^b\Lambda n_1\mu_1 s^2}{(\Lambda+s)(n_1\mu_1+s)(\Lambda+n_1\mu_1+s)^2} & \text{otherwise,} \end{cases} \tag{A.1}$$

where $R = \frac{\pi_1(0)n_1^{n_1}}{n_1!}$ *is used for notational convenience.*

Proof: The above Lemma describes stage-2 input process for both the inflexible and the flexible workforce models. In the interest of brevity, we present a detailed proof for the inflexible workers model only, i.e., when $n = 2$ and $n_1 = 1$. Clearly, if $b = 0$, the stage-2 input process coincides with the departure process of stage-1. Here Burke's theorem applies that therefore $A_2^*(s) = \Lambda/\Lambda + s$.

Consider the jth arrival to stage-2 under steady state operations when $b \geq 1$. It coincides either with an external arrival (this happens when the intermediate buffer is not altogether empty), or with the moment of a service completion at stage-1. Let τ_j denote the arrival epoch of the jth arrival, and $A_{2,j}$ the inter-arrival time. Clearly,

$$A_{2,j} = \tau_j - \tau_{j-1}, \quad \forall\ j \geq 1,$$

where we define $\tau_0 = 0$. Let $Q_1^{(j)}$ denote the steady state number of customers in stage-1 at the moment of jth arrival to stage-2, and Q_1 denote the steady state number in stage-1 at an arbitrary observation epoch. Since stage-1 queue is an M/M/1 queue and each stage-2 arrival epoch coincides either with a stage-1 arrival epoch or with a stage-1 departure epoch, it can be argued that $Q_1^{(j)}$ is identical in distribution to Q_1 for all j.

We define three random variables X, X_1 and X_2. All three are exponentially distributed with parameters $\Lambda + \mu_1$, Λ and μ_1 respectively. Thus, X_1 is the time between consecutive external arrivals and X_2 is the time between departures from stage-1 when $Q_1^{(j)} > 0$. Similarly, X is the time until next event which could either be an arrival or a service completion at stage-1. Consider τ_j, $\forall\ j \geq 0$, the moment at which jth customer has just arrived. The time until next arrival to stage-2, $A_{2,j+1}$, can be written as:

$$A_{2,j+1} = \begin{cases} X_1 & \text{if } Q_1^{(j)} < b, \\ X + (\frac{\mu_1}{\Lambda+\mu_1})(X_1 \mid_{X_1>X_2}) & \\ \quad +(\frac{\Lambda}{\Lambda+\mu_1})(X_2 \mid_{X_2>X_1}) & \text{if } Q_1^{(j)} = b, \text{ and} \\ X_2 & \text{if } Q_1^{(j)} > b. \end{cases} \tag{A.2}$$

In A.2, the notation $X_1 \mid_{X_1>X_2}$ denotes the duration of X_1 given that $X_1 > X_2$. The following contingencies give rise to relationship A.2:

1. If at τ_j, $Q_1^{(j)} < b$, next arrival to stage-2 will coincide with the next external arrival. Since the forward recurrence time of X_1 is identical in distribution to X_1, it follows that $A_{j+1} = X_1$ whenever $Q_1^{(j)} < b$.

2. If at τ_j, $Q_1^{(j)} = b$, next arrival to stage-2 occurs after 2 events take place. The first event happens when either an arrival or a service completion occurs. Owing to the memoryless property of the exponential distribution, the time until this event has distribution X. The second event is either an arrival or a service completion depending on whether $X_1 > X_2$ or $X_2 > X_1$. This explains the two terms, $X_1 \,|_{X_1>X_2}$ and $X_2 \,|_{X_2>X_1}$, which represent the partial distributions of X_1 and X_2 respectively. $X_1 \,|_{X_1>X_2}$ follows X if the first event is a service completion which occurs with probability $\frac{\mu_1}{\Lambda+\mu_1}$. Similarly, $X_2 \,|_{X_2>X_1}$ follows X with probability $\frac{\Lambda}{\Lambda+\mu_1}$.

3. If at τ_j, $Q_1^{(j)} > b$, next arrival to stage-2 coincides with the next departure from stage-1. Since the latter has a forward recurrence time of X_2, it is clear that in this situation $A_{2,j+1}$ equals X_2.

Let $G_{X_1|X_1>X_2}(a)$ denote the cumulative distribution function of X_1 given that $X_1 > X_2$. Using notation g_Y to denote the probability density function of a random variable Y, we get

$$\begin{aligned}
G_{X_1|X_1>X_2}(a) &= \frac{\text{Prob}(X_1 \le a \text{ and } X_1 > X_2)}{\text{Prob}(X_1 > X_2)} \\
&= \frac{\int_{y=0}^{\infty} \text{Prob}(X_1 \le a \text{ and } X_1 > X_2 \mid X_2 = y) g_{X_2}(y) dy}{\int_{y=0}^{\infty} \text{Prob}(X_1 > X_2 \mid X_2 = y) g_{X_2}(y) dy} \\
&= \frac{\int_{y=0}^{a} (\int_{x=y}^{a} \Lambda e^{-\Lambda x} dx)\ \mu_1 e^{-\mu_1 y} dy}{\int_{y=0}^{\infty} e^{-\Lambda y} \mu_1 e^{-\mu_1 y} dy} \\
&= 1 - e^{-(\Lambda+\mu_1)a} - \left(\frac{\Lambda+\mu_1}{\mu_1}\right)(1 - e^{-\mu_1 a}) e^{-\Lambda a}.
\end{aligned}$$

Upon differentiating the above, we get the following expression for the PDF of $X_1 \,|_{X_1>X_2}$:

$$g_{X_1|X_1>X_2}(a) = \left(\frac{\Lambda+\mu_1}{\mu_1}\right)(1 - e^{-\mu_1 a})\Lambda e^{-\Lambda a}. \tag{A.3}$$

Let $\pi_1(r)$ denote the probability that stage-1 has $r \ge 0$ customers at an arbitrary observation epoch. Since $\pi_1(r) = \rho_1^r(1-\rho_1)$, we can write

$$\begin{aligned}
g_{A_{2,j+1}}(x) &= (1-\rho_1^b) g_{X_1}(x) + \rho_1^b (1-\rho_1) \left(\frac{\mu_1}{\Lambda+\mu_1} g_{X+X_1|X_1>X_2}(x) + \frac{\Lambda}{\Lambda+\mu_1} g_{X+X_2|X_2>X_1}(x) \right) \\
&+ \rho_1^{b+1} g_{X_2}(x).
\end{aligned} \tag{A.4}$$

Next, we take the Laplace Transforms on both sides of A.4 and simplify to obtain the following relationship:

$$\begin{aligned}
A_{2,j+1}{}^*(s) &= (1-\rho_1^b)\tfrac{\Lambda}{\Lambda+s} + \rho_1^b(1-\rho_1)\left(\tfrac{\Lambda+\mu_1}{\Lambda+\mu_1+s}\right)\left(\tfrac{\Lambda}{\Lambda+s} - \tfrac{\Lambda}{\Lambda+\mu_1+s} + \tfrac{\mu_1}{\mu_1+s} - \tfrac{\mu_1}{\Lambda+\mu_1+s}\right) \\
&+ \rho_1^{b+1}\left(\tfrac{\mu_1}{\Lambda+s}\right).
\end{aligned} \tag{A.5}$$

Since the distribution of $A_{2,j+1}$ (the RHS of A.5) is independent of the index j, we drop this subscript and simplify A.5 to obtain:

$$A_2^*(s) = \frac{\Lambda}{\Lambda+s} - \frac{\Lambda\rho_1^b(\mu_1-\Lambda)s^2}{(\Lambda+s)(\mu_1+s)(\Lambda+\mu_1+s)^2}. \tag{A.6}$$

Notice that $A_2^*(s) \rightarrow \frac{\Lambda}{\Lambda+s}$ as $b \rightarrow \infty$. This agrees with intuition: when buffer size is large, the two stages operate like two independent M/M/1 queues. Hence proved. #.

Next we show that $C_{A_2}^2 \approx 1$ and that the stage-2 can be approximated by a M/M/1 queue. From A.1, it is easy to obtain the following additional properties of A_2 for the inflexible workers model: $\bar{A}_2 = -A_2^{*\prime}(0) = \frac{1}{\Lambda}$, $E(A_2^2) = A_2^{*\prime\prime}(0) = \frac{2}{\Lambda^2}$ if $b = 0$, and $\frac{2}{\Lambda^2} - \frac{2\rho_1^b(1-\rho_1)}{\mu_1^2(1+\rho_1)^2}$ if $b \geq 1$. Similarly, $C_{A_2}^2 = 1$ if $b = 0$, and $1 - \frac{2\rho_1^{b+2}(1-\rho_1)}{(1+\rho_1)^2}$ if $b \geq 1$. Thus, we have a $M/M/1$ queue at stage-1 and a $G/M/1$ queue at stage-2. The steady-state queue length distribution at stage-2, $\pi_2(r)$, can be obtained as $\pi_2(r) = x^r(1-x)$, for $r = 0, 1, \cdots$, where $x \in (0,1)$ is a solution to the equation: $x = A_2^*(\mu_2 - \mu_2 x)$ (see, for example, Kleinrock, 1975, pp. 251).

Since we need closed form expressions to be able perform optimization, we set $C_{A_2}^2 \approx 1$, i.e., it treat stage-2 queue as a $M/M/1$ queue. The reasonableness for this approximation can be seen upon calculating $\Delta = [(1 - C_{A_2}^2)/C_{A_2}^2] \times 100$, or the percent error upon assuming $C_{A_2}^2 = 1$. Notice that for $b \geq 1$, the percent error Δ is monotonically decreasing as $b \rightarrow \infty$. Therefore, we set $b = 1$ and find that the maximum error is 7.7% and that it occurs when $\rho_1 = 0.686$. In fact for b greater than 1, the error quickly goes to zero, as shown below.

	$b=1$	$b=5$	$b=10$	$b=20$	$b=40$	$b=80$
$\rho_1 = 0.2$	0.897	0.001	0.000	0.000	0.000	0.000
$\rho_1 = 0.3$	2.288	0.018	0.000	0.000	0.000	0.000
$\rho_1 = 0.4$	4.078	0.100	0.001	0.000	0.000	0.000
$\rho_1 = 0.5$	5.882	0.348	0.011	0.000	0.000	0.000
$\rho_1 = 0.6$	7.239	0.883	0.068	0.000	0.000	0.000
$\rho_1 = 0.7$	7.667	1.740	0.288	0.008	0.000	0.000
$\rho_1 = 0.8$	6.747	2.658	0.856	0.091	0.001	0.000
$\rho_1 = 0.9$	4.209	2.722	1.590	0.549	0.066	0.001

For most practical situations, we expect b to be greater than 1 in order to keep response times fast. Also, the intended use of our models lies in pre-design evaluation of buffer size and location alternatives. Given these facts, we believe assuming $C_{A_2}^2 \approx 1$ is a reasonable approximation. A close match was also observed between exact mean order fulfillment times (obtained by solving the G/M/1 queue), and the corresponding approximate values obtained from assuming $C_{A_2}^2 = 1$. For example, if we set $b = 1$, $\Lambda = 0.686$, and $\mu_1 = 1$ (notice that this is a case in which our approximation is most inaccurate), we observe that our approximation overestimates the exact value of the mean order fulfillment time at stage-2 by 2.85%.

For the flexible workforce model, the stage-2 queueing process is a $G/M/n_2$ process with $C_{A_2}^2 = 1 - \frac{2R(\rho_1/n_1)^{b+2}}{(1+\rho_1/n_1)^2}$ (follows from A.1). Since the exact mathematical expression for the mean order delay for such systems is not known, we once again approximate $C_{A_2}^2 \approx 1$. So long as we maintain $n_1 \leq b$ (which is a reasonable assumption in this case), the maximum overestimation error Δ is still 7.7%, and it happens when $b = n_1 = 1$ and $\rho_1 = 0.686$. Numerical experiments also confirm that $C_{A_2}^2$ rapidly approaches 1 whenever $n_1 > 1$.

B Performance Metrics for the Inflexible Workforce Model

Recall that Q_i denote the steady state number of in-process jobs at stage-i at an arbitrary moment of observation. It is easy to see that the number of semi-finished items in the intermediate buffer $= \max(0, b - Q_1)$ and that Q_1 is independent of stage-2. Therefore, the average buffer inventory can

be written as:

$$\bar{I} = \sum_{r=0}^{b} (b-r)\pi_1(r), \tag{B.7}$$

where $\pi_i(r) = \text{Prob}\{Q_i = r\}$, for $r \geq 0$. We obtain 6 upon simplifying the above sum after substituting $\pi_1(r) = (\Lambda t)^r(1-\Lambda t)$. Similarly, since $S(b,t) = Q_2 + \max(0, Q_1 - b)$, we have

$$\bar{S} = E(Q_2) + \sum_{r=b}^{\infty} (r-b)\pi_1(r) \tag{B.8}$$

After substituting for $E(Q_2)$ and $\pi_1(r)$, and simplifying we obtain

$$\bar{S} = \frac{\rho_2}{1-\rho_2} + \frac{\rho_1^{b+1}}{1-\rho_1}, \tag{B.9}$$

which is 9.

To obtain $\bar{F}$, note that each arrival must wait for processing at stage-2, whereas it waits at stage-1 only when the intermediate buffer is empty. Furthermore, since each external arrival is a Poisson arrival, it sees time average behavior. Thus,

$$\bar{F} = \bar{F}_1 + \bar{F}_2 = \sum_{r=b}^{\infty} (r-b+1)t\pi_1(r) + \frac{(T-t)}{1-\Lambda(T-t)}, \tag{B.10}$$

where the second term on the right hand side of B.10 is the expected order delay in a $M/M/1$ queue. Simplification of equation B.10 yields 7.

Since F_2 is the delay experienced by a job in a $M/M/1$ queue with arrival rate Λ and service rate μ_2, F_2 is exponentially distributed as follows (see, for example, Buzacott and Shanthikumar (1993), pp. 56):

$$\text{Prob}\{F_2 \leq y\} = 1 - e^{-\mu_2(1-\rho_2)y}. \tag{B.11}$$

The delay an arriving demand experiences at stage-1 depends on Q_1 and its conditional distribution can be written as follows:

$$\text{Prob}\{F_1 \leq y \mid Q_1 = r\} = \begin{cases} 1 & \text{if } r < b, \\ \sum_{\ell=r-b+1}^{\infty} \frac{(\mu_1 y)^\ell}{\ell!} e^{-\mu_1 y} & \text{otherwise.} \end{cases}, \quad \text{for all } y \geq o. \tag{B.12}$$

Clearly,

$$\begin{aligned} \text{Prob}\{F \leq x\} &= \text{Prob}\{F_1 = 0 \text{ and } F_2 \leq x\} + \textstyle\int_{y=0^+}^{x} \text{Prob}\{F_1 = y\}\text{Prob}\{F_2 \leq x-y\}dy, \\ &= (1-\rho_1^b)(1-e^{-\mu_2(1-\rho_2)x}) + \textstyle\int_{y=0}^{x} (\rho_1^b \mu_1(1-\rho_1)e^{-\mu_1(1-\rho_1)y})(1-e^{-\mu_2(1-\rho_2)(x-y)})dy. \end{aligned} \tag{B.13}$$

The above expression, after some simplification, yields 8.

Let S_i denote the number of jobs backordered at stage-i. Then,

$$\text{Prob}\{S_1 = k\} = \begin{cases} \text{Prob}\{Q_1 = b+k\} & \text{if } k > 0, \\ \text{Prob}\{Q_1 \leq b\} & \text{otherwise,} \end{cases} \quad \text{for all } k \geq 0. \tag{B.14}$$

$$\text{Prob}\{S_2 = k\} = \text{Prob}\{Q_2 = k\}, \qquad \text{for all } k \geq 0. \tag{B.15}$$

Since $S = S_1 + S_2$, we have

$$\begin{aligned} \text{Prob}\{S = \ell\} &= \sum_{k=0}^{\ell} \text{Prob}\{S_1 = k\}\ \text{Prob}\{S_2 = \ell - k\} \\ &= (1-\rho_1)(1-\rho_2)\left(\sum_{k=1}^{\ell} \rho_1^{b+k}\rho_2^{\ell-k} + \sum_{i=0}^{b} \rho_1^i \rho_2^\ell\right). \end{aligned} \tag{B.16}$$

Upon simplifying the right hand side above, we obtain

$$\text{Prob}\{S = \ell\} = \begin{cases} \frac{(1-\rho_1)(1-\rho_2)(\rho_2^\ell - \rho_1^\ell)\rho_1^{b+1}}{\rho_2 - \rho_1} + (1-\rho_2)(1-\rho_1^{b+1})\rho_2^\ell & \text{if } \rho_1 \neq \rho_2, \\ \ell(1-\rho)^2\rho^{b+m} + (1-\rho)(1-\rho^{b+1})\rho^m & \text{when } \rho_1 = \rho_2 = \rho. \end{cases} \tag{B.17}$$

Clearly, $Sx = \sum_{\ell=x}^{\infty} \text{Prob}\{S = \ell\}$. Summing both sides of B.17 we obtain 10.

C Properties of Performance Metrics for the Inflexible Workforce Model

First, we reiterate the properties of functions $\bar{I}$, $\bar{F}$, $\bar{S}$, Fx, and Sx as mentioned in Proposition 1. These properties are:

a For each fixed t, $\bar{I}$ is an increasing convex function of b.

b For each fixed b, $\bar{I}$ is decreasing in t. However, it is neither convex not concave in t.

c For each fixed t, $\bar{F}$, $\bar{S}$, Fx and Sx are a decreasing convex functions of b.

d For each fixed b, $\bar{F}$ is a convex function of t. However, it is not monotonic in t.

Proof: In order to prove the assertions above, we differentiate $\bar{I}$ and $\bar{F}$ with respect to b and t to obtain:

$$\bar{I}_b = \frac{\partial \bar{I}}{\partial b} = 1 + \frac{(\Lambda t)^{b+1}\ln(\Lambda t)}{1-\Lambda t}. \tag{C.18}$$

$$\bar{I}_{bb} = \frac{\partial^2 \bar{I}}{\partial b^2} = \frac{(\Lambda t)^{b+1}(\ln(\Lambda t))^2}{1-\Lambda t}. \tag{C.19}$$

$$\bar{I}_t = \frac{\partial \bar{I}}{\partial t} = \Lambda\left[\frac{b(\Lambda t)^b(1-\Lambda t) - (1-(\Lambda t)^b)}{(1-\Lambda t)^2}\right]. \tag{C.20}$$

$$\bar{I}_{tt} = \frac{\partial^2 \bar{I}}{\partial t^2} = \frac{(\Lambda)^2\{b^2(\Lambda t)^{b-1}(1-\Lambda t)^2 + b(\Lambda t)^{b-1}[1-(\Lambda t)^2] - 2[1-(\Lambda t)^b]\}}{(1-\Lambda t)^3}. \tag{C.21}$$

$$\bar{F}_b = \frac{\partial \bar{F}}{\partial b} = \frac{t(\Lambda t)^b \ln(\Lambda t)}{1-\Lambda t}. \tag{C.22}$$

$$\bar{F}_{bb} = \frac{\partial^2 \bar{F}}{\partial b^2} = \frac{t(\Lambda t)^b[\ln(\Lambda t)]^2}{1-\Lambda t}. \tag{C.23}$$

$$\bar{F}_t = \frac{\partial \bar{F}}{\partial t} = \frac{(\Lambda t)^b[b(1-\Lambda t)+1]}{(1-\Lambda t)^2} - \frac{1}{(1-\Lambda(T-t))^2}. \tag{C.24}$$

$$\bar{F}_{tt} = \frac{\partial^2 \bar{F}}{\partial t^2} = \frac{\Lambda b^2 (\Lambda t)^{b-1}}{1-\Lambda t} + \frac{\Lambda b (\Lambda t)^{b-1}(1+\Lambda b)}{(1-\Lambda t)^2} + \frac{2\Lambda(\Lambda t)^b}{(1-\Lambda t)^3} + \frac{2\Lambda}{(1-\Lambda(T-t))^3}. \tag{C.25}$$

$$\bar{S}_b = \frac{\partial \bar{S}}{\partial b} = \frac{\rho_1^{b+1}\ln(\rho_1)}{1-\rho_1} \tag{C.26}$$

$$\bar{S}_{bb} = \frac{\partial^2 \bar{S}}{\partial b^2} = \frac{\rho_1^{b+1}(\ln(\rho_1))^2}{1-\rho_1} \tag{C.27}$$

$$(Fx)_b = \frac{\partial Fx}{\partial b} = \begin{cases} \left(\frac{(1/(T-t)-\Lambda)[\ln(\Lambda t)](\Lambda t)^b}{1/t-1/(T-t)}\right)\left(e^{-[1/(T-t)-\Lambda]x} - e^{-[1/t-\Lambda]x}\right) & \text{if } t \neq T/2, \\ e^{-[2/T-\Lambda]x}\left(x(\Lambda T/2)^b[2/T-\Lambda]\ln(\Lambda T)\right) & \text{otherwise.} \end{cases} \tag{C.28}$$

$$(Fx)_{bb} = \frac{\partial^2 Fx}{\partial b^2} = \begin{cases} \left(\frac{(1/(T-t)-\Lambda)[\ln(\Lambda t)]^2(\Lambda t)^b}{1/t-1/(T-t)}\right)\left(e^{-[1/(T-t)-\Lambda]x} - e^{-[1/t-\Lambda]x}\right) & \text{if } t \neq T/2, \\ e^{-[2/T-\Lambda]x}\left(x(\Lambda T/2)^b[2/T-\Lambda][\ln(\Lambda T)]^2\right) & \text{otherwise.} \end{cases} \tag{C.29}$$

$$(Sx)_b = \frac{\partial Sx}{\partial b} = \begin{cases} \left(\frac{\rho_1^{b+1}\ln(\rho_1)(1-\rho_2)}{\rho_2-\rho_1}\right)(\rho_2^x - \rho_1^x) & \text{if } \rho_1 \neq \rho_2, \\ x(1-\rho)\rho^{b+x}\ln(\rho) & \text{if } \rho_1 = \rho_2 = \rho. \end{cases} \tag{C.30}$$

$$Sx_{bb} = \frac{\partial^2 Sx}{\partial b^2} = \begin{cases} \left(\frac{\rho_1^{b+1}[\ln(\rho_1)]^2(1-\rho_2)}{\rho_2-\rho_1}\right)(\rho_2^x - \rho_1^x) & \text{if } \rho_1 \neq \rho_2, \\ x(1-\rho)\rho^{b+x}[\ln(\rho)]^2 & \text{if } \rho_1 = \rho_2 = \rho. \end{cases} \tag{C.31}$$

Since, for any $x < 1$ and $x \neq 0$, $x < -\ln(1-x) < x/(1-x)$, letting $x = 1-\Lambda t$, we see that $1 < \frac{-\ln(\Lambda t)}{1-\Lambda t} < \frac{1}{\Lambda t}$. Therefore, the RHS of equation C.18, which can also be written as $1-(\Lambda t)^b\left(\frac{-\ln(\Lambda t)}{1-\Lambda t}\right)$, takes values in the range $1-(\Lambda t)^{b+1}$ and $1-(\Lambda t)^b$. Since this range is entirely positive for all feasible values of Λt and b, this proves that $\bar{I}$ is increasing in b. Furthermore, since the RHS of equation C.19 is positive, this implies that $\bar{I}$ is increasing convex in b.

In order to prove that $\bar{I}$ is decreasing in t, we need to show that $b(\Lambda t)^b \leq (1-(\Lambda t)^b)/(1-\Lambda t)$. Notice that the right hand side of this inequality can be written as $\sum_{i=0}^{b-1}(\Lambda t)^i$. Since, $\Lambda t < 1$, it is now easy to see that the inequality holds for all b. Examining the RHS of equation C.21, it is revealed that it can take positive as well as negative values depending on the values of parameters b and t. Thus, whereas $\bar{I}$ is decreasing in t, it is neither convex nor concave in the entire range of b and t.

Notice that the RHS of equations C.22, C.26, C.28 and C.30 is negative since $\rho_1 = \Lambda t < 1$ and therefore $\bar{F}$, $\bar{S}$, Fx, and Sx are decreasing in b. Similarly, the RHS of C.23, C.27, C.29, and C.31 is always positive, and this confirms that $\bar{F}$, $\bar{S}$, Fx and Sx are decreasing convex in b.

Finally, from equation C.24 we see that the derivative of $\bar{F}$ with respect to t is a large positive number when $\Lambda t \to 1$ and it is a large negative number when $\Lambda(T-t) \to 1$. However, the second derivative of $\bar{F}$ is always positive. Thus, $\bar{F}$ is a convex non-monotonic function of t. This completes the proofs of all of our assertions above. #

D Algorithm for obtaining optimal b and t with inflexible workforce

A complete solution to optimization problem 11 - 17 can be obtained by applying the optimality conditions. The following algorithm, presented with constraint 12 in mind, results in identification of

optimal b and t. Similar algorithms also exist when 12 is replaced by other service level constraints.

1. For any t, $\bar{F}$ is decreasing in b and achieves its minimum at $b = \infty$. Furthermore, $\bar{F}(\infty, t) = (T - t)/[1 - \Lambda(T - t)]$ is positive and strictly decreasing in t, and t in turn is bounded above by $1/\Lambda$. Therefore, the minimum mean order fulfillment time achievable is $\bar{F}_{\min} = \max\{0, (T - 1/\Lambda)/(2 - \Lambda T)\}$. Therefore, if $\alpha \leq \max\{0, (T - 1/\Lambda)/(2 - \Lambda T)\}$, then there is no feasible solution to the problem described in 11 - 17, since constraint 12 cannot be satisfied. In the remaining three instances listed below, we assume that $\alpha > \max\{0, (T - 1/\Lambda)/(2 - \Lambda T)\}$.

2. If $\bar{F}(0,t) = \frac{t}{1-\Lambda t} + \frac{T-t}{1-\Lambda(T-t)} \leq \alpha$, then $b = 0$ is feasible. It is easy to confirm that $\bar{F}(0,t)$ is convex in t and achieves its minimum value of $\frac{2T}{2-\Lambda T}$ at $t = T/2$. Furthermore, if $T < 1/\Lambda$, $\bar{F}(t,0)$ has a finite maximum value of $\frac{T}{1-\Lambda T}$. This implies that if $T < 1/\Lambda$ and $\alpha \geq \frac{T}{1-\Lambda T}$, then make all items to order, i.e., $b^* = 0$ and $t^* = 0$.

3. If either (a) $\alpha \geq \frac{2T}{2-\Lambda T}$ and $T > 1/\Lambda$, or (b) $\frac{2T}{2-\Lambda T} \leq \alpha < \frac{T}{1-\Lambda T}$ and $T < 1/\Lambda$, then it is possible to satisfy order lead time constraint while keeping $b = 0$. However, $t > 0$ and as a result overall costs may be lower if by choosing $b > 0$, the redesign costs, $R(t)$, could be lowered sufficiently to more than offset the increased cost of holding inventory.

   Let t_1 and t_2 be the roots obtained in $[0, T]$ by solving the quadratic equation $\bar{F}(0,t) = \frac{t}{1-\Lambda t} + \frac{T-t}{1-\Lambda(T-t)} = \alpha$. Then $t(0)$, the optimal work content of undifferentiated items in a make-to-order system, equals $\min[t_1, t_2]$. If, on the other hand, we choose $b > 0$, the order delay constraint must be tight. This allows us to find the optimal t by first substituting b in terms of t and solving a single variable optimization problem over a finite range. Overall optimal (b, t) pair is the one that minimizes $K(b, t)$.

4. If $\max\{0, \frac{T-1/\Lambda}{2-\Lambda T}\} < \alpha < \frac{2T}{2-\Lambda T}$, then $b^* > 0$, i.e., it is optimal to keep some semi-processed inventory. In this case, the mean order fulfillment time constraint 12 is tight. Like in the previous step, here we solve a minimization problem in one variable by first substituting b in terms of t.

Steps 1-4 above divide the parameter space (determined by parameters T, Λ, M, and α) into four contiguous regions. In each case, we obtain a strategy for finding the optimal b, t pair. In the first region the problem has no solution. In the second region $b = t = 0$ is feasible. It is also optimal since $K(0,0) = 0$. In regions 3 and 4, either $b^* = 0$, $t^* > 0$, which yields $K(0, t^*) = R(t^*)$, or $b^* > 0$ in which case constraint 12 is tight. This reduces the optimization problem to a minimization problem in a single variable, after we first substitute b in terms of t.

We need to accommodate the fact that b can only take integer values in all those steps that involve substitution of b in terms of t. We suggest that both integer ceiling and floor of non-integer be considered and that we choose the value that minimizes $K(b, t)$.

E Properties of Performance Metrics for the Flexible Workforce Model

First, we reiterate the properties of functions $\bar{I}$ and $\bar{F}$ mentioned in Proposition 3. The properties are:

a For fixed t and n_1, $\bar{I}$ is an increasing convex function of b.

b For fixed b and n_1, $\bar{I}$ is decreasing in t. However, it is neither convex not concave in t.

c For fixed t and n_1, $\bar{F}$ is a decreasing convex function of b.

d For fixed b and n_1, $\bar{F}$ is a convex function of t. However, it is not monotonic in t.

e For fixed b and t, $\bar{I}$ is increasing in n_1. However, it is neither convex nor concave.

Proof: In order to prove the assertions above, we differentiate $\bar{I}$ and $\bar{F}$ from expressions 33 and 35 with respect to b and t to obtain:

$$\frac{\partial \bar{I}}{\partial b} = 1 + \frac{B_1(\rho_1/n_1)^{b-n_1+1}\ln(\rho_1/n_1)}{1-\rho_1/n_1}. \tag{E.32}$$

$$\frac{\partial^2 \bar{I}}{\partial b^2} = \frac{B_1(\rho_1/n_1)^{b-n_1+1}(\ln(\rho_1/n_1))^2}{1-\rho_1/n_1}. \tag{E.33}$$

$$\begin{aligned}\frac{\partial \bar{I}}{\partial t} &= -\lambda + \left(\frac{1}{(1-\rho_1/n_1)^2}\right)\Big(B_1(\lambda/n_1)[(b+1-n_1)(\rho_1/n_1)^{b-n_1}(1-\rho_1/n_1) \\ &- (1-\rho_1/n_1)^{b+1-n_1}] - B_1'(\rho_1/n_1)(1-\rho_1/n_1)(1-(\rho_1/n_1)^{b+1-n_1})\Big),\end{aligned} \tag{E.34}$$

where we use the prime notation to denote derivative with respect to t, i.e., $B_i' = \frac{\partial B_i}{\partial t}$.

$$\frac{\partial \bar{F}}{\partial b} = \frac{B_1 t(\rho_1/n_1)^{b-n_1}\ln(\rho_1/n_1)}{n_1-\lambda t}. \tag{E.35}$$

$$\frac{\partial^2 \bar{F}}{\partial b^2} = \frac{B_1 t(\rho_1/n_1)^{b-n_1}[\ln(\rho_1/n_1)]^2}{n_1-\lambda t}. \tag{E.36}$$

$$\frac{\partial \bar{F}}{\partial t} = \frac{U}{(n_1-\lambda t)^2} + \frac{V}{[(n-n_1)-\lambda(T-t)]^2} - 1, \tag{E.37}$$

where

$$U = B_1(\lambda t/n_1)^{b-n_1}\{(b+1-n_1)(n_1-\lambda t)+\lambda t\} + B_1' t(\lambda t/n_1)^{b-n_1}(n_1-\lambda t), \tag{E.38}$$

and

$$V = B_2'(T-t)\{(n-n_1)-\lambda(T-t)\} - B_2(n-n_1). \tag{E.39}$$

$$\frac{\partial^2 \bar{F}}{\partial t^2} = \frac{U'[(n_1-\lambda t)^2]+2\lambda(n_1-\lambda t)U}{(n_1-\lambda t)^4} + \frac{V'[(n-n_1)-\lambda(T-t)]^2 - 2\lambda V[(n-n_1)-\lambda(T-t)]}{[(n-n_1)-\lambda(T-t)]^4}, \tag{E.40}$$

where

$$\begin{aligned}U' &= 4B_1'(\lambda t/n_1)^{b-n_1} + B_1\{(b-n_1)(\lambda t/n_1)^{b-n_1-1}(\lambda/n_1)(b+1-n_1)(n_1-\lambda t\} \\ &+ B_1'' t(\lambda t/n_1)^{b-n_1}(n_1-\lambda t),\end{aligned} \tag{E.41}$$

and

$$V' = B_2''(T-t)[(n-n_1)-\lambda(T-t)] - 2B_2'[(n-n_1)-\lambda(T-t)]. \tag{E.42}$$

$$\frac{\partial \bar{I}}{\partial n_1} = \frac{W'(1-\rho_1/n_1) - W(\rho_1/n_1^2)}{(1-\rho_1/n_1)^2}, \tag{E.43}$$

where

$$W = B_1\{(\rho_1/n_1)^{b+1-n_1} - (\rho_1/n_1)\}, \tag{E.44}$$

and

$$W' = \frac{\partial W}{\partial n_1} = (\partial B_1/\partial n_1)\{(\rho_1/n_1)^{b+1-n_1} - (\rho_1/n_1)\} + B_1\{(b+1-n_1)(\rho_1/n_1)^{b-n_1} - \ln(\rho_1/n_1)(\rho_1/n_1)^{b+1-n_1} + \rho_1/n_1^2\}. \quad \text{(E.45)}$$

Upon observing that $1 < \frac{-\ln(\rho_1/n_1)}{(1-\rho_1/n_1)} < 1/(\rho_1/n_1)$, we can use these bounds in E.32 and show that $\partial\bar{I}/\partial b$ lies in the interval $(1 - B_1(\rho_1/n_1)^{b+1-n_1},\ 1 - B_1(\rho_1/n_1)^{b-n_1})$. Since this interval is entirely positive for all feasible values of ρ_1, n_1, and b, we have thus proved that $\bar{I}$ is increasing in b. Also, notice that the right hand side of E.33 is always positive which confirms that $\bar{I}$ is convex in b.

From previously known convexity results for $M/M/n$ queues, it is well known that B_i is increasing convex in ρ_i/n_i (see, for example, Harel and Zipkin (1987)). Therefore, $B_1' \geq 0$, and from E.34, $\bar{I}$ is decreasing in t if and only if

$$(b+1-n_1)(\rho_1/n_1)^{b-n_1}(1-\rho_1/n_1) - (1-(\rho_1/n_1)^{b+1-n_1}) \leq 0.$$

The above inequality always holds since $\rho_1/n_1 \leq 1$ and therefore for any integer $x \geq 0$, the following must hold: $x(\rho_1/n_1)^{x-1} \leq \sum_{i=0}^{x-1}(\rho_1/n_1)^i$. Next, the second derivative of $\bar{I}$ with respect to t can be shown to take both negative and positive values. Therefore, $\bar{I}$ is neither convex nor concave in t.

That $\bar{F}$ is decreasing and convex in b follows directly from expressions E.35 and E.36 which show that $\partial\bar{F}/\partial b \leq 0$ and $\partial^2\bar{F}/\partial b^2 \geq 0$.

While both B_i's are increasing convex in ρ_i/n_i, ρ_2/n_2 is decreasing in t. Therefore, $B_2' \leq 0$ and as a result $U \geq 0$ whereas $V \leq 0$. From here we see that $\partial\bar{F}/\partial t$ is a large positive number when $\lambda t/n_1 \to 1$ and a large negative number when $\lambda(T-t)/(n-n_1) \to 1$. Thus, we have proved that average order fulfillment time is not monotonic in t. Using relationships E.41 and E.42 and the facts that B_1', $B_1'' \geq 0$ whereas $B_2' \leq 0$ and $B_2'' \geq 0$, it is possible to show that both U' and V' are positive. These results are sufficient to prove that the right hand side of E.40 is positive, which proves that $\bar{F}$ is convex in t.

Applying the chain rule of differentiation, we see that $\partial B_1/\partial n_1 \leq 0$ which proves that $W' \geq 0$ (see the right hand side of E.45). Furthermore since $W \leq 0$ (from E.44), we have that the right hand side of E.43 is positive. Thus, $\bar{I}$ is increasing in n_1. Continuing in this way, it can be shown that the second derivative of $\bar{I}$ with respect to n_1 takes both positive and negative values. This confirms that $\bar{I}$ is neither convex nor concave in n_1. #

F Algorithm for obtaining optimal b and t with flexible workforce

Consider first the case when $n_1 = 0$ or $n_1 = n$. Let notation $t^*(m)$ and $b^*(m)$ denote the optimal values of t and b when $n_1 = m$. If $n_1 = 0$, no work can be assigned to stage-1, and hence $t^*(0) = 0$, $\bar{I}(b,0,0) = b$ whereas $\bar{F}(b,0,0) = \frac{B_2 T}{n-\Lambda T} + T$. This corresponds to a completely make-to-order situation and in this case it is clear that $b^*(0) = 0$ since b has no effect on mean order fulfillment time. If feasible, i.e., if $\frac{B_2 T}{n-\Lambda T} + T \leq \alpha$, then this is also the overall optimum solution, since $K(0,0,0) = 0$ is the smallest attainable cost.

On the other hand, when $n_1 = n$, no work can be assigned to stage-2, i.e., $t^*(n) = T$. This results in a completely make-to-stock system with $\bar{I}(b,n,T) = b - \Lambda T - \frac{B_1(\Lambda T/n)}{(1-\Lambda T/n)}\{1-(\Lambda T/n)^{b-n}\}$ and $\bar{F}(b,n,T) = \frac{B_1(\Lambda T/n)^{b-n}T}{n-\Lambda T}$. Since $\bar{I}(b,n,t)$ is increasing in b, $b^*(n)$ is the smallest integer b for which

$\bar{F}(b, n, T) \leq \alpha$. Solving for $b^*(n)$, we obtain:

$$b^*(n) = \begin{cases} \lceil \ln\left(\frac{\alpha(n-\Lambda T)}{B_1 T}\right) / \ln(\Lambda T/n) + n \rceil & \text{if } \alpha < B_1 T/(n - \Lambda T), \\ n & \text{otherwise.} \end{cases} \tag{F.46}$$

Notice that in this case, a feasible b always exists so long as $\alpha > 0$ since $\bar{F}(b, n, T)$ can be made arbitrarily small by choosing a large b.

Next, we outline steps necessary to find optimal b and t when $n_1 = m$, $1 \leq m \leq n-1$. As with the inflexible workforce model, here too either $b^*(m) = m$ or constraint 37 is binding. Thus, for each fixed n_1, the optimization problem 36 - 45 can be reduced to that of finding the minimum of a function of t over a finite range. This results in an efficient computational algorithm.

Let m denote a value of n_1 in the interval $[1, n-1]$, $t_m = argmin_t \bar{F}(m, m, t)$ (i.e., the value of t that minimizes 35 after setting $b = n_1 = m$), $\hat{t}_m = argmin_t \bar{F}(\infty, m, t) = \min(T, m/\Lambda)$[1], and let $b^*(m)$, $t^*(m)$ be the optimal values of b and t with $n_1 = m$. Then, for $1 \leq m \leq n-1$, $b^*(m)$ and $t^*(m)$ can be found as follows:

1. If $\bar{F}(m, m, 0) \leq \alpha$ and $\Lambda T \leq n - m$, then the order fulfillment time constraint can be met even when we idle m workers and produce to order. Although the requirement that $b \geq m$ implies that we will set $b = m$, this is a technical point. In practice, no work is assigned to stage-1 and therefore no buffer is needed. We resolve this situation by setting $t^*(m) = 0$, $b^*(m) = m$, and $K(m, m, 0) = 0$. We can write the requirement $\bar{F}(m, m, 0) \leq \alpha$ in the following alternate form
$$\frac{B_2 T}{n - m - \Lambda T} + T \leq \alpha. \tag{F.47}$$
On the other hand, if $\bar{F}(m, m, 0) > \alpha$, then $t^*(m) > 0$ and the following cases arise.

2. If $\bar{F}(m, m, t_m) \leq \alpha < \bar{F}(m, m, 0)$, then the equation $\bar{F}(m, m, t) = \alpha$ has at least one and at most two real root in $[0, T]$. Two roots exist in all cases except the following two: (a) $\Lambda T < m$ and $\bar{F}(m, m, T) < \alpha$, or (b) $\alpha = \bar{F}(m, m, t_m)$. Let the roots be denoted by t_1 and t_2 where $t_2 = T$ if only one root exists. Two cases now arise. If $b^*(m) = m$, then the optimal t is a value of t in the range $[t_1, t_2]$ which minimizes $K(m, m, t)$. On the other hand, if $b^*(m) > m$, then the response time constraint 37 is binding. In that case, we can substitute b in terms of t in the objective function and solve a single variable optimization model to obtain the optimal t, as we did in model 1. Overall optimum b, and t are obtained by comparing the objective function value at the two relative minima (one with $b^*(m) = m$ and the other with $b^*(m) > m$).

3. If $\bar{F}(\infty, m, \hat{t}_m) < \alpha < \bar{F}(m, m, t_m)$, then $b^*(m) > m$ and the response time constraint 37 is binding. Therefore, we can substitute b in terms of t in the objective function and solve a single variable optimization model, similar to the approach taken in model 1.

4. Finally, if $\alpha \leq \bar{F}(\infty, m, \hat{t}_m)$, a feasible solution with $n_1 = m$ does not exist. In this situation, we set $\bar{I}(b^*(m), m, t^*(m)) = \infty$.

Whenever constraint 37 is binding, we can obtain b as a function of t as shown below:

$$b = n_1 + \ln\left[\left(\alpha - (T-t) - \frac{B_2(T-t)}{n - n_1 - \Lambda(T-t)}\right)\left(\frac{n_1 - \Lambda t}{B_1 t}\right)\right] / \ln(\rho_1/n_1). \tag{F.48}$$

Substituting this into 36, and setting $n_1 = m$, we obtain $K_m(t)$. In steps 2 and 3 of the algorithm we need to search for a t that minimizes $K_m(t)$ in the feasible range described by 38 - 45.

[1]Since $\bar{F}(\infty, m, t) = B_2(T-t)/[n - m - \Lambda(T-t)] + (T-t)$ is decreasing in t, it is minimized at the largest possible of value of t, which is the smaller of T and m/Λ.

G On the point of differentiation that minimizes order delay

PROPOSITION 5 *If $n_1 \geq n/2$, then $t(b)/T \geq n_1/n$, where $t(b) = argmin_t \bar{F}(b,t)$, is the point of differentiation that minimizes mean order delay.*

Proof: We show the result holds for the flexible workforce model, of which the inflexible worker assignment model is a special case.

For notational convenience, we use $\nu_i = \rho_i/n_i$. Casting equation 35 into this new notation, we can simplify it as follows:

$$\bar{F} = \bar{F}_1 + \bar{F}_2, \tag{G.49}$$

where

$$\bar{F}_1 = \frac{B_1 \nu_1^{b+1-n_1}}{\Lambda(1-\nu_1)}, \tag{G.50}$$

and

$$\bar{F}_2 = \frac{B_2 \nu_2}{\Lambda(1-\nu_2)} + \frac{\nu_2(n-n_1)}{\Lambda}. \tag{G.51}$$

We differentiate $\bar{F}_i$ using the chain rule as follows:

$$\frac{\partial \bar{F}_i}{\partial t} = (\frac{\partial \bar{F}_i}{\partial \nu_i})(\frac{\partial \nu_i}{\partial t}). \tag{G.52}$$

Notice that B_i are increasing and convex in ν_i (Harel and Zipkin, 1987). Differentiating with respect to ν_i yields

$$\frac{\partial \bar{F}_1}{\partial \nu_1} = \frac{\{B_1'\nu_1(1-\nu_1) + B_1[(1-\nu_1)(b+1-n_1)+\nu_1]\}\nu_1^{b-n_1}}{\Lambda(1-\nu_1)^2}. \tag{G.53}$$

$$\frac{\partial \bar{F}_2}{\partial \nu_2} = \frac{B_2'\nu_2(1-\nu_2) + B_2}{\Lambda(1-\nu_2)^2} + \frac{n-n_1}{\Lambda}. \tag{G.54}$$

Notice that since $\nu_1 < 1$, $\sum_{i=0}^{b-n_1} \nu_1^i$ equals $[1-\nu_1^{b+1-n_1}]/(1-\nu_1)$ which is greater than or equal to $(b+1-n_1)\nu_1^{b-n_1}$. Substituting above, we obtain:

$$\begin{aligned} \frac{\partial \bar{F}_1}{\partial \nu_1} &\leq \frac{B_1'\nu_1(1-\nu_1)\nu_1^{b-n_1} + B_1}{\Lambda(1-\nu_1)^2}, \\ &\leq \frac{B_1'\nu_1(1-\nu_1) + B_1}{\Lambda(1-\nu_1)^2}. \end{aligned} \tag{G.55}$$

Let

$$X_i = \frac{B_i'\nu_i(1-\nu_i) + B_i}{\Lambda(1-\nu_i)^2}. \tag{G.56}$$

Since B_i's are increasing convex in ν_i, it is possible to show that $X_1 \leq X_2$ so long as $\nu_1 \leq \nu_2$. Using this notational simplification, we can also write

$$\frac{\partial \bar{F}}{\partial t} \leq X_1(\frac{\Lambda}{n_1}) - (\frac{\Lambda}{n-n_1})(X_2 + \frac{n-n_1}{\Lambda}). \tag{G.57}$$

Clearly, when $n_1 \geq n - n_1$, and $\nu_1 \leq \nu_2$, the RHS of the above inequality is strictly negative. This implies that the $t(b)$ is such that $\nu_1 \geq \nu_2$ which is equivalent to the condition $t(b)/T \geq n_1/n$.

Notice that when workforce is inflexible, $n_1 = 1$ and $n = 2$. Therefore, $t(b)/T \geq 0.5$.

H Performance metrics for systems with finished goods inventories

We are concerned here with the analysis of two systems that maintain finished goods inventories to achieve quick response. In system 1, shown in Figure 6, all items are differentiated and therefore, we need a separate buffer for each of the M different types of products. In contrast, the point of differentiation has been pushed to the retailer level in system 2, as shown in Figure 7. Here, we have a common buffer and items are customized by the retailer. This can be accomplished, for example, by performing easy-to-do insertions or deletions of modular components. Such a practice is common in personal computer industry.

In both the systems described above, the two production stages behave like M/M/1 queues in tandem. Therefore, $\pi(r)$, the probability that there are a total of r jobs in process is

$$\begin{aligned}\pi(r) &= \sum_{r_1=0}^{r} \pi_1(r_1)\pi_2(r-r_1) \\ &= \begin{cases} (r+1)(1-\rho)^2\rho^r & \text{if } \rho_1=\rho_2=\rho, \\ \frac{(1-\rho_1)(1-\rho_2)[1-(\rho_1/\rho_2)^{r+1}]\rho_2^r}{(1-\rho_1/\rho_2)} & \text{otherwise.} \end{cases}\end{aligned} \tag{H.58}$$

Consider the size of the average inventory for type j finished goods. The inventory buffer is non-empty only if the number of type j jobs in process is at most b. Thus, by definition, the average inventory of type j finished items can be written as:

$$\bar{I}_{1,j}(b,t) = \sum_{k_j=0}^{b}\sum_{r=k_j}^{\infty}(b-k_j)p_{k_j,r}\pi(r), \tag{H.59}$$

where $p_{k_j,r}$ is the conditional probability that k_j jobs are present at stages 1 and 2, given that there are a total of r jobs. Furthermore, owing to the equal arrival rates assumption, we have

$$p_{k_j,r} = \frac{r!}{k_j!(r-k_j)!}\left(\frac{1}{M}\right)^{k_j}\left(\frac{M-1}{M}\right)^{r-k_j}. \tag{H.60}$$

Due to symmetry, the total average inventory, $\bar{I}_1(Mb,t)$, is simply M times $\bar{I}_{1,j}(b,t)$. Upon substituting from H.58 and H.60 into H.59, and simplifying, we obtain the expression for $\bar{I}_1(Mb,t)$ shown in 60.

Next, we derive an expression for the order delay experienced by an arbitrary type j product demand (tagged customer) arriving to the system depicted in Figure 6. It experiences delay only if the total number of type j jobs in the two production stages exceeds b. Let there be $k \geq b$ type j jobs in the system at the moment of arrival of the tagged customer. Notice that we have dropped the subscript j on account of symmetry. Then, the tagged arrival will experience a delay which equals $(k-b+1)+Y$ service completions, where Y is a random variable with support $[0, r-k]$. It represents all type ℓ, $\ell \neq j$, jobs that will have to be processed, on account of first-in-first-out service discipline, until the $(k-b+1)^{\text{th}}$ type j job is processed. The latter will be used to satisfy the tagged customer. Then, using combinatorial arguments, the probability that $Y=y$, denoted by p_y, can be written as follows:

$$p_y = \left(\frac{(k-b+y)!(r+b-k-y-1)!}{y!(k-b)!(b-1)!(r-k-y)!}\right)\left(\frac{k!(r-k)!}{r!}\right) \tag{H.61}$$

Let $\bar{D}(r)$ denote the average inter-departure time from the production system when there are r jobs in it. Since arrivals are Poisson, we have:

$$\bar{D}(r) = \begin{cases} T-t & \text{if } Q_2>0, \\ (T-t)+t & \text{otherwise.} \end{cases} \tag{H.62}$$

From earlier arguments, the conditional probability that $Q_2 = 0$, given that $Q_1 + Q_2 = r$, denoted by $\pi_{r,0}$, can be written as follows:

$$\pi_{r,0} = \begin{cases} \frac{1}{r+1} & \text{if } \rho_1 = \rho_2 = \rho, \\ \frac{(\rho_1/\rho_2)^r[1-(\rho_1/\rho_2)]}{1-(\rho_1/\rho_2)^{r+1}} & \text{otherwise.} \end{cases} \tag{H.63}$$

Let $\bar{F}_{1,j}(b,t)$ denote the average waiting time experienced by a type j tagged customer. Then,

$$\bar{F}_{1,j} = \sum_{r=b}^{\infty} \sum_{k=b}^{r} \sum_{y=0}^{r-k} (p_{k,r})(p_y) \left(\sum_{\ell=0}^{k-b+y} \bar{D}(r-\ell) \right) \pi(r). \tag{H.64}$$

Due to symmetry, an arbitrary type j customer experiences the same average delay as an arbitrary arrival, irrespective of class, i.e., $\bar{F}_1(Mb,t) = \bar{F}_{1,j}(b,t)$. Therefore, upon substituting from H.58, H.60, H.61, and H.62 into H.64, and simplifying, we obtain 61.

The derivation of expressions 62 and 63 is similar. In fact, it is somewhat easier since there is a single buffer and all items in it can satisfy any customer's demand. Hence the details of this derivation are omitted.

cost functions: $h(t) = t$, $R(t) = 10(e^{1.0t} - 1)$, $W(b) = (100)(b)$									
	$T = 0.7$			$T = 1.2$			$T = 1.8$		
α	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$
α_ℓ	11	0.962	9518.6236	92	0.772	115224.9363	283	0.543	205279.9623
α_m	2	0.567	720.5346	4	0.516	5272.5134	14	0.501	83555.9266
α_h	1	0.297	170.2217	1	0.414	1522.9592	1	0.495	73624.0921
cost function $h(t) = t$, $R(t) = 0$, $W(b) = 0$									
	$T = 0.7$			$T = 1.2$			$T = 1.8$		
α	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$
α_ℓ	11	0.962	6.0375	92	0.777	72.9238	283	0.544	231.9867
α_m	2	0.567	0.5738	4	0.602	1.5207	14	0.512	5.4944
α_h	1	0.297	0.1647	1	0.609	0.1970	1	0.506	0.0817
cost functions: $h(t) = e^{1.0t} - 1$, $R(t) = 10(e^{1.0t} - 1)$, $W(b) = (100)(b)$									
	$T = 0.7$			$T = 1.2$			$T = 1.8$		
α	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$
α_ℓ	11	0.962	17052.7300	92	0.772	955832.0274	283	0.543	4422907.9403
α_m	2	0.567	795.1134	4	0.516	6543.9691	14	0.501	141022.9045
α_h	1	0.297	175.6053	1	0.414	1594.3464	1	0.495	74430.6130
cost function $h(t) = e^{1.0t} - 1.0$, $R(t) = 0$, $W(b) = 0$									
	$T = 0.7$			$T = 1.2$			$T = 1.8$		
α	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$
α_ℓ	12	0.945	7488.7899	92	0.772	840680.6308	283	0.543	4217861.3656
α_m	3	0.516	88.9534	4	0.516	1273.0736	14	0.501	57473.2841
α_h	1	0.297	5.5484	1	0.414	71.6371	1	0.506	805.7070
cost functions: $h(t) = \ln(t + 1)$, $R(t) = 10(e^{1.0t} - 1)$, $W(b) = (100)(b)$									
	$T = 0.7$			$T = 1.2$			$T = 1.8$		
α	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$
α_ℓ	11	0.962	9517.2018	92	0.772	115203.4374	283	0.543	205209.3464
α_m	2	0.567	720.4441	4	0.516	5272.1548	14	0.501	83554.1160
α_h	1	0.297	170.2066	1	0.414	1522.9122	1	0.495	73624.0643
cost function $h(t) = \ln(t + 1.0)$, $R(t) = 0$, $W(b) = 0$									
	$T = 0.7$			$T = 1.2$			$T = 1.8$		
α	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$	b^*	t^*/T	$K(b^*, t^*)$
α_ℓ	11	0.962	4.6158	92	0.777	51.5204	283	0.544	161.7597
α_m	2	0.778	0.5038	4	0.602	1.1443	14	0.512	3.8933
α_h	1	0.297	0.1497	1	0.609	0.1479	1	0.506	0.0581

Table 1: Optimal $(b^*, t^*/T, K(b^*, t^*))$ for inflexible work force model.

cost function $h(t)=t$, $R(t)=10(e^{10t}-1)$, $W(b)=100b$, $\alpha=0.5$									
	$T=1.75$			$T=3.0$			$T=4.5$		
n_1	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$
0	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-
2	**24**	**7.1600e-01**	**2.7675e+06**	-	-	-	-	-	-
3	13	7.3023e-01	3.5482e+06	**70**	**8.4263e-01**	**9.5177e+11**	-	-	-
4	10	8.0954e-01	1.4213e+07	27	8.8890e-01	3.8136e+12	-	-	-
5	5	1.0000e+00	3.9825e+08	5	1.0000e+00	1.0686e+14	**30**	**1.0000e+00**	**3.4934e+20**
cost function $h(t)=t$, $R(t)=0$, $W(b)=0$, $\alpha=0.5$									
	$T=1.75$			$T=3.0$			$T=4.5$		
n_1	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$
0	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-
2	9	8.5314e-01	8.7648e+00	-	-	-	-	-	-
3	**3**	**1.0000e+00**	**2.1875e+00**	32	8.7417e-01	6.3265e+01	-	-	-
4	4	8.1897e-01	3.6787e+00	8	9.9997e-01	1.1866e+01	-	-	-
5	5	1.0000e+00	5.6875e+00	**5**	**1.0000e+00**	**6.0000e+00**	**30**	**1.0000e+00**	**8.6086e+01**
cost function $h(t)=e^{10t}-1$, $R(t)=10(e^{10t}-1)$, $W(b)=100b$, $\alpha=0.5$									
	$T=1.75$			$T=3.0$			$T=4.5$		
n_1	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$
0	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-
2	12	7.2063e-01	5.9724e+06	-	-	-	-	-	-
3	**6**	**7.3497e-01**	**5.6268e+06**	**37**	**8.4657e-01**	**4.3349e+12**	-	-	-
4	5	8.1274e-01	2.0374e+07	12	8.9370e-01	8.1789e+12	-	-	-
5	5	1.0000e+00	5.2768e+08	5	1.0000e+00	1.2824e+14	**30**	**1.0000e+00**	**1.0176e+21**
cost function $h(t)=e^{10t}-1.0$, $R(t)=0$, $W(b)=0$, $\alpha=0.5$									
	$T=1.75$			$T=3.0$			$T=4.5$		
n_1	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$
0	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-
2	10	7.2954e-01	2.7579e+06	-	-	-	-	-	-
3	**5**	**7.4206e-01**	**1.5688e+06**	**35**	**8.4870e-01**	**3.2415e+12**	-	-	-
4	4	8.1897e-01	4.3023e+06	10	9.0107e-01	3.6025e+12	-	-	-
5	5	1.0000e+00	1.2943e+08	5	1.0000e+00	2.1373e+13	**30**	**1.0000e+00**	**6.6830e+20**
cost function $h(t)=\ln(t+1.0)$, $\alpha=0.5$									
cost function $h(t)=\ln(1+t)$, $R(t)=10(e^{10t}-1)$, $W(b)=100b$, $\alpha=0.5$									
	$T=1.75$			$T=3.0$			$T=4.5$		
n_1	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$
0	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-
2	**24**	**7.1600e-01**	**2.7675e+06**	-	-	-	-	-	-
3	13	7.3023e-01	3.5482e+06	**70**	**8.4263e-01**	**9.5177e+11**	-	-	-
4	10	8.0954e-01	1.4213e+07	27	8.8890e-01	3.8136e+12	-	-	-
5	5	1.0000e+00	3.9825e+08	5	1.0000e+00	1.0686e+14	**30**	**1.0000e+00**	**3.4934e+20**
cost function $h(t)=\ln(t+1.0)$, $R(t)=0$, $W(b)=0$, $\alpha=0.5$									
	$T=1.75$			$T=3.0$			$T=4.5$		
n_1	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$	b^*	t^*/T	$K(b^*,n_1,t^*)$
0	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-
2	9	8.5314e-01	5.3627e+00	-	-	-	-	-	-
3	**3**	**1.0000e+00**	**1.2645e+00**	32	8.7417e-01	3.1052e+01	-	-	-
4	4	1.0000e+00	2.2761e+00	8	9.9997e-01	5.4834e+00	-	-	-
5	5	1.0000e+00	3.2877e+00	**5**	**1.0000e+00**	**2.7726e+00**	**30**	**1.0000e+00**	**3.2612e+01**

Table 2: Optimal b^*, t^*/T, and $K(b^*,n_1,t^*)$ for each n_1 with $n=5$ and $\alpha=0.5$. Overall optimal configuration is shown in bold font.

cost function $h(t) = t$, $R(t) = 0$, $W(b) = 0$, $\alpha = 1.0$									
	$T = 3.5$			$T = 6.0$			$T = 9.0$		
n_1	b^*	t^*/T	$K(b^*, n_1, t^*)$	b^*	t^*/T	$K(b^*, n_1, t^*)$	b^*	t^*/T	$K(b^*, n_1, t^*)$
0	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-
3	30	7.3997e-01	5.8904e+01	-	-	-	-	-	-
4	8	8.6334e-01	1.1779e+01	-	-	-	-	-	-
5	**5**	**9.9997e-01**	**5.2502e+00**	-	-	-	-	-	-
6	6	9.9997e-01	8.7501e+00	28	8.7365e-01	9.5988e+01	-	-	-
7	7	7.2780e-01	1.1342e+01	12	9.1112e-01	2.9536e+01	-	-	-
8	8	7.6389e-01	1.4241e+01	**9**	**1.0000e+00**	**1.6394e+01**	-	-	-
9	9	8.5731e-01	1.8002e+01	9	1.0000e+00	1.8000e+01	96	9.5416e-0 1	6.0168e+02
10	10	1.0000e+00	2.2750e+01	10	1.0000e+00	2.4000e+01	**28**	**1.0000e+00**	**1.2496e+02**
cost function $h(t) = e^{10t} - 1.0$, $R(t) = 0$, $W(b) = 0$, $\alpha = 1.0$									
	$T = 3.5$			$T = 6.0$			$T = 9.0$		
n_1	b^*	t^*/T	$K(b^*, n_1, t^*)$	b^*	t^*/T	$K(b^*, n_1, t^*)$	b^*	t^*/T	$K(b^*, n_1, t^*)$
0	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-
3	34	7.1880e-01	2.3448e+12	-	-	-	-	-	-
4	11	7.2066e-01	7.1464e+11	-	-	-	-	-	-
5	7	7.2491e-01	4.5520e+11	-	-	-	-	-	-
6	**6**	**7.2660e-01**	**3.8303e+11**	35	8.3732e-01	1.7729e+23	-	-	-
7	7	7.2780e-01	5.1453e+11	**18**	**8.4342e-01**	**1.1470e+23**	-	-	-
8	8	7.6389e-01	2.1764e+12	13	8.6615e-01	2.7777e+23	-	-	-
9	9	8.5731e-01	6.4498e+13	10	9.2253e-01	4.7950e+24	**107**	**9.4631e-01**	**8.1751e+38**
10	10	1.0000e+00	1.0309e+16	10	1.0000e+00	4.5680e+26	28	1.0000 e+00	1.6945e+40
cost function $h(t) = \ln(t + 1.0)$, $R(t) = 0$, $W(b) = 0$, $\alpha = 1.0$									
	$T = 3.5$			$T = 6.0$			$T = 9.0$		
n_1	b^*	t^*/T	$K(b^*, n_1, t^*)$	b^*	t^*/T	$K(b^*, n_1, t^*)$	b^*	t^*/T	$K(b^*, n_1, t^*)$
0	-	-	-	-	-	-	-	-	-
1	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-
3	30	7.3997e-01	2.9070e+01	-	-	-	-	-	-
4	8	8.6334e-01	5.4251e+00	-	-	-	-	-	-
5	**5**	**9.9997e-01**	**2.2562e+00**	-	-	-	-	-	-
6	6	9.9997e-01	3.7603e+00	28	8.7365e-01	3.3534e+01	-	-	-
7	7	9.9997e-01	5.2643e+00	12	9.1112e-01	1.0085e+01	-	-	-
8	8	9.9997e-01	6.7684e+00	**9**	**1.0000e+00**	**5.3167e+00**	-	-	-
9	9	9.9997e-01	8.2725e+00	9	1.0000e+00	5.8377e+00	96	9.5416e-0 1	1.5838e+02
10	10	1.0000e+00	9.7765e+00	10	1.0000e+00	7.7836e+00	**28**	**1.0000e+00**	**3.1971e+01**

Table 3: Optimal b^*, t^*/T, and $K(b^*, n_1, t^*)$ for each n_1 with $n = 10$ and $\alpha = 1.0$. Overall optimal configuration is shown in bold font.

$h(t)=t,\ \alpha=1$													
M	T	*System 1*			*System 2*				*System 3*				
		b_1	I_1	$h(T)I_1$	b_2	I_2	$h(T)I_2$	Δ_1	b_3	t_3^*	I_3	$h(T)I_3$	Δ_2
2	0.80	5	3.93	3.15	10	3.533	2.827	10.11	90	0.471	73.552	34.678	-1126.68
3	0.53	3	3.45	1.85	7	2.094	1.117	39.46	23	0.293	16.049	4.708	-321.51
4	0.40	2	3.00	1.20	6	1.625	0.650	45.83	11	0.212	6.342	1.344	-106.74
5	0.32	1	1.54	0.49	4	0.791	0.253	48.79	6	0.166	2.732	0.453	-78.94
6	0.27	1	2.16	0.57	3	0.463	0.124	78.40	4	0.136	1.531	0.208	-67.92
7	0.23	1	2.83	0.65	2	0.218	0.050	92.25	2	0.115	0.540	0.062	-24.70
8	0.20	1	3.56	0.71	1	0.064	0.013	98.17	1	0.100	0.200	0.020	-53.85
9	0.18	1	4.31	0.77	1	0.064	0.011	98.56	0	0.082	0.000	0.000	100.00
10	0.16	1	5.10	0.82	0	0.000	0.000	100.00	0	0.072	0.000	0.000	
$h(t)=e^{10t}-1,\ \alpha=1$													
M	T	*System 1*			*System 2*				*System 3*				
		b_1	I_1	$h(T)I_1$	b_2	I_2	$h(T)I_2$	Δ_1	b_3	t_3^*	I_3	$h(T)I_3$	Δ_2
2	0.80	5	3.93	11715.00	10	3.533	10528.889	10.12	90	0.471	73.965	8121.220	22.87
3	0.53	3	3.45	711.62	7	2.094	431.717	39.33	23	0.292	16.273	285.426	33.89
4	0.40	2	3.00	160.79	6	1.625	87.079	45.84	11	0.211	6.444	46.614	46.47
5	0.32	1	1.54	36.32	4	0.791	18.615	48.74	6	0.165	2.771	11.660	37.36
6	0.27	1	2.16	28.93	3	0.463	6.205	78.55	4	0.136	1.546	4.455	28.20
7	0.23	1	2.83	25.04	2	0.218	1.922	92.32	2	0.115	0.543	1.176	38.82
8	0.20	1	3.56	22.72	1	0.064	0.409	98.20	1	0.100	0.200	0.344	15.98
9	0.18	1	4.31	21.21	1	0.064	0.315	98.51	0	0.082	0.000	0.000	100.00
10	0.16	1	5.10	20.17	0	0.000	0.000	100.00	0	0.072	0.000	0.000	
$h(t)=\ln(1+t),\ \alpha=1$													
M	T	*System 1*			*System 2*				*System 3*				
		b_1	I_1	$h(T)I_1$	b_2	I_2	$h(T)I_2$	Δ_1	b_3	t_3^*	I_3	$h(T)I_3$	Δ_2
2	0.80	5	3.93	2.31	10	3.5330	2.0770	10.13	90	0.472	73.531	28.404	-1267.57
3	0.53	3	3.45	1.48	7	2.0940	0.8950	39.49	23	0.293	16.035	4.126	-361.01
4	0.40	2	3.00	1.01	6	1.6250	0.5470	45.79	11	0.212	6.338	1.218	-122.72
5	0.32	1	1.54	0.43	4	0.7910	0.2200	48.60	6	0.166	2.732	0.419	-90.37
6	0.27	1	2.16	0.51	3	0.4630	0.1100	78.39	4	0.136	1.530	0.195	-77.39
7	0.23	1	2.83	0.58	2	0.2180	0.0450	92.24	2	0.115	0.540	0.059	-31.09
8	0.20	1	3.56	0.65	1	0.0640	0.0120	98.15	1	0.100	0.200	0.019	-58.83
9	0.18	1	4.31	0.70	1	0.0640	0.0100	98.58	0	0.082	0.000	0.000	100.00
10	0.16	1	5.10	0.76	0	0.0000	0.0002	99.97	0	0.072	0.000	0.000	100.00

Table 4: Comparison of systems 1 through 3 for different values of M when $\alpha = 1.0$. $\Delta_1 = h(T)\{\bar{I}_1 - \bar{I}_2\} \times 100/h(T)\bar{I}_1$ denotes the inventory cost savings of system 2 over system 1. $\Delta_2 = \{h(T)\bar{I}_2 - h(t_3^*)\bar{I}_3\} \times 100/h(T)\bar{I}_2$ denotes the inventory cost savings of system 3 over system 2. In order to maintain $M\lambda T$ constant as M increases, T is decreased by a proportional amount.

$h(t)=t,\ \alpha=1$													
M	T	*System 1*			*System 2*				*System 3*				
		b_1	I_1	$h(T)I_1$	b_2	I_2	$h(T)I_2$	Δ_1	b_3	t_3^*	I_3	$h(T)I_3$	Δ_2
2	0.80	5	3.93	3.15	10	3.533	2.827	10.13	90	0.471	73.552	34.678	-1126.86
3	0.80	4	5.58	4.47	10	3.533	2.827	36.73	90	0.471	73.552	34.678	-1126.86
4	0.80	3	5.75	4.60	10	3.533	2.827	38.55	90	0.471	73.552	34.678	-1126.86
5	0.80	3	8.29	6.63	10	3.533	2.827	57.37	90	0.471	73.552	34.678	-1126.86
6	0.80	3	10.97	8.78	10	3.533	2.827	67.80	90	0.471	73.552	34.678	-1126.86
7	0.80	2	7.73	6.18	10	3.533	2.827	54.30	90	0.471	73.552	34.678	-1126.86
8	0.80	2	9.48	7.59	10	3.533	2.827	62.74	90	0.471	73.552	34.678	-1126.86
9	0.80	2	11.28	9.03	10	3.533	2.827	68.68	90	0.471	73.552	34.678	-1126.86
10	0.80	2	13.12	10.50	10	3.533	2.827	73.07	90	0.471	73.552	34.678	-1126.86
$h(t)=e^{10t}-1,\ \alpha=1$													
M	T	*System 1*			*System 2*				*System 3*				
		b_1	I_1	$h(T)I_1$	b_2	I_2	$h(T)I_2$	Δ_1	b_3	t_3^*	I_3	$h(T)I_3$	Δ_2
2	0.80	5	3.93	11715.45	10	3.533	10528.889	10.13	90	0.471	73.965	8121.220	22.87
3	0.80	4	5.58	16640.39	10	3.533	10528.889	36.73	90	0.471	73.965	8121.220	22.87
4	0.80	3	5.75	17134.76	10	3.533	10528.889	38.55	90	0.471	73.965	8121.220	22.87
5	0.80	3	8.29	24696.73	10	3.533	10528.889	57.37	90	0.471	73.965	8121.220	22.87
6	0.80	3	10.97	32698.48	10	3.533	10528.889	67.80	90	0.471	73.965	8121.220	22.87
7	0.80	2	7.73	23038.14	10	3.533	10528.889	54.30	90	0.471	73.965	8121.220	22.87
8	0.80	2	9.48	28254.42	10	3.533	10528.889	62.74	90	0.471	73.965	8121.220	22.87
9	0.80	2	11.28	33619.13	10	3.533	10528.889	68.68	90	0.471	73.965	8121.220	22.87
10	0.80	2	13.12	39095.66	10	3.533	10528.889	73.07	90	0.471	73.965	8121.220	22.87
$h(t)=\ln(1+t),\ \alpha=1$													
M	T	*System 1*			*System 2*				*System 3*				
		b_1	I_1	$h(T)I_1$	b_2	I_2	$h(T)I_2$	Δ_1	b_3	t_3^*	I_3	$h(T)I_3$	Δ_2
2	0.80	5	3.93	2.31	10	3.533	2.077	10.13	90	0.472	73.531	28.404	-1267.70
3	0.80	4	5.58	3.28	10	3.533	2.077	36.73	90	0.472	73.531	28.404	-1267.70
4	0.80	3	5.75	3.38	10	3.533	2.077	38.55	90	0.472	73.531	28.404	-1267.70
5	0.80	3	8.29	4.87	10	3.533	2.077	57.37	90	0.472	73.531	28.404	-1267.70
6	0.80	3	10.97	6.45	10	3.533	2.077	67.80	90	0.472	73.531	28.404	-1267.70
7	0.80	2	7.73	4.54	10	3.533	2.077	54.30	90	0.472	73.531	28.404	-1267.70
8	0.80	2	9.48	5.57	10	3.533	2.077	62.74	90	0.472	73.531	28.404	-1267.70
9	0.80	2	11.28	6.63	10	3.533	2.077	68.68	90	0.472	73.531	28.404	-1267.70
10	0.80	2	13.12	7.71	10	3.533	2.077	73.07	90	0.472	73.531	28.404	-1267.70

Table 5: Comparison of systems 1 through 3 for different values of M when $\alpha = 1.0$. $\Delta_1 = h(T)\{\bar{I}_1 - \bar{I}_2\} \times 100/h(T)\bar{I}_1$ denotes the inventory cost savings of system 2 over system 1. $\Delta_2 = \{h(T)\bar{I}_2 - h(t_3^*)\bar{I}_3\} \times 100/h(T)\bar{I}_2$ denotes the inventory cost savings of system 3 over system 2. In order to maintain $M\lambda T$ constant as M increases, λ is decreased by a proportional amount.

Faculty of Business
McMaster University

WORKING PAPERS - RECENT RELEASES

407. Ali R. Montazemi and Kalyan Moy Gupta, "A Framework for Retrieval in Case-Based Reasoning Systems", June, 1995.

408. Ali R. Montazemi and Kalyan Moy Gupta, "An Adaptive Agent for Case Description in Diagnostic CBR Systems", June, 1995.

409. Roy J. Adams, Noel Cowell and Gangaram Singh, "The Making of Industrial Relations in the Commonwealth Caribbean", June, 1995.

410. Jiang Chen and George Steiner, "Approximation Methods for Discrete Lot Streamingin Flow Shops", June, 1995.

411. Harish C. Jain and S. Muthuchidambaram, "Bill 40 Amendments to Ontario Labour Relations Act: An Overview and Evaluation", June, 1995.

412. Jiang Chan and George Steiner, "Discrete Lot Streaming in Three-Machine Flow Shops", July, 1995.

413. J. Brimberg, A. Mehrez and G.O. Wesolowsky, "Allocation of Queueing Facilities Using a Minimax Criterion", January, 1996.

414. Isik Zeytinoglu and Jeanne Norris, "Global Diversity in Employment Relationships: A Typology of Flexible Employment", March, 1996.

415. N. Archer, "Characterizing World Wide Web Search Strategies", April, 1996.

416. J. Rose, "Immediacy and Saliency in Remedying Employer Opposition to Union Organizing Campaigns", July, 1996.

417. Roy J. Adams and Parbudyal Singh, "Worker Rights Under NAFTA: Experience With the North American Agreement on Labor Cooperation", September, 1996.

418. George Steiner and Paul Stephenson, "Subset-Restricted Interchange for Dynamic Min-Max Scheduling Problems", September, 1996.

419. Robert F. Love and Halit Uster, "Comparison of the Properties and the Performance of the Criteria Used to Evaluate the Accuracy of Distance Predicting Functions", November, 1996.

420. Harish C. Jain and Simon Taggar, "The Status of Employment Equity in Canada", December, 1996.

421. Harish C. Jain and Parbudyal Singh, "Beyond The Rhetoric: An Assessment of the Political Arguments and Legal Principles on Strike Replacement Laws in North America", January, 1997.

422. Jason Schwandt, "Electronic Data Interchange: An Overview of Its Origins, Status, and Future", March, 1997.

423. Christopher K. Bart with John C. Tabone, "Mission Statement Rationales and Organizational Alignment in the Not-for-profit Healthcare Sector", November, 1997.

424. Harish C. Jain, Michael Piczak, Işik Urla Zeytinoğlu, "Workplace Substance Testing - An Exploratory Study", November, 1997.

425. S. Suarga, Yufei Yuan, Joseph B. Rose, and Norman Archer, "Web-based Collective Bargaining Support System: A Valid Process Support Tool for Remote Negotiation", January, 1998.

426. Pawan S. Budhwar and Harish C. Jain, "Evaluating Levels of Strategic Integration and Development of Human Resource Management in Britain", March, 1998.

427. Halit Üster and Robert F. Love, "Application of Weighted Sums of Order p to Distance Estimation", April, 1998.

428. Halit Üster and Robert F. Love, "On the Directional Bias of the ℓ_{bp}-norm", April, 1998.

429. Milena Head, Norm Archer, and Yufei Yuan, "MEMOS: A World Wide Web Navigation Aid", October, 1998.

430. Harish C. Jain and Parbudyal Singh, "The Effects of the Use of Strike Replacement Workers on Strike Duration in Canada", February, 1999.

431. Parbudyal Singh and Harish C. Jain, "Strike Replacements in the United States, Canada and Mexico: A Review of the Law and Empirical Research", February, 1999.

432. John W. Medcof and Jeremy Boyko, "Reinforcing, Revising and Reconciling Attributions in the Employment Interview", March, 1999.

433. Norm Archer, "World Wide Web Business Catalogs in Business-to-Business Procurement", March, 1999.