Managing Capacity Flexibility in Make-to-Order Production Environments

Fehmi Tanrisever [a], Douglas Morrice [b], David Morton [c]

[a] *School of Industrial Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands*
[b] *Red McCombs School of Business, The University of Texas at Austin, Austin, TX 78712*
[c] *Graduate Program in Operations Research and Industrial Engineering, The University of Texas at Austin, Austin, TX 78712*

Abstract: This paper addresses the problem of managing flexible production capacity in a make-to-order (MTO) manufacturing environment. We present a multi-period capacity management model where we distinguish between process flexibility (the ability to produce multiple products on multiple production lines) and operational flexibility (the ability to dynamically change capacity allocations among different product families over time). For operational flexibility, we consider two polices: a fixed allocation policy where the capacity allocations are fixed throughout the planning horizon and a dynamic allocation policy where the capacity allocations change from period to period. The former approach is modeled as a single-stage stochastic program and solved using a cutting-plane method. The latter approach is modeled as a multi-stage stochastic program and a sampling-based decomposition method is presented to identify a feasible policy and assess the quality of that policy. A computational experiment quantifies the benefits of operational flexibility and demonstrates that it is most beneficial when the demand and capacity is well-balanced and the demand variability is high. Additionally, our results reveal that myopic operating policies may lead a firm to adopt more process flexibility and form denser flexibility configuration chains. That is, process flexibility may be over-valued in the literature since it is assumed that a firm will operate optimally after the process flexibility decision. We also show that the value of process flexibility increases with the number of periods in the planning horizon if an optimal operating policy is employed. This result is reversed if a myopic allocation policy is adopted instead.

Keywords
Supply Chain Management, Capacity Flexibility, Stochastic Programming, Make-to-Order, Assignment

1. INTRODUCTION AND PROBLEM DEFINITION

As the competition in high-tech markets becomes more and more intense, product differentiation and customization become a top priority for the many companies. For instance, today, most companies in the computer manufacturing industry allow their customers to customize nearly every component of their products. While product customization is a must for strategic competition in these markets, increased levels of customization also come with their own operational-level challenges.

This paper studies such an operational challenge recently faced by a high-tech make-to-order manufacturing firm: Managing multiple flexible production lines to produce multiple product families so as to minimize the total operating cost (including the cost of managing

process flexibility and the backlogged demand), over multiple production periods where the demand for the products is highly uncertain.

The firm which motivated this research is a manufacturer of electronic devices that consist of a single chassis and a set of parts assembled on it. Products are grouped into families depending on the chassis that they are built onto and each family requires a different set of parts. While this work was motivated by a firm in the electronics industry, many of the same issues studied here are also faced by make-to-order manufacturing firms in other industries.

On the demand side, customers are allowed to choose almost every part of their products. In particular, a customer order includes a selection of chassis type and a set of parts that are available for that chassis. Since the number of possible product configurations that can be formed by the customers is large, it is possible to start the final assembly of a product only after a firm customer order is received. On the supply side, customer orders are produced on multiple production lines, which may be adjusted to manufacture any set of product families prior to the start of production. The adjustments are time consuming and costly; hence it is not practical to change them once the production is started. The same set of assignments is preserved over multiple production periods, until a significant change in the demand pattern is observed. If the firm is short of capacity in one period, then excess demand is backlogged and carried over to the next period. Since the customers are placing orders for highly customized products, they are usually willing to wait for their orders. Cancelling an order, in case of a delay, is not very desirable for the customers since there is no other competitor with which they can place the same order and receive it immediately.

Prior to the start of production, the firm decides a product-to-line assignment, which we refer to as the *process flexibility* of the firm. Process flexibility refers to the ability of a firm to produce multiple products on multiple production facilities or lines, as described by the process-flexibility literature (see Jordan and Graves 1995). As greater process flexibility is adopted by the firm, i.e., as more products are assigned to more lines, the firm's ability to match capacity with demand improves. However, process flexibility comes at a cost. In particular, assigning product i to line j involves a certain cost depending on i and j due to: (1) pre-positioning the related parts and chassis inventory next to the production line, (2) computer programming and setup, which are time consuming, and (3) dedicating labor and material handling equipment to produce family i on line j during the planning horizon, which increases direct manufacturing

expenses. Hence, in our model, process flexibility is a tactical level decision, which can only be revised in response to major changes in the demand pattern. The capacity investment decisions are, however, fixed in the medium- and short-term.

Once the process flexibility decision is made, operating the system by allocating capacity to demand is another practical challenge in a multi-period planning horizon. In particular, the *operational flexibility* of the firm, i.e., the ability to dynamically change capacity allocations among different product families over time, plays a critical role in the selection of capacity allocations. Further, operating decisions also affect the choice of process flexibility *ex ante.*

Regarding the operational flexibility of the firm, we consider two basic modeling approaches: (1) a Dynamic Allocation Model (DAM), where the allocation decisions are made after observing the demand at the beginning of each production period and (2) a Fixed Allocation Model (FAM), where the allocation decisions are made at the beginning of the planning horizon together with the assignment decisions and these decisions do not change in response to demand realizations from period to period.

The sequence of decisions for our firm is shown in Figure 1. First, based on the forecasted demand, the firm commits to a process flexibility configuration prior to the start of production and incurs a certain flexibility cost. Next, at the beginning of every production period t, demand is realized and the production capacity is allocated to meet that demand, and the existing backlog, subject to the process flexibility configuration and the operational flexibility of the firm. Unmet demand from period t is backlogged. The overall objective (under both DAM and FAM) is to minimize the total operating cost over the planning horizon, which includes the cost of process flexibility and the expected cost of total backlog.

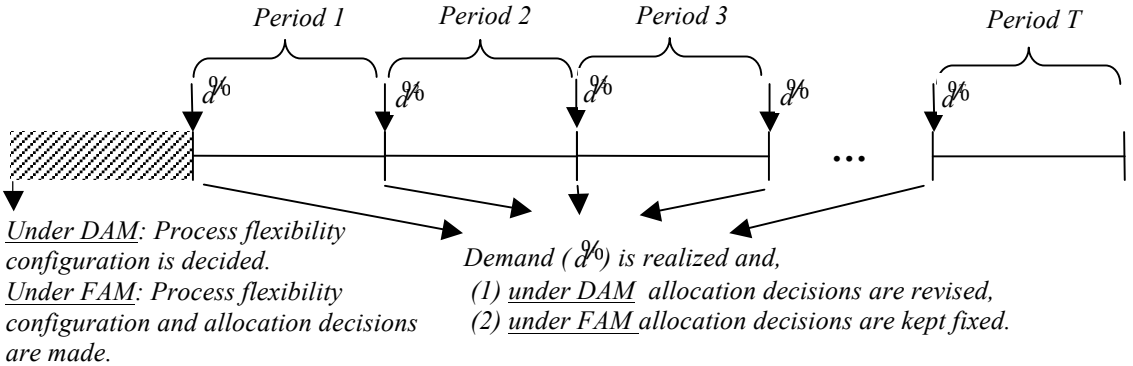

Figure 1: A Graphical Representation of DAM and FAM

As the sequence of decisions suggests, we model DAM as a multi-stage stochastic integer program with binary decisions only in the first stage and FAM is modeled as a single stage stochastic integer program. We also provide effective procedures to solve our mathematical models. Regarding our solution methods, the solution methodology developed for FAM handles non-identical and correlated demand both across time and product families. We require demand to be independent across time when solving DAM, but it need not be identically distributed and we can handle inter-product dependencies. Assuming independence across time is reasonable for a make-to-order firm involved in mass customization facing an aggregate demand that comes from a large number of customers who act independently.

Note that FAM has no operational flexibility since each line is allocated a fixed time to produce a certain family, while DAM has full operational flexibility. Fixing allocation decisions may have significant operational benefits including: reduced scheduling problems, operational standardization and increased efficiency (Li and Tirupati 1997). However, in our setting, quantifying these benefits is not straightforward since it is not easy to incorporate them in a mathematical decision model. In practice, our firm employs an operating policy that is close to FAM (allocations are rarely changed in response to demand). So, in this paper, FAM serves as a benchmark to evaluate the potential benefits of operational flexibility observed under DAM.

We provide two sets of computational analyses. First, we quantify the potential benefits of operational flexibility by comparing the performance of DAM and FAM. These models simultaneously optimize for process flexibility and the capacity allocation decisions. Second, we investigate the value of process flexibility in a multi-period production framework under different dynamic operating policies. For this purpose we introduce the myopic version of DAM as a third operating model (MDAM) where the firm may change the allocations at the beginning of each period, but does so without taking the impact on future periods into account. By comparing the value of process flexibility under DAM and MDAM, we show that process flexibility may not only be used to hedge against the demand uncertainty, but may also be employed to protect against possible suboptimal operating decisions in the future.

The rest of the paper is configured as follows: In §2, we provide a brief review of the related literature and outline our contributions. In §3, FAM is explained in detail and an effective solution algorithm is presented. §4 explains the DAM and presents a sampling-based decomposition method to find a near-optimal solution. §5.1 presents a computational study of the

benefits of operational flexibility by comparing the performance of FAM and DAM. §5.2 is dedicated to the analysis of the value of process flexibility and operating policies. We conclude with a brief discussion of results and future research directions in §6.

2. LITERATURE REVIEW

Our paper is most closely related to the capacity-flexibility literature. The literature on capacity flexibility is extensive, but the related literature can be categorized in two main streams. The first stream focuses on investment in resources that are dedicated versus totally flexible. The second stream explores process flexibility, i.e., the ability of a firm to produce multiple products on multiple production facilities or lines. The former stream includes Fine and Freund (1990), Van Mieghem (1998), Li and Tirupati (1994, 1995, 1997) and Van Mieghem and Rudi (2002).

Fine and Freund (1990) consider a firm which has the option to invest in both product-dedicated capacity and flexible capacity. They provide a two-stage stochastic program in which the capacity investment decisions are made in the first stage and the production decisions are made in the second stage. They investigate the cost-benefit tradeoff of flexible capacity. Van Mieghem (1998) shows the impact of price- and cost-mix differentials when investing in flexible and dedicated capacity. Like Fine and Freund (1990), he considers a model in which investment decisions are made at the beginning of the time horizon. Van Mieghem and Rudi (2002) extend Van Mieghem (1998) to a dynamic setting with multiple products, and multiple processing and storage points, which is called a newsvendor network.

Li and Tirupati (1994) address investment in flexible and non-flexible technology in a multi-period problem with deterministic demand. Their objective is to minimize the the cost of technology investment and the operating cost over the planning horizon. Li and Tirupati (1995) consider the same problem with two products and stochastic demand. However, in this setting unlike Van Mieghem and Fine and Freund, demand uncertainty is addressed by specifying a target service level. Li and Tirupati (1997) extend their previous two papers by explicitly considering two kinds of operating policies, which refer to the allocation of flexible capacity among different products arising from: (1) a Static Allocation Model (SAM) and (2) a Dynamic Allocation Model (DAM). In SAM, the allocation of flexible capacity to product lines is made at the beginning of the planning horizon. In DAM they permit dynamic allocations of flexible capacity in each period after demand realizations are observed. In both models the objective is to minimize the investment cost subject to service level constraints. For SAM, an exact exponential

time algorithm is provided. Assuming a proportional allocation rule, the DAM is approximated with a single period model and a heuristic is given to generate good solutions for special cases.

The literature mentioned above is similar to our paper in the sense that the cost and benefits of flexibility are considered explicitly. However, they only consider two types of capacity: fully flexible and non-flexible capacity. In our case, the capacity (i.e., the assembly lines) can be adjusted to any intermediate level of flexibility at a certain cost. In addition, except for Li and Tirupati (1994) who model demand as being deterministic, the above papers only consider single period models, which may not be sufficiently realistic for many practical situations as mentioned by Van Mieghem (1998). With its multi-period structure and stochastic demand our models are arguably more realistic for practical purposes.

The second stream of capacity-flexibility literature originates with Jordan and Graves (1995), (referred to as J&G) allows for choosing among resources with an intermediate level of flexibility. J&G focus on process flexibility. They show that: (1) limited flexibility (each plant builds only a few products) can achieve almost all the benefits of total flexibility (each plant builds all the products), and (2) limited flexibility should be configured to chain products and plants together as much as possible. A main focus of their paper is a measure to quantify the benefits of the given product-plant configuration, and they use this measure to guide the search for a good limited-flexibility configuration. While a configuration, which yields almost all the benefits of total flexibility is identified, the authors do not explicitly study associated cost trade-offs. We explicitly model the cost of process flexibility, and in this case, the flexibility measure of J&G cannot be used to guide a search for a good configuration due to the combinatorial nature of the problem. Additionally, J&G assume that the demand uncertainty is revealed at a single time point, i.e., immediately after the flexibility configuration decision. Our model addresses this restriction by considering a multi-period model in which the unmet demand is backlogged at the end of each period.

The results of J&G are based on the assumption that the firm optimally allocates capacity after the process flexibility decision has been made. We reconsider this issue in a multi-period framework and study the manner in which the value of process flexibility depends on the operating policies employed. Indeed, we show that a myopic operating policy (commonly practiced) may significantly reduce the value of a process flexibility configuration and increase the need for more process flexibility. We note that Bish et al. (2005) also consider the impact of

allocation policies on system performance in a two-product, two-firm case with lost sales under fully flexible and dedicated manufacturing settings.

The work by Graves and Tomlin (2003) extends the chaining ideas of J&G to multi-stage supply chains. In contrast to the above work, Garavelli (2003) considers the logistics aspect of process flexibility. Further, Gurumurthi and Benjaafar (2004) show the effectiveness of chaining in queueing systems under varying control policies. Worker cross-training and skill chaining are also studied in the queueing literature by Hopp et al. (2004) and Iravani et al. (2007).

Finally, the DAM can be interpreted as a risk mitigating strategy to reduce the mismatch between supply and demand. Hence, our paper is also linked to the operational hedging literature (e.g., Huchzermeier and Cohen 1996, Van Mieghem 2003 and Chod et al. 2010). See Boyabatli and Toktay (2004) for a recent review on this topic.

3. FIXED ALLOCATION MODEL

In this section, we develop and analyze the fixed allocation model (FAM), which closely reflects current practice at the firm. In this model, both the product-to-line assignments and the capacity allocations are decided before the production starts. Then, the allocation decisions as well as the assignment decisions are kept fixed throughout the planning horizon. The objective is to minimize the total assignment and expected backlogging costs. Details together with a list of notation are presented below:

***Indices*:**

i, M = i indexes the product families, which total M in number
j, N = j indexes the production lines, which total N in number
t, T = t indexes time periods, which total T in number
k = k indexes demand realizations for period t

***Data*:**

K_j = capacity of production line j, per period (in time units)
e_{ij} = amount of time needed to produce one unit of family i on line j
a_{ij} = assignment/flexibility cost incurred to produce family i on line j
s_i = per unit per period backlogging cost for family i
$c_i(.,.)$ = backlogging cost function for family i (defined below)
$\tilde{d}_i^t$ = random demand for product family i in period t
$\tilde{d}^t$ = $(\tilde{d}_1^t,...,\tilde{d}_M^t)$: vector of product family demands in period t
$\tilde{d}_i$ = $(\tilde{d}_i^1,...,\tilde{d}_i^T)$: demand for family i from period 1 to T (notation d without "~" refers to a general demand realization)
$d^{t,k}$ = a particular realization of demand vector in period t
$d_i^{t,k}$ = a particular realization of demand for family i in period t

***Decision variables*:**

y_{ii} = capacity of line j allocated to produce family i, in production units (allocation decisions)
x_{ij} = 1 if product family i is assigned to line j; 0 otherwise (assignment decisions)

Fixed Allocation Model (FAM):

$$z_f = \min_{x,y} \sum_{j=1}^{N} \sum_{i=1}^{M} a_{ij} x_{ij} + \sum_{i=1}^{M} s_i Ec_i(\sum_{j=1}^{N} y_{ij}, \tilde{d}_i) \quad (1)$$

$$\text{s.t.} \quad \sum_{i=1}^{M} e_{ij} y_{ij} \le K_j \qquad \forall j \quad (2)$$

$$y_{ij} \le (K_j / e_{ij}) x_{ij} \qquad \forall i, j \quad (3)$$

$$x_{ij} \in \{0,1\}, y_{ij} \ge 0 \qquad \forall i, j \quad (4)$$

We use x and y to denote the vectors whose components are x_{ij} and y_{ij}, respectively.

The first term in the objective function is the assignment cost and the second term is the expected total backlogging cost over the planning horizon. The first set of constraints, (2), limits the capacity of each line j to K_j. Constraints in (3) are the design constraints, which allow a line to produce only the products that are assigned to it. We develop the backlogging function as follows. Let $b_i^1(y_i, d_i^1) = [d_i^1 - y_i]^+$ be the backlog for family i, in period $t = 1$, where $y_i = \sum_{j=1}^{N} y_{ij}$ and where $[.]^+$ is the larger of its argument and 0. Then, for $t = 2,\ldots,T$ the backlog is recursively defined as:

$$b_i^t(y_i, d_i^1, ..., d_i^t) = [b_i^{t-1}(y_i, d_i^1, ..., d_i^{t-1}) + d_i^t - y_i]^+ . \quad (5)$$

Finally, $c_i(y_i, d_i) \equiv \sum_{t=1}^{T} b_i^t(y_i, d_i^1, ..., d_i^t)$. The following proposition characterizes convexity of the associated expected backlogging cost. (See Appendix A for the proofs of all propositions.)

Proposition 1: Assume $\tilde{d}_i, i = 1,...,M,$ have finite mean and that $s_i \ge 0, i = 1,...,M.$ Then, the expected total backlogging cost, i.e., $f(y) = \sum_{i=1}^{M} s_i Ec_i(\sum_{j=1}^{N} y_{ij}, \tilde{d}_i)$, is convex.

The expectations in the objective function, (1), are with respect to the joint distributions of $\tilde{d}_i = (\tilde{d}_i^1, ..., \tilde{d}_i^T), i = 1,...,M.$ If each $\tilde{d}_i$ has a modest number of realizations then it is easy to reformulate FAM as a mixed-integer linear program by introducing additional decision variables to linearize $c_i(y_i, d_i)$. If $\tilde{d}_i$ has many realizations or is continuous, then this is not a viable approach. In this case, by Proposition 1 we can instead view FAM as a mixed-integer nonlinear program (MINLP) whose continuous relaxation is a convex nonlinear program. That said, it is

not possible to solve such an instance of FAM by commercially-available MINLP solvers since we do not have an analytical expression for $f(y) = \sum_{i=1}^{M} s_i Ec_i(\sum_{j=1}^{N} y_{ij}, \tilde{d}_i)$. So, we instead develop a cutting-plane algorithm to solve FAM.

3.1 A Cutting-plane Algorithm for FAM

A cutting-plane algorithm for FAM does not require an analytical expression for *f*(*y*). Rather, it requires being able to evaluate (or estimate) *f*(*y*) and its gradient $\nabla f(y)$, when *y* is fixed to a specific value. In general, *f*(*y*) is not differentiable because its definition includes nested functions involving positive-part operations. The following proposition gives conditions under which $f(y)$ is differentiable.

Proposition 2: Assume $\tilde{d}_i$ has finite mean and an absolutely continuous distribution for each *i* = 1,..,*M*. Then, $f(y) = \sum_{i=1}^{M} s_i Ec_i(\sum_{j=1}^{N} y_{ij}, \tilde{d}_i)$ is differentiable.

Even though *f*(*y*) can be differentiable, in general we cannot evaluate it (or its gradient) exactly. That said, we can estimate each expectation $Ec_i(y_i, \tilde{d}_i)$ by Monte Carlo sampling. Let $\tilde{d}_{i,r}$, $r = 1,\ldots,R$, be independent and identically distributed (i.i.d.) as $\tilde{d}_i$ and estimate $Ec_i(y_i, \tilde{d}_i)$ via $\frac{1}{R}\sum_{r=1}^{R} c_i(y_i, \tilde{d}_{i,r})$. We let $\bar{f}_R(y) = \sum_{i=1}^{M} s_i \frac{1}{R}\sum_{r=1}^{R} c_i(y_i, \tilde{d}_{i,r})$ and we define FAM_R as FAM, except that the objective function is replaced by $\sum_{j=1}^{N}\sum_{i=1}^{M} a_{ij}x_{ij} + \bar{f}_R(y)$. The following proposition characterizes solutions of FAM_R as the number of replications *R* grows large.

Proposition 3: Let $\tilde{d}_r = (\tilde{d}_{1,r},\ldots,\tilde{d}_{M,r})$, $r = 1,\ldots,R$, satisfy $\lim_{R\to\infty} \bar{f}_R(y) = f(y)$, with probability one (w.p.1). Let (x_R^*, y_R^*) denote an optimal solution to FAM_R. Then every limit point of $\{(x_R^*, y_R^*)\}_{R=1}^{\infty}$ solves FAM, w.p.1.

As indicated above, we select $\tilde{d}_{i,r}$, $r = 1,\ldots,R$, to be i.i.d. from the distribution of $\tilde{d}_i$. In this case, the pointwise convergence hypothesis of Proposition 3 holds by the strong law of large numbers for a sample mean of i.i.d. random variables. In what follows, our Monte Carlo sampling scheme generates the demand observations according to this i.i.d. scheme.

Proposition 3 justifies replacing FAM with FAM_R when the number of replications R is sufficiently large. Fortunately, given the definition of $c_i(y_i, \tilde{d}_i)$ we can choose R quite large. For any finite R, $\bar{f}_R(y)$ is convex but non-smooth. We can solve FAM_R using Kelley's (1960) cutting-plane method, adapted to deal with integer-valued decision variables x (see, e.g., Westerlund and Pettersson 1995). At iteration κ of the algorithm the following problem (Master-κ) is solved:

$$\underline{z}_\kappa = \min_{x,y,\theta} \sum_{j=1}^{N} \sum_{i=1}^{M} a_{ij} x_{ij} + \theta \tag{6}$$

s.t. (2)-(4)

$$\theta \geq \bar{f}_R(y^l) + g^l(y - y^l), \qquad l = 1, ..., \kappa - 1, \tag{7}$$

where $g^l \in \partial \bar{f}_R(y^l)$, i.e., g^l is a subgradient of $\bar{f}_R(y)$ at $y = y^l$.

The cutting-plane algorithm at each iteration forms a first-order Taylor approximation, i.e., a cut, at the current iterate $y^l : \bar{f}_R(y^l) + g^l(y - y^l)$. In what follows, the cut's gradient, g^l, and its intercept $\bar{f}_R(y^l) - g^l y^l$, are called *cut coefficients*. When we solve Master-κ we therefore have an outer piecewise linear approximation of $\bar{f}_R(y)$ given by $\max_{l=1,...,\kappa-1} [\bar{f}_R(y^l) + g^l(y - y^l)]$. The formulation in Master-κ linearizes this piecewise linear approximation via decision variable θ and constraints (7). See Appendix B for algorithm details.

4. DYNAMIC ALLOCATION MODEL

In this section, we develop a time-dynamic allocation model. Like the FAM of the previous section, product-to-line assignments must be decided at the beginning of the planning horizon. Unlike the FAM, in DAM the capacity allocation decisions can adapt to the demand in each period. DAM is a multi-stage stochastic program with binary first stage decision variables representing product-to-line assignments. Each of the subsequent stages is constrained by these first stage binary decisions. The model is given below with the following additional notation:

b_i^t = backlogged demand for family i in period t ($b_i^0 \equiv 0 \ \forall i$)

y_{ij}^t = capacity of line j allocated to produce family i in period t, in production units

Throughout this section, we assume that $\tilde{d}^t$, $t = 1,\ldots,T$ are independent and identically distributed random vectors.

Dynamic Allocation Model (DAM):

$$z^* = \min_{x\in\{0,1\}^{M\times N}} \sum_{j=1}^{N}\sum_{i=1}^{M} a_{ij}x_{ij} + E_{\tilde{d}^1} h^1(x, b^0, \tilde{d}^1) \tag{8}$$

where for $t = 1,\ldots,T$,

$$h^t(x, b^{t-1}, \tilde{d}^t) = \min_{y^t, b^t} \sum_{i=1}^{M} s_i b_i^t + E_{\tilde{d}^{t+1}} h^{t+1}(x, b^t, \tilde{d}^{t+1}) \tag{9a}$$

$$\left.\begin{array}{lll} \text{s.t.} & \displaystyle\sum_{j=1}^{N} y_{ij}^t + b_i^t = d_i^t + b_i^{t-1}, & \forall i \\ & \displaystyle\sum_{i=1}^{M} e_{ij} y_{ij}^t \le K_j, & \forall j \\ & y_{ij}^t \le (K_j / e_{ij}) x_{ij}, & \forall i, j \\ & y_{ij}^t, b_i^t \ge 0, & \forall i, j \end{array}\right\} (y^t, b^t) \in Y(x, b^{t-1}, d^t) \tag{9b}$$

and where $h^{T+1} \equiv 0$.

In the DAM, the process flexibility configuration is selected via x in (8) to minimize the cost of that configuration plus the expected operations cost to the planning horizon. That operations cost is captured in the recursion specified by (9), which takes x as input, and makes the allocation and resulting backlogging decisions in each time period, $t = 1,\ldots,T$. When selecting x the demand process $\{\tilde{d}^t\}_{t=1}^T$ is known only through its distribution. When deciding y^t and b^t in period t, we know the current period's demand realization, $\tilde{d}^t$, demand backlog from the previous period, b^{t-1}, and the distribution governing the future demand process, $\{\tilde{d}^{t+1},\ldots,\tilde{d}^T\}$. Beyond the fact that DAM has greater operational flexibility than FAM (see Figure 1), the structural form of the constraints in (9b) is the same as that in the FAM.

Multi-stage stochastic programs, such as the one in (8)-(9) represent significant computational challenges, even when the demands in each period are independent. When the demands have a continuous distribution, as we assume in our computational study in the next section, model (8)-(9) is intractable. Even if $\tilde{d}^t$ has a finite number of realizations in each time period, the size of the scenario tree grows exponentially with the number of time periods, and hence the model quickly becomes intractable. The fact that DAM has binary first stage decision variables adds further computational challenges.

When an exact solution of a multi-stage stochastic program is not computationally viable, we turn to approximations. If $\tilde{d}^t$ has a continuous distribution then we could replace it with a manageable number of realizations in each time period. There are multiple ways to generate such discrete approximations, but we do so using Monte Carlo sampling.

The requirement of having a modest number of realizations in each stage precludes direct application of the multi-stage decomposition algorithms in the stochastic programming literature to our DAM. So, we proceed in this section in four steps as follows: First in §4.1, we construct what we call an empirical scenario tree by replacing the true demand distribution at each stage by an empirical distribution constructed by Monte Carlo sampling. We call the dynamic allocation problem defined on this empirical tree EDAM. Second, we extend the sampling-based algorithm of Pereira and Pinto (1991) to solve EDAM in §4.2. Their algorithm requires extension because in addition to the standard staircase structure in which backlogged inventory is carried between adjacent time periods we also have binary first-stage flexibility decisions that are carried to all of the periods. This requires that we construct a non-standard cut, which we describe in detail. Third, in §4.3 we describe how we can construct a feasible policy for DAM using the cuts generated in solving EDAM. In the fourth and final step, we seek to establish whether our feasible policy is near-optimal. To do so, we first describe how to estimate the policy's expected cost in §4.4. Then, in §4.5 we show how to construct a confidence interval on the policy's optimality gap using a lower bound estimator again formed using EDAM. The solution validation ideas we use rely on Chiralaksanakul and Morton (2003), but have not been previously extended to problems with integer design decisions or decisions that directly affect all the time periods.

4.1 Empirical Scenario Tree Construction

In order to generate a sample scenario tree, we generate a set (indexed by *S*) of i.i.d. observations of the demand $\tilde{d}^{1,k}, k \in S$, in period 1. We then use this same set of observations to represent the realizations in each time period *t*. So, the first period sampled observations are $\tilde{d}^{1,k}, k \in S$. And, in period 2, each of these realizations has $\tilde{d}^{2,k} = \tilde{d}^{1,k}, k \in S$, as its descendent nodes, etc. In this way, our empirical scenario tree, like its "true" counterpart, exhibits interstage independence with identically distributed demand in each period. Hence, the dynamic allocation model defined

on an empirical scenario tree (EDAM) takes the following form after each expectation is replaced with the corresponding sample mean.

(EDAM)

$$\hat{z} = \min_{x \in \{0,1\}^{M \times N}} \sum_{j=1}^{N} \sum_{i=1}^{M} a_{ij} x_{ij} + \frac{1}{|S|} \sum_{k \in S} \hat{h}^1(x, b^0, \tilde{d}^{1,k}) \tag{10}$$

where for $t = 1, \ldots, T$, and $k \in S$,

$$\hat{h}^t(x, b^{t-1}, \tilde{d}^{t,k}) = \min_{y^t, b^t} \sum_{i=1}^{M} s_i b_i^t + \frac{1}{|S|} \sum_{k \in S} \hat{h}^{t+1}(x, b^t, \tilde{d}^{t+1,k}) \tag{11}$$

s.t. $(y^t, b^t) \in Y(x, b^{t-1}, \tilde{d}^{t,k})$,

where $\hat{h}^{T+1} \equiv 0$ and where the constraint set Y is defined in (9b).

Solving EDAM is of central importance in generating near optimal polices for DAM. Next, we present a method using sampling to provide near optimal solutions to EDAM.

4.2 An Algorithm to Solve EDAM

In this section, we develop a multi-stage nested decomposition algorithm to solve EDAM. Our algorithm is based on the idea of sequentially approximating the expected cost-to-go function in each stage with a piecewise linear function, similar to what we described in §3. Hence, one may use the approximate cost-to-go functions to make the x decisions at stage 0 and the y^t decisions at stage $t \geq 1$. Further, we note that due to the presence of first-stage binary assignment decisions, which feed each of the problems in the subsequent time periods, as in (11), EDAM is significantly harder to solve than a standard multi-stage stochastic linear program.

Here we present an algorithm extending that of Pereira and Pinto (1991) to handle binary first stage decisions, which feed all subsequent periods. First, the algorithm decomposes EDAM into a subproblem for each time period, including what we label $t = 0$, where x is selected. Then, the algorithm iteratively applies forward and backward phases. During a single forward pass, a demand realization $\tilde{d}^{t,r}$ is drawn from the sample set S and the stage t subproblem is solved using the current piecewise linear approximation of the cost-to-go function. In the first iteration this subproblem is solved myopically but as the algorithm proceeds the piecewise linear function better approximates the sample-mean functions in (10) and (11). Solving the subproblems leads to a backlog demand, $b^{t,r}$, being passed to stage t+1, where an independent realization $\tilde{d}^{t+1,r}$ is drawn from S, until we reach the final period T. So, during the forward phase the multi-stage

problem is solved along a given sample path of demand, knowing only the approximate cost-to-go function, and not the future period demands, at each period.

During the backward phase of the algorithm, given x and $b^{T-1,r}$ the stage T subproblem is first solved for all $\tilde{d}^{T,r} \in S$ and an optimality cut is passed to stage T-1. Next, given x and $b^{T-2,r}$ the stage T-1 subproblem is solved for all $\tilde{d}^{T-1,r} \in S$, and an optimality cut is passed to stage T-2. This backward pass continues until a cut is passed to stage $t = 1$ and finally to $t = 0$. The cuts accumulated in each stage represent a piecewise outer linear approximation of the cost-to-go function at that stage. Hence in each iteration of the algorithm, assignment decisions, x, are selected to minimize the objective (10) where the cost-to-go function is replaced by a piecewise linear approximation. As we iterate, the approximating functions become more precise and hence the assignment and allocation policies improve. Upon termination, the feasible policy obtained by the algorithm is evaluated by drawing independent demand samples from set S to provide an upper bound estimate. The details of the algorithm are given in Appendix C.

The algorithm we have just specified approximately solves EDAM. The output of the algorithm is a policy for EDAM. Specifically, at the end of the algorithm, the process flexibility configuration is given by the solution x to the stage 0 subproblem. Note that the subproblems replace the exact cost-to-go function with a set of linear cuts that have accumulated in the course of the algorithm. And, for the allocation policy, we first solve the stage 1 subproblem with its cuts, given x and $\tilde{d}^1$ to obtain y^1 and b^1 (note that when doing so $\tilde{d}^1, \ldots, \tilde{d}^T$ need not be sampled yet). Next, we solve the stage 2 subproblem with its cuts, given x, b^1 and $\tilde{d}^2$, etc., until we finally solve the stage T subproblem for y^T.

4.3 Near Optimal Policy Generation for DAM

In this section, we present a procedure for generating a near optimal feasible policy for DAM. For this purpose, we first generate an empirical scenario tree and the associated EDAM. The approximate model (EDAM) is then solved with the algorithm given in §4.2. As we have described, when the algorithm terminates, the subproblems at each stage t contain a set of cuts generated during the backward passes of the algorithm. Since these cuts approximate the cost-to-go functions of EDAM, they may also be used to approximate the cost-to-go functions of DAM. Hence, they define a feasible policy for the actual model, DAM as well.

In particular, we use the optimization models in Figure 2 to generate a good feasible policy for DAM and determine decisions (y^t, b^t) for each period $t = 1,\ldots,T$ and x at period $t = 0$.

For $t = 0$:	For $t = 1, \ldots, T$:
$\min_{x,\theta^1} \quad ax + \theta^1$	$\min_{y^t, b^t, \theta^{t+1}} \quad sb^t + \theta^{t+1}$
s.t. $\quad \theta^1 + \mu^{0,l} x \geq \alpha^{0,l} \quad l = 1, \ldots, R,$	s.t. $\quad \theta^{t+1} + \beta^{t,l} b^t \geq \alpha^{t,l} + \mu^{t,l} x \quad l = 1, \ldots, R,$
$x \in \{0,1\}^{M \times N}$	$(y^t, b^t) \in Y(x, \hat{b}^{t-1}, d^t),$
	(For $t=T$, the cut constraints and θ^{T+1} are absent)

Figure 2: Models for Generating a Feasible Policy for DAM

The vectors β, α and μ contain the cut coefficients and R is the total number of cuts obtained while solving EDAM with the method of §4.2 (see Appendix C.2 for details). Note that the cuts in each period $t > 1$ are parameterized by the first stage process flexibility decisions x.

4.4 Policy Cost Estimation (Upper Bound Estimation)

Once a feasible policy is identified for DAM, the next step is to evaluate its cost to obtain an upper bound on the optimal value of DAM, z^*. In particular, for a given demand sample path i, $(\tilde{d}^{1,i}, \ldots, \tilde{d}^{T,i})$, our policy generates a stream of feasible solutions, $\hat{x}$, $\hat{b}^1(\tilde{d}^{1,i})$, …, $\hat{b}^T(\tilde{d}^{T,i})$ for DAM and, the cost of the policy for that sample path is given by: $U^i = U(\tilde{d}^{1,i}, \ldots, \tilde{d}^{T,i})$ $= a\hat{x} + \sum_{t=1}^{T} s\hat{b}^t(\tilde{d}^{t,i})$. Since the identified policy is not necessarily optimal, the expected cost of the policy exceeds DAM's optimal value, i.e., $E\tilde{U} = EU(\tilde{d}^1, \ldots, \tilde{d}^T) \geq z^*$.

Next, to obtain a point estimate of $E\tilde{U}$, we generate η i.i.d. demand sample paths, $(\tilde{d}^{1,i}, \ldots, \tilde{d}^{T,i})$, $i = 1, \ldots, \eta$, and evaluate the cost of the policy. Then, an approximate one-sided $100(1-\alpha)\%$ confidence interval for $E\tilde{U}$ is $(-\infty, \overline{U}_\eta + z_\alpha s_u/\sqrt{\eta}]$, where $\overline{U}_\eta = (1/\eta)\sum_{i=1}^{\eta} U^i$ and $s_u^2 = (1/(\eta-1))\sum_{i=1}^{\eta}(U^i - \overline{U}_\eta)^2$. Here, z_α is the (1-α)-level quantile for a standard normal.

4.5 Lower Bound Estimation

This section explains how to develop a probabilistic lower bound for z^*, the optimal value of DAM. Our goal is to combine this bound with the one in §4.4 and to develop a confidence interval for the optimality gap of the policy generated in §4.3. Our lower bound estimator is based on the following proposition.

Proposition 4: Let $\tilde{L}$ denote the final lower bound generated by the algorithm of §4.2 for the optimal value of an EDAM. Then, $z^* \geq E\tilde{L}$.

Next, we develop a point estimate of $E\widetilde{L}$ to establish a lower bound for z^*. Hence, we construct ν i.i.d. sample scenario trees, $\Gamma^1,...,\Gamma^\nu$, and the associated EDAMs as explained in §4.1. Then, we solve these problems with the method of §4.2 to obtain the lower bound estimators $L^1,...,L^\nu$. Then, by the standard central limit theorem for i.i.d. random variables, an approximate one-sided 100(1-α)% confidence interval for z^* (also for $E\widetilde{L}$) is given by $[\overline{L}_\nu - z_\alpha s_l / \sqrt{\nu}, +\infty)$, where $\overline{L}_\nu = \frac{1}{\nu}\sum_{i=1}^{\nu} L^i$ and $s_l^2 = \frac{1}{\nu-1}\sum_{i=1}^{\nu}(L^i - \overline{L}_\nu)^2$. Finally, by combining this confidence interval with the one in §4.4, using the Boole-Bonferroni inequality, we obtain a confidence interval for the optimality gap of the feasible policy, i.e., for $E\widetilde{U} - z^*$. Specifically, an approximate 100(1-2α)% confidence interval for $E\widetilde{U} - z^*$ is given by $[0, (\overline{U}_\eta - \overline{L}_\nu)^+ + z_\alpha s_u / \sqrt{\eta} + z_\alpha s_l / \sqrt{\nu}]$.

5. COMPUTATIONAL RESULTS AND ANALYSIS

The purpose of the computational study in this section is two-fold. First, in §5.1, we investigate the value of operational flexibility by comparing the expected performance of the dynamic and fixed allocation models. Then, in §5.2, we specifically focus on the value of process flexibility in a multi-period decision environment, ignoring the cost of the assignments. Further, the computational results show that our algorithms are very effective to solve real-sized problems. Below for simplicity, we assume that production of a unit of any product requires the same amount of capacity and the production lines are identical, i.e., $e_{ij} = 1 \ \forall i, j$ and $K_j = 100 \ \forall j$. Indeed, in our motivating electronics manufacturer case, the number of assembly components and the complexity of the assembly operations do not significantly differ from one family to another, which is typical in high-volume mass customization. Accordingly, we do not observe a significant difference between the capacity requirements of different families.

For the computational study, we consider a problem with M = 6 families, N = 3 lines and T = 5 production periods. In practice, the firm has 10-15 product families and 6 production lines. However, from the manufacturing point of view similar product families that share a significant number of components are aggregated. Moreover, some set of production lines are also dedicated for producing certain product families. Hence for practical purposes it is sufficient to consider a 6-product 3-line problem. Unless otherwise stated, for ease of analysis, we assumed that

backlogging cost for all families is s_i = \$1 while the assignment cost for any line-product pair is a_{ij} = \$10 (while not shown here we have found similar results for different combinations of backlog and setup costs). Product demands, $\tilde{d}_i$, for the test problem are normally distributed, and generated to simulate the real situation at the firm. In particular, the product families with higher mean demand have lower coefficient of variation. Base mean (μ_{Base}) and variance (σ^2_{Base}) demand data for our analyses are given in Table 1.

Table 1: Base Demand Data for the Test Problem

	Family 1	Family 2	Family 3	Family 4	Family 5	Family 6	Total
Base Mean	5	35	40	45	85	90	300
Base Variance	0.5	22	26	31	97	102	

5.1 Value of Operational Flexibility: Comparing DAM and FAM

In this section, we numerically compare and analyze FAM and DAM developed in Sections 3 and 4, respectively. We setup an experimental design scaling the test problem given above. In particular, we consider two factors that scale: (*i*) the size of demand (ρ), and (*ii*) its coefficient of variation (β). We conjecture that these two factors should have a significant impact on the performance of the policies under investigation. We consider seven demand sizes (for ρ = 0.8, 0.9, 0.933, 1, 1.067, 1.1 and 1.2), as well as three levels of coefficient of variation (for β = 1, 3 and 5). More generally, for each experimental setting demand data $\tilde{d}_i$ is given by $\tilde{d}_i(\beta,\rho)$ such that $\mu_i = E\tilde{d}_i(\beta,\rho) = \rho\mu_{Base,i}$ and $\sigma_i^2 = \text{Var}\,\tilde{d}_i(\beta,\rho) = \rho^2\beta\sigma^2_{Base,i}$. The coefficient of variation of $\tilde{d}_i(\beta,\rho)$ is constant as we scale ρ and grows in β. Also, note that, when ρ=1 mean demand is equal to the capacity.

Below, we summarize computational results. The optimal operating costs under FAM are presented in Table 2. These are estimates obtained with a sample size of R=10^6. The associated standard deviations are about 0.03% of the sample mean estimates. In our example, the operating cost increases with demand size, ρ, and the coefficient of variation of demand, β (henceforth referred as to variability in demand). In addition, the introduction of variability (to a deterministic system) has the highest impact on the operating costs when ρ=1. Intuitively, when we have plenty of capacity, demand variability can be buffered with capacity, similarly when we have the capacity well below the demand then the variability will not affect the expected backlog too much, since the capacity is already fully utilized.

Table 2: Optimal Operating Cost under Fixed Allocation Model (z_f)

ρ	Deterministic (β=0)	Low Variability (β=1)	Moderate Variability (β=3)	High Variability (β=5)
0.8	60	71.65	90.11	116.09
0.9	70	97.61	165.79	223.56
0.933	70	126.45	214.20	281.33
1	80	238.95	353.58	431.98
1.067	370	443.81	547.68	626.33
1.1	520	566.18	659.48	735.30
1.2	940	981.56	1041.25	1102.78

The results for DAM are summarized in the next three tables. We use a sample size of 100 in each time period in our Monte Carlo procedures (details of the computational parameters are given in Appendix D). In Table 3, we provide the operating cost estimates of the near optimal policy identified by the solution method of §4.3. Then, in Table 4, we provide the lower bound estimates on the true optimal value of DAM as explained in §4.5. Finally, we present the optimality gap estimates of the identified policies in Table 5. (In Table 5, %Gap is calculated by dividing the length of the confidence interval by the mean upper bound estimate, $\bar{U}_\eta$.)

Table 3: Cost of the Identified Feasible Policy for DAM ($\bar{U}_\eta, s_u/\sqrt{\eta}$)

ρ	Deterministic (β=0)		Low Variability (β=1)		Moderate Variability (β=3)		High Variability (β=5)	
	$\bar{U}_n$	$s_u/\sqrt{\eta}$	$\bar{U}_n$	$s_u/\sqrt{\eta}$	$\bar{U}_n$	$s_u/\sqrt{\eta}$	$\bar{U}_n$	$s_u/\sqrt{\eta}$
0.8	60	0	70.02	0.00	72.33	0.04	78.68	0.10
0.9	70	0	74.54	0.06	90.28	0.14	104.84	0.28
0.933	70	0	84.48	0.07	107.59	0.29	132.16	0.49
1	80	0	153.91	0.47	207.41	0.82	248.25	1.06
1.067	370	0	386.88	0.81	419.46	1.34	453.74	1.62
1.1	520	0	525.46	0.92	550.31	1.43	576.86	1.82
1.2	940	0	971.10	1.05	976.40	1.76	990.69	2.20

In the worst case, the optimality gap of the identified policy is around 5% of the estimated cost and for low and medium variability cases the gap is less than 2.5%. Hence, we conclude that, for our example, the approach presented in §4 generates near optimal policies for DAM. In addition, since our gap generation mechanism is based on sampling, it is natural to observe that the algorithm performance slightly degrades as the variability of demand grows.

Table 4: Lower Bound for the Optimal Operating Cost of DAM ($\bar{L}_\nu, s_l/\sqrt{\nu}$)

ρ	Deterministic (β=0)		Low Variability (β=1)		Moderate Variability (β=3)		High Variability (β=5)	
	$\bar{L}_\nu$	$s_l/\sqrt{\nu}$	$\bar{L}_\nu$	$s_l/\sqrt{\nu}$	$\bar{L}_\nu$	$s_l/\sqrt{\nu}$	$\bar{L}_\nu$	$s_l/\sqrt{\nu}$
0.8	60	0	69.80	0.09	71.84	0.15	77.35	0.32
0.9	70	0	74.70	0.22	90.45	0.60	103.90	1.00
0.933	70	0	84.56	0.32	107.37	0.92	128.91	1.44
1	80	0	153.90	0.85	207.57	1.50	245.17	3.84
1.067	370	0	386.67	0.60	420.07	1.20	445.51	1.88
1.1	520	0	524.92	0.40	550.17	1.05	570.02	1.61
1.2	940	0	969.71	0.46	975.31	0.66	986.27	1.08

Table 5: Approximate 95% CI for the Optimality Gap of the Feasible Policy

ρ	Deterministic (β=0)		Low Variability (β=1)		Moderate Variability (β=3)		High Variability (β=5)	
	95%CI	% Gap	95%CI	% Gap	95%CI	% Gap	95%CI	% Gap
0.8	N/A	N/A	[0, 0.40]	0.57	[0, 0.87]	1.20	[0, 2.15]	2.74
0.9	N/A	N/A	[0, 0.55]	0.74	[0, 1.46]	1.62	[0, 3.46]	3.30
0.933	N/A	N/A	[0, 0.76]	0.90	[0, 2.60]	2.42	[0, 7.02]	5.31
1	N/A	N/A	[0, 2.61]	1.69	[0, 4.54]	2.19	[0, 12.70]	5.12
1.067	N/A	N/A	[0, 2.99]	0.77	[0, 4.97]	1.18	[0, 15.08]	3.32
1.1	N/A	N/A	[0, 3.11]	0.59	[0, 5.00]	0.91	[0, 13.57]	2.35
1.2	N/A	N/A	[0, 4.33]	0.45	[0, 5.85]	0.60	[0, 10.86]	1.10

Comparing the results in Table 2 and Table 3, it is clear that operational flexibility (i.e., using a dynamic allocation model) significantly reduces the negative impact of variability on the operating cost. In particular, it becomes more beneficial to use DAM instead of FAM as the system becomes more variable and the mean demand is close to the capacity. The maximum difference is observed when $\beta = 5$ and $\rho = 1$. Table 6 summarizes the absolute benefits of using DAM over FAM under our experimental settings.

Table 6: Expected Absolute Benefits of using DAM over FAM (z_f-$\overline{U}$)

ρ	Deterministic (β=0)	Low Variability (β=1)	Moderate Variability (β=3)	High Variability (β=5)
0.8	0	1.63	17.78	37.41
0.9	0	23.07	75.51	118.72
0.933	0	41.97	106.61	149.17
1	0	85.04	146.17	183.73
1.067	0	56.93	128.22	172.59
1.1	0	40.72	109.17	158.44
1.2	0	10.46	64.85	112.09

Intuitively, if the mean demand is well above the capacity, then operational flexibility does not provide much benefit since the system capacity is already fully utilized and dynamically changing the allocations does not help to decrease the expected backlog. When the mean demand is well below the capacity, then again operational flexibility is not very beneficial. If the capacity is well-balanced with respect to the demand, then there is significant opportunity for decreasing the expected backlog by revising the allocations periodically. In addition, the absolute benefits of operational flexibility are increasing with the variability of the system.

In Table 7, we provide the expected percentage benefits of using DAM over FAM. The effect of variability on the percentage and absolute benefits is similar: the percentage benefits also increase with the variability of the system. The impact of capacity availability is slightly different in this case, i.e., the percentage benefits are maximized when the firm has some slack capacity. For moderate and high variability cases, the percentage benefits are maximized around

$\rho = 0.9$. Although it is not presented in the table, percentage benefits under the low variability case are also maximized when ρ is slightly below 1.

When we have some slack capacity, demand and capacity mismatches do not occur as often as in the case when $\rho = 1$, and hence DAM has relatively fewer opportunities to adaptively change allocations to decrease the backlog. This reduces the absolute benefits. However, as ρ decreases, DAM becomes more efficient in eliminating demand and capacity mismatches (i.e., it can eliminate a bigger proportion of these mismatches due to the slack capacity) and hence the backlogging cost under DAM decreases faster than FAM. Accordingly, the percentage savings in backlogging cost increases as ρ decreases. On the other hand, as slack capacity increases, the assignment costs shrink to a certain positive value while the backlogging cost approaches zero in both cases, so the assignment cost becomes the dominant term after a threshold level of ρ. Hence, the overall relative performance difference $((z_f - \overline{U}) / z_f)$ is maximized when we have some slight slack capacity and then it goes to zero as the assignment costs dominate.

Table 7: Expected Percentage Benefits of using DAM over FAM $(z_f - \overline{U}) / z_f$

ρ	Deterministic (β=0)	Low Variability (β=1)	Moderate Variability (β=3)	High Variability (β=5)
0.8	0	2.27	19.73	32.22
0.9	0	23.63	45.55	53.10
0.933	0	33.19	49.77	53.02
1	0	35.59	41.34	42.53
1.067	0	12.83	23.41	27.56
1.1	0	7.19	16.55	21.55
1.2	0	1.07	6.23	10.16

Table 8 shows the number of assignments for the optimal process flexibility configurations obtained by DAM for our experimental design. When ρ is sufficiently small the optimal number of assignments is $M = 6$ (the number of families). When ρ is large and β is small the number of assignments is $N = 3$ (the number of lines). This occurs for ρ larger than 1.2, even when $\beta = 0$. When $\beta = 0$ and $\rho = 1$ the total demand is 300, exactly matching total capacity, and the capacity is fully utilized with 8 assignments. More generally, Table 8 shows that denser chains like this are needed when the variability is high and demand and capacity are well-balanced. This is due to the greater need, in these cases, for splitting the families among the lines to utilize the capacity effectively. Further, we note that even the densest of our optimal configurations, are even less dense than a symmetrical J&G chain which has 12 assignments. This is primarily driven by the asymmetric problem structure and the tradeoff between backlogging and setups costs.

Table 8: Optimal Number of Assignments obtained by DAM

ρ	Deterministic ($\beta = 0$)	Low Variability ($\beta = 1$)	Moderate Variability ($\beta = 3$)	High Variability ($\beta = 5$)
0.8	6	7	7	7
0.9	7	7	8	8
0.933	7	8	8	8
1	8	8	8	8
1.067	7	7	8	8
1.1	7	7	7	8
1.2	4	7	7	7

5.2 Value of Process Flexibility

In this section, we investigate the value of process flexibility, under optimal and myopic operating policies. Ignoring the cost of assignments in our case, allows us to extend the results of J&G to a multi-period framework in which the firm must also decide how to allocate capacity to demand over time, i.e., the allocation (operating) policies.

First, we consider an *optimal dynamic allocation* policy, where the firm decides allocations to minimize its expected backlogging cost over the full planning horizon (as in the DAM). In this case, we solve DAM for a given flexibility configuration to find the optimal allocations. Next, we consider a *myopic dynamic allocation* policy, where the firm minimizes its backlogging cost myopically in each period. We label this latter policy MDAM. Mathematically, MDAM can be stated as a special case of DAM as

$$z^* = \min_{x \in \{0,1\}^{M \times N}} \sum_{j=1}^{N} \sum_{i=1}^{M} a_{ij} x_{ij} + E_{\tilde{d}^1, \ldots, \tilde{d}^T} \sum_{t=1}^{T} h^t(x, b^0, \tilde{d}^t)$$

where for $t = 1,\ldots,T$,

$$h^t(x, b^{t-1}, \tilde{d}^t) = \min_{y^t, b^t} \sum_{i=1}^{M} s_i b_i^t$$

$$s.t. \qquad (y^t, b^t) \in Y(x, b^{t-1}, \tilde{d}^t)$$

We start with the same example as in the previous section and consider a particular experimental setting with $\rho = 1$ and $\beta = 5$. The demand and cost data are summarized in Table 9. (We again use a sample size of 100 in each time period in our Monte Carlo procedures.)

Table 9: Test Problem Demand and Cost Data

	Family 1	Family 2	Family 3	Family 4	Family 5	Family 6	Total
Mean demand	5	35	40	45	85	90	300
Variance of demand	2.5	110	130	155	485	510	
Backlogging cost	1	1	1	1	1	1	

In this section, the purpose of the numerical analyses is to demonstrate that the value of process flexibility depends on the firm's operating policies. First, we compare the value of the

partial-flexibility configuration shown in Figure 3.b to the full process flexibility case given in Figure 3.a under DAM and MDAM. In practice, Figure 3.b may represent the current operating configuration of the firm, or it may be implied by an asymmetric cost of assignments. We also explore how the results change under a symmetric J&G chain and a reduced chain given in Figures 3.c and 3.d, respectively. The comparisons are with respect to expected backlogging cost, i.e., $a_{ij} = 0,\ \forall i, j,$ and the results are summarized in Table 10.

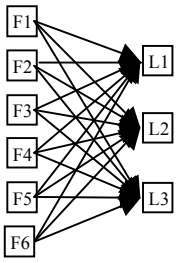

Figure 3.a: Full Process Flexibility

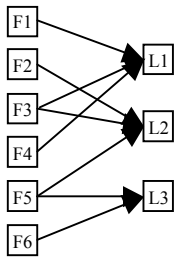

Figure 3.b: Partial Process Flexibility

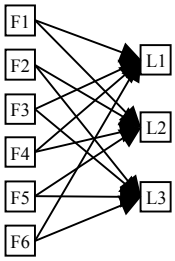

Figure 3.c: J&G Chain Process Flexibility

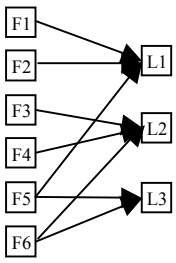

Figure 3.d: Reduced Chain Process Flexibility

Table 10: Value of Flexibility Configurations in Figures 3.a-d under DAM and MDAM

	Figure 3.a	Figure 3.b	Figure 3.c	Figure 3.d
Optimal Allocation Cost (DAM)	164.96	257.88	164.96	179.98
Myopic Allocation Cost (MDAM)	164.96	263.68	164.96	180.16
(MDAM-DAM)/MDAM	0.00%	2.20%	0.00%	0.10%
Performance gap relative to Full-Flex under DAM		56.33%	0.00%	9.11%
Performance gap relative to Full-Flex under MDAM		59.84%	0.00%	9.21%

Under the optimal dynamic allocation policy, the expected backlogging cost of the partial flexibility configuration of Figure 3.b is 56.33% higher than the cost of the full flexibility configuration. However, this difference in backlogging cost grows to 59.84% under the myopic dynamic allocation policy. Hence, the value of the partial configuration is reduced under the myopic allocation case. Specifically, the expected backlogging cost for the partial configuration is 2.20% higher under the MDAM compared to DAM.

The intuition for the results in Table 10 is as follows. Since the myopic allocation policy minimizes the immediate cost in every period, it does not take into account the impact of the backlogging plan on future operations. For instance, in this particular example, the assignment configuration in Figure 3.b implies that the total capacity accessible to produce families F2, F5 and F6 (first group) is 200 units while their total mean demand is 210. On the other hand, for the remaining families, F1, F3 and F4 (second group), the total accessible capacity is 200 units while their expected demand is only 90 units. Therefore, backlogging the first group, when it is also possible to backlog the second group, creates an increased risk of backlogging in the future since

such a policy further distorts the capacity and demand imbalance in the system. Hence, the myopic solution deteriorates over time.

We now repeat the example with partial flexibility with a different cost structure that is, in a relative sense, biased to encourage backlogging of the families with restricted capacity via $s_2 = s_5 = s_6 = 1$ and $s_1 = s_3 = s_4 = 1.005$. In this case, the percentage difference between the performance of DAM and MDAM increases from 2.20% to 4.76%. On the other hand, if the backlogging cost scheme is reversed (i.e., $s_2 = s_5 = s_6 = 1.005$ and $s_1 = s_3 = s_4 = 1$) the difference is close to zero. In general, under any arbitrary backlogging cost scheme, the value of the process flexibility configuration in Figure 3.b would differ under DAM and MDAM since the optimal policy identified by DAM is a forward-looking policy, which dynamically prioritizes families to backlog by taking the current backlog levels, backlogging cost scheme and the future demand into account. Hence, a myopic rule may fail to perform well.

On the other hand, a symmetric J&G chain proves to be a robust configuration with respect to the operational decisions in our multi-period setting. The backlogging cost (in other words, the value of process flexibility) does not change whether the firm operates under myopic or optimal operating policies. That is, myopic capacity allocation decisions do not lead to important backlogging consequences since such suboptimal decisions are absorbed by the process flexibility configuration. However, the J&G chain involves 1.5 times more assignments than the configuration in Figure 3.b.

The J&G chain, in our case, behaves almost identically to a full flexibility configuration due to the relatively small size of the problem. Figure 3.d presents a reduced form of the J&G chain, involving significantly fewer links. The backlogging cost of this reduced-chain is only 9.11% higher than that of the full flexibility configuration, while the performance gap between DAM and MDAM under this reduced-chain is virtually zero (0.1%).

When we consider the assignment costs ($a_{ij} = \$10, \forall\, i,j$) we obtain the results in Table 11. In this case, full flexibility is the most expensive solution, while the reduced-chain has the minimum cost since it achieves the best balance between backlogging and assignment costs among the configurations in Figure 3. We note that with arbitrary assignment costs, finding the optimal process flexibility configuration is a combinatorial problem, and the DAM in Section 3 is essential in addressing the tradeoff between these two costs and achieving an optimal process flexibility configuration.

Table 11: Total Cost of Configurations in Figures 3.a-d under DAM and MDAM

	Figure 3.a	Figure 3.b	Figure 3.c	Figure 3.d
Optimal Allocation (DAM)	344.96	337.88	284.96	259.98
Myopic Allocation (MDAM)	344.96	343.68	284.96	260.16
(MDAM-DAM)/MDAM	0.00%	1.69%	0.00%	0.07%
Performance gap relative to Full-Flex under DAM		-2.05%	-17.39%	-24.63%
Performance gap relative to Full-Flex under MDAM		-0.37%	-17.39%	-24.58%

Next, we illustrate that the performance gap under DAM and MDAM may still exist even for very efficient chain flexibility configurations. For this purpose we extend our numerical example in the previous case to larger scale so that a chain configuration is significantly different than a full process flexibility configuration. We present an example with 6 families, 6 production lines (each with 100 units of capacity) and 10 time periods. Demand (again, normally distributed) and backlogging cost information for each family is provided in Table 12.

Table 12: Test Problem Demand and Cost Data

	Family 1	Family 2	Family 3	Family 4	Family 5	Family 6	Total
Mean Demand	30	80	80	150	110	150	600
Variance of Demand	80	500	550	1000	600	1200	
Backlogging cost	1	1	1	1	1	1	

Below, we investigate the value of the flexibility scheme in Figure 4.b, configured as a chain that achieves almost all the benefits of full flexibility under DAM. Table 13 shows that under DAM the chain in Figure 4.b is very effective and its cost deviates from the full flexibility case by only 1%. The effectiveness of the chain decreases significantly under MDAM and its cost deviates from the full flexibility case by over 4%. Hence, even if a chain is very efficient, under DAM, its efficiency may degrade significantly under MDAM.

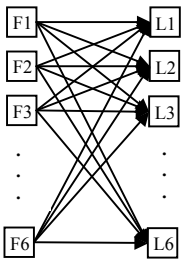

Figure 4.a: Full Process Flexibility

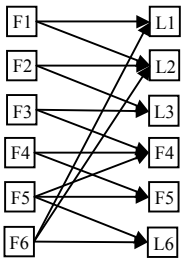

Figure 4.b: Chain Process Flexibility

Table 13: Value of Flexibility Configurations in Figures 4.a and 4.b under DAM and MDAM

Allocation Policy	Figure 4.a Backlog Cost	Figure 4.b Backlog Cost	(Chain − Full)/Full
Optimal Allocation (DAM)	837.17	845.625	1.01%
Myopic Allocation (MDAM)	837.17	873.169	4.30%
(MDAM-DAM)/MDAM	0.00%	3.15%	

In Figure 4.b, each product family is either assigned to 2 or 3 production lines. When we increase the number of links in this configuration, we observe that the performances of the

chaining configuration under MDAM and DAM grow closer. This is because increased process flexibility reduces the need for forward-looking capacity allocation decisions, and hence the performance of myopic allocation policies improves. Note that in Tables 10 and 13, MDAM performs as well as DAM under full process flexibility. Finally, in Table 14 we analyze the effectiveness of the chaining configuration in Figure 4.b as the number of time periods changes.

Table 14: Relative Value of Flexibility Configuration in Figure 4.b for T = 2, 8 and 14

	2-Period Problem	8-Period Problem	14-Period Problem
(Chain-Full)/Full under DAM	1.64%	1.09%	0.89%
(Chain-Full)/Full under MDAM	2.52%	4.13%	4.63%
(MDAM-DAM)/MDAM under Chain Flexibility	0.85%	2.92%	3.57%
(MDAM-DAM)/MDAM under Full Flexibility	0.00%	0.00%	0.00%

Two things stand out in Table 14. First, the effectiveness of the chaining configuration improves as the number of periods grows under DAM. This is because the system becomes more congested over time as the demand is backlogged, and hence the performance difference between the chaining and full-flexibility configurations shrinks. Second, as the number of periods increases the effectiveness of the chaining configuration deteriorates under MDAM. This is due to the fact that it becomes more important to look forward when making allocation decisions over longer planning horizons. In particular, for a single-period problem, a myopic allocation is optimal, but as the number of periods grows the suboptimality of myopic decisions becomes more severe. On the other hand, the congestion argument mentioned above also applies here and eventually as the number periods keeps growing, the performance difference between the chaining and full flexibility configurations shrinks, as well. In our example, up to 14 periods, the first factor dominates and the performance gap increases with the number of periods.

6. DISCUSSION

In this paper, we addressed a real life production and capacity management problem motivated by a high-tech electronic device manufacturer, which produces multiple product families on multiple flexible assembly lines over multiple time periods under demand uncertainty. While our motivation stems from the electronics industry, many of the same issues considered here extend to a wide range of MTO manufacturing environments. For example, BMW employs a business model which allows customers to make changes to their vehicle up to six days before final assembly. With such flexibility, it is economical for BMW to offer its customers 550,000 variants of the Z3 vehicle.

We specifically modeled and analyzed the value of process and operational flexibility, in a multi-period manufacturing environment with stochastic demand. Regarding the operational flexibility of the firm, we studied a fixed allocation model (FAM), a fully optimized dynamic allocation model (DAM) and a myopically optimized dynamic allocation model (MDAM). We formulated FAM as a single-stage stochastic program and developed a cutting plane algorithm to solve it. In the two dynamic allocation models, the firm may change the allocation decisions from period to period after observing the demand. DAM is formulated as a multi-stage stochastic program with binary first stage decision variables. For DAM we outlined a method to obtain near optimal policies and to generate bounds on the optimality gap for a given feasible solution.

In contrast to FAM, DAM allows the company to utilize operational flexibility. By comparing the performance of these two models we show that operational flexibility is most valuable under high demand variability and when mean demand and capacity is well-balanced.

Ignoring the cost of assignments, we analyze the value of process flexibility under optimal and myopic dynamic operating policies. Our computational results show that the value of process flexibility may depend significantly on the operating policy employed by the firm to allocate capacity in a multi-period production environment. In particular, we show that a firm operating with a myopic allocation policy may require the adoption of more process flexibility to hedge against demand uncertainty, compared to the optimal dynamic allocation case.

Finally, the impact of the number of time periods is revealing. Under an optimal dynamic allocation policy, the effectiveness of a chain flexibility configuration improves as the number of periods increases, since the system becomes more utilized over time. However, if a myopic operating policy is used, then the myopic solutions become more and more distorted as the number of periods increases and hence, the effectiveness of the chaining configuration decreases.

Our computational results suggest that our methods are effective for solving real-sized problems. Decomposition methods in the two-stage setting have benefitted from the use of a trust region or a quadratic proximal term to speed convergence, and we could similarly benefit from using these in the multi-stage setting. Moreover, we have extended the decomposition algorithm of Pereira and Pinto (1991) because of its (relative) simplicity to describe. Enhanced versions designed to reduce computational effort have been developed by Chen and Powell (1999) and Donohue and Birge (2006), and we could benefit from these enhancements.

We consider a flexibility management problem given that the firm already invested in capacity. In a more strategic-level investment problem, one may also endogenize the capacity decision of the firm. In our case, as it is suggested by the numerical results in Section 5, capacity is a substitute for flexibility. In addition, we have assumed that the demand is independent across time. Relaxing this assumption may also generate additional insights. For example, when the demand is positively correlated, we expect a reduction in DAM's ability to smooth out the production over time - and mitigate backlogging - compared to FAM.

We note that operating a DAM is more complex and may involve intangible costs such as reduced learning effects and increased scheduling problems. Exploring and modeling such intangible costs of flexibility offers a new and potentially fruitful future research direction. Further, exploring the impact of a capacity requirement mix on the product families may provide additional managerial insights. Finally, we have shown the value of flexible manufacturing configurations under demand uncertainty. Similar benefits from flexible configurations can emerge under supply uncertainty. Hence, we intend to design multi-stage flexible supply chains, which are robust to both demand and capacity fluctuations in an MTO environment.

Acknowledgements The authors thank three anonymous referees for helpful comments that improved the paper. This work has been partially supported by the National Science Foundation through grants CMMI-0653916 and CMMI-0855577.

References

Bish, E.K., Muriel, A. and Biller, S. (2005) Managing flexible capacity in a make-to-order environment. *Management Sci.*, 51, 167-180.

Boyabatli, O. and Toktay, B. (2004, January) Operational hedging: A review with discussion. *Working Paper.*

Chen, Z.L. and Powell, W.B. (1999) Convergent cutting plane and partial-sampling algorithm for multistage stochastic linear programs with recourse. *Journal of Optimization Theory and Applications*, 102, 497–524.

Chiralaksanakul, A. and Morton, D.P. (2004) Assessing policy quality in multi-stage stochastic programming. *Stochastic Programming E-Print Series*.

Chod, J., Rudi, N. and Van Mieghem, J.A. (2010) Operational flexibility and financial hedging: Complements or substitutes? *Management Sci.,* 56, 1030–1045.

Donohue, C.J. and Birge, J.R. (2006) The abridged nested decomposition method for multistage stochastic linear programs with relatively complete recourse. *Algorithmic Operations Res.*, 1, 20-30.

Fine, C.H. and Freund, R.M. (1990) Optimal investment in product-flexible manufacturing capacity. *Management Sci.*, 36, 449–466.

Garavelli, A.C. (2003) Flexibility configurations for the supply chain management. *Int. J. Production Economics*, 85, 141–153.

Graves, S.C. and Tomlin, B.T. (2003) Process flexibility in supply chains. *Management Sci.*, 49, 907-919.

Gunasekaran, A. and Ngai, E.W.T. (2005) Build-to-order supply chain management: A literature review and framework for development. *Journal of Operations Management*, 23, 423–451.

Gurumurthi, S. and Benjaafar, S. (2004) Modeling and analysis of flexible queueing systems. *Naval Research Logistics*, 51, 755-782.

Hopp, W.J., Tekin, E. and Van Oyen, M.P. (2004) Benefits of skill chaining in production lines with cross-trained workers. *Management Sci.*, 50, 83–98.

Huchzermeier, A. and Cohen, M.A. (1996) Valuing operational flexibility under exchange rate risk. *Operations Research*, 44, 100-113.

Iravani, S.M.R., Kolfal, B. and Van Oyen, M.P. (2007) Call center labor cross-training: It's a small world after all. *Management Sci.*, 53, 1102–1112.

Jordan, W.C. and Graves, S.C. (1995) On the principles of the benefits of manufacturing process flexibility. *Management Sci.*, 41, 577–594.

Kelley, J.E. (1960) The cutting plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8, 703–712.

Li, S. and Tirupati, D. (1994) Dynamic capacity expansion problem with multiple products: technology selection and timing of capacity additions. *Operations Research*, 42, 958–976.

___ and ___ (1995) Technology choice with stochastic demands and dynamic capacity allocation: a two product analysis. *Journal of Operations Management*, 12, 239–258.

___ and ___ (1997) Impact of product mix flexibility and allocation policies on technology. *Computers and Operations Research*, 24, 611–626.

Pereira, M.V.F. and Pinto, L.M.V.G. (1991) Multistage stochastic optimization applied to energy planning. *Mathematical Programming*, 52, 359–375.

Westerlund, T. and Pettersson, F. (1995) An extended cutting plane method for solving convex MINLP problems. *Computers and Chemical Engineering Sup*. 19, 131–136.

Van Mieghem, J.A. (1998) Investment strategies for flexible resources. *Management Sci.*, 44, 1071–1078.

___ and Rudi, N. (2002) Newsvendor networks: Inventory management and capacity investment with discretionary activities. *M&SOM*, 4, 313-335.

___ (2003) Capacity management, investment, and hedging: Review and recent development. *M&SOM*, 5, 269-302.