CORE

Metadata, citation and similar papers at core.ac.uk

Provided by Calhoun, Institutional Archive of the Naval Postgraduate School

Calhoun

Institutional Archive of the Naval Postgraduate School

Calhoun: The NPS Institutional Archive

Faculty and Researcher Publications Faculty and Researcher Publications

2008

Allocation of flexible and indivisible resources with decision postponement and demand learning

Bish, Ebru K.

DUDLEY
KNOX
LIBRARY

Calhoun is a project of the Dudley Knox Library at NPS, furthering the precepts and goals of open government and government transparency. All information contained herein has been approved for release by the NPS Public Affairs Officer.

Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943

http://www.nps.edu/library

Available online at www.sciencedirect.com

ScienceDirect

European Journal of Operational Research 187 (2008) 429–441

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

www.elsevier.com/locate/ejor

Production, Manufacturing and Logistics

Allocation of flexible and indivisible resources with decision postponement and demand learning ☆

Ebru K. Bish [a,*], Kyle Y. Lin [b,1], Seong-Jong Hong [a,2]

[a] *Grado Department of Industrial and Systems Engineering (0118), 250 Durham Hall, Virginia Tech, Blacksburg, VA 24061, USA*
[b] *Operations Research Department, 1411 Cunningham Road, Naval Postgraduate School, Monterey, CA 93943, USA*

Received 12 September 2005; accepted 7 March 2007
Available online 11 April 2007

Abstract

We consider a firm that uses two perishable resources to satisfy two demand types. Resources are *flexible* such that each resource can be used to satisfy either demand type. Resources are also *indivisible* such that the entire resource must be allocated to the same demand type. This type of resource flexibility can be found in different applications such as movie theater complexes, cruise lines, and airlines. In our model, customers arrive according to independent Poisson processes, but the arrival rates are uncertain. Thus, the manager can learn about customer arrival rates from earlier demand figures and potentially increase the sales by postponing the resource allocation decision. We consider two settings, and derive the optimal resource allocation policy for one setting and develop a heuristic policy for the other. Our analysis provides managerial insights into the effectiveness of different resource allocation mechanisms for flexible and indivisible resources.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Retailing; Resource flexibility; Demand forecast; Capacity allocation; Perishable inventory

1. Introduction

In this paper, we study the optimal capacity allocation[3] decision and the value of postponement for *flexible* and *indivisible* resources in the presence of *forecast uncertainty*. Although indivisible and flexible resources are utilized in various service industries, issues dealing with the management of such resources have not received much attention in the operations management literature. Our objective is to provide managerial implications and guidelines on how to manage this type of resources.

Specifically, we consider a service system that utilizes two capacitated and perishable resources to satisfy two types of consumers (demand streams) arriving stochastically and dynamically over a selling season. Each consumer type requires a different type of service, to be provided at the end of the

☆ This research is supported in part by NSF Grants DMI-0010032, DMI-0223314, and DMI-0245643.

* Corresponding author. Tel.: +1 540 231 7099; fax: +1 540 231 3322.

E-mail addresses: ebru@vt.edu (E.K. Bish), kylin@nps.edu (K.Y. Lin), sehong1@vt.edu (S.-J. Hong).

[1] Tel.: +1 831 656 2648; fax: +1 831 656 2595.

[2] Fax: +1 540 231 3322.

[3] We use the terms "capacity allocation" and "resource allocation" interchangeably throughout the paper.

0377-2217/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2007.03.013

selling season. Resources are (i) *flexible* because each resource can provide either type of service, (ii) *indivisible* because each resource can only provide one type of service in its entirety, and (iii) *perishable* because resource capacity cannot be stored in the form of inventory. We assume resources are identical except for their capacities. In order to maximize the expected revenue at the end of the selling season, the revenue manager faces two operational-level questions: (i) whether or not to grant the service request of each arriving consumer, and (ii) how to allocate the resources to the demand streams. Obviously, these two questions are closely related: Once the capacity allocation decision is given, the manager knows exactly up to how many consumers of each type he can admit into the system. What complicates this decision problem is the fact that in most real-world applications, arrival rates of the demand streams are not known, and can only be estimated with uncertainty at the beginning of the selling season.

To simplify the analysis, we assume there are no cancelations (hence, no overbooking); that is, every accepted consumer will show up at the time of service and be granted service. In addition, we assume that there is no consumer-driven substitution; that is, a consumer will not switch to a different service type if what she desires is unavailable. We will discuss the implications of these assumptions in detail in Section 6.

One commonly used strategy is to allocate resources to the demand streams at the beginning of the selling season in a way that maximizes the expected total revenue; we refer to this strategy as the *no postponement* strategy. Although this strategy is convenient and easy to implement, it is often difficult to make the "right" allocation that early, especially when the demand forecast is subject to errors. By postponing the allocation decision, it is possible for the manager to learn about the demand pattern from early sales figures to make a better allocation decision at a later time. In this paper, we study this capacity allocation problem and devise two postponement strategies. Our objective is to analyze the effectiveness of each postponement strategy, and the value of postponement in general, for managing flexible and indivisible resources.

The capacity allocation problem of flexible and indivisible resources arises in many real-world situations. For example, consider a multiplex movie theater that has several screens of different seating capacities. The manager starts selling movie tickets on the Internet and over the phone well in advance.[4] By scheduling different movies around the same time, it is possible for the manager not to assign movies to screens until close to the show time. The screens in this example are the flexible, divisible, and perishable resources. Because screens have different seating capacities, the manager would like to allocate a screen of a larger seating capacity to a more "popular" movie in order to increase sales. However, before the ticket sale begins, the manager need not have a good idea about each movie's popularity. For another example, a transportation service provider (a cruise company, a shuttle service provider, or an airline) usually has vehicles of different sizes that need to be assigned to different services around the same time. Still another example can be found in the manufacturing industry, where a production facility has different production lines that can be set up to produce one of many different products.

As discussed above, resource flexibility allows the manager to postpone the resource allocation decision to exploit benefits of learning through early demand figures and hedge against demand and capacity imbalances, thus providing a *risk pooling* effect. While several strategies for risk pooling – such as centralization of inventory, delayed product differentiation, component commonality, and lateral transshipments – have been analyzed extensively in the literature (see de Kok and Graves, 2003; Tayur et al., 1999, and the references therein), only recently have flexible resource management issues been incorporated into operations management models; see Van Mieghem (2003) for an excellent review of research in this area. However, most research concerning capacity allocation mechanisms for flexible resources supposes that the capacity of a flexible resource can be shared between multiple products; that is, flexible resources are divisible. This type of (divisible) resource flexibility – also referred to as "process flexibility" or "product mix flexibility" (Sethi and Sethi, 1990) – is commonly encountered in flexible plants/assembly lines that can produce multiple products at the same time (see, for instance, Bish and Wang, 2004; Chod and Rudi, 2005; Fine and Freund, 1990; Van Mieghem, 1998) as well as in environments where a higher

[4] An example in the United States is www.fandango.com, which sells movie tickets online.

value resource/inventory can also be used to satisfy the demand for a lower level resource/inventory, so that the total capacity of each resource can be split between different demands; see, for instance, Netessine et al. (2002).

To our knowledge, there is very limited research that studies flexible resources that are also indivisible, with a few exceptions that focus on the aircraft swapping problem in the airline industry. In particular, most of this research studies algorithmic approaches on *how* to swap a *given pair of aircraft* initially assigned to a pair of flights, while conserving the flow balance in the flight network; see, for instance, Talluri (1996) for the aircraft swapping problem, and Sherali et al. (2006) and the references cited therein for a comprehensive review on airline fleeting and swapping approaches. From a different perspective, Bish et al. (2004) study the benefits of several demand driven aircraft swapping mechanisms characterized in terms of their *frequency* (how often the swapping decision should be revised), but under the assumption that demand parameters are known with certainty at the outset. In summary, the existing research does not study the issue of how to manage indivisible flexible resources, which includes decisions regarding whether to accept each arriving customer and which resource to allocate to each demand stream. These operational issues are the focus of this paper.

Our objective is to obtain managerial insights and guidelines on how the capacity allocation decision should be managed. As such, our model is rather stylized and generic, with only two resources and two demand streams. We recognize that for each application area, additional constraints need to be imposed in order for the model to produce practical results. For instance, in the case of a multiplex movie theater, there are typically more than two screens, and the sizes of the screens and their sound systems may be different. In addition, there are logistics issues concerning switching the films quickly between screens. Nevertheless, we hope that our work demonstrates the value of allocation postponement, and will spur interest from industries to invest in technologies that make allocation postponement possible for their respective areas.

The remainder of this paper is organized as follows. In Section 2, we present our assumptions and the demand forecast framework. In Section 3, we consider a *single-decision setting*, in which after the manager rejects a customer, he has to reject all future customers who request the same type of service. In Section 4, we extend our analysis to a *repeated-decision setting*, in which the manager can accept any customer as long as the capacity allows. In Section 5, we perform numerical experiments to study the value of these two postponement strategies. Finally in Section 6, we offer some discussion on our model, the assumptions we make, and the implications of these assumptions.

2. Assumptions and demand forecast framework

We consider a manager who uses two flexible resources – with respective capacities of C_1 and C_2, $C_1 < C_2$ – to satisfy two types of customers (we use "customer" as the generic term for "demand") who arrive over the selling season $[0, T]$. Type i, $i = 1, 2$, customers arrive according to a Poisson process whose rate can only be forecasted, with possible errors, at time 0. Denote the arrival rate of type i customers by Λ_i, $i = 1, 2$, and assume they are independent random variables.

Resources are *indivisible* in the sense that the entire capacity of each resource must be allocated to the same type of customers. However, resources are *flexible* in the sense that each resource can be allocated to either type of customers. Furthermore, resources are *perishable* because any resource capacity not used at the end of the selling season is lost. When a customer arrives, the manager needs to decide immediately whether to accept or reject the customer, *with the constraint that at time T all accepted customers must be satisfied with the service type that they require*. We assume that there is no *consumer-driven substitution*, i.e., a consumer who is not accepted for her service type will not switch to the other service type. We also assume that there is no cancelation, i.e., each accepted consumer will show up for service at the end of the selling season; hence, there is no need for overbooking. We discuss the implications of these assumptions in Section 6.

The goal of the manager is to utilize the early sales figures to better understand both demand arrival rates and make decisions (i.e., whether to accept or reject each arriving customer) in order to maximize the expected revenue when the sale ends at time T. While doing this, the manager may utilize the resource flexibility in the system, which allows the resource allocation decision (i.e., which resource is assigned to which demand stream) to be postponed to a later time during the selling season. Observe that in practice, resource allocation decision may have to be made by a certain time, $T - \tau$, for some $0 < \tau < T$, although customer demands may continue to be accepted in the interval $(T-\tau, T]$. The cutoff point, $T-\tau$, depends on system constraints as well as managerial decisions. For instance, in a multiplex, technological constraints dictate that the allocation decision has to be made by a certain time to allow for the set-up of the film to be shown. In our model, we consider $\tau = 0$. It is possible to perform a similar analysis for different values of τ; obviously, the expected revenue benefit of a postponed capacity allocation policy should be nonincreasing in τ. In addition, each choice of τ may have different costs, such as the cost related to the loss of customer goodwill (due to the delayed resource allocation decision). Such cost parameters will be industry specific; in addition, the values of these cost parameters is a question better answered empirically. In order to keep our analysis free of such cost parameter estimations, we consider the revenue side. The trade-off between costs and revenues can then be analyzed by an empirical study in the context of each possible application. Furthermore, in order to simplify the exposition, we assume that the prices are the same for both service types (same-price policies are common in the movie theater business; moreover, our analysis can be easily extended to the case where prices are different), so that maximizing the expected total revenue is equivalent to maximizing the expected total number of sales.

At time 0, the manager can estimate the first two moments of the arrival rates Λ_1 and Λ_2. We use a gamma distribution to describe the prior of the customer arrival rate, because gamma distribution has a flexible form and can represent a wide variety of functional forms depending on the values of its two parameters (k_i, a_i) (e.g., for $k_i = 1$, it reduces to the exponential distribution, and for k_i integer, it is the Erlang distribution), and it has positive support, which is appropriate for an arrival rate distribution. Specifically, we assume that before the sale begins, the arrival rate of type i customers follows a gamma distribution with parameters (k_i, a_i) with the probability density function

$$f_{\Lambda_i}(\lambda) = \frac{a_i e^{-a_i\lambda}(a_i\lambda)^{k_i-1}}{\Gamma(k_i)} \quad \text{for } \lambda \geqslant 0,$$

where $a_i > 0$ and $k_i > 0$, for $i = 1, 2$. With a gamma distribution, the first two moments of Λ_i, $i = 1, 2$, are given by

$$E[\Lambda_i] = \frac{k_i}{a_i} \text{ and } \mathrm{Var}(\Lambda_i) = \frac{k_i}{a_i^2}. \tag{1}$$

Let $\{N_i(t), 0 \leqslant t \leqslant T\}$ denote the arrival process of type i customers, with $N_i(t)$ denoting the number of type i customers arriving up to time t.

As discussed above, under the *no postponement* policy the manager makes the capacity allocation decision at time 0 in order to maximize the expected sales, whose expression is given by[5]

$$\begin{aligned}\max\{&E[\min\{N_1(T), C_1\}\\&+\min\{N_2(T), C_2\}], E[\min\{N_1(T), C_2\}\\&+\min\{N_2(T), C_1\}]\}.\end{aligned} \tag{2}$$

Next, we discuss how demand forecasts are updated in a postponement setting. If j type i customers arrive in $[0, t]$, we can use Bayes' rule to determine the posterior density function of Λ_i, $i = 1, 2$, as follows:

$$f_{\Lambda_i|N_i(t)=j}(\lambda) = \frac{f_{\Lambda_i}(\lambda)e^{-\lambda t}\frac{(\lambda t)^j}{j!}}{\int_0^\infty f_{\Lambda_i}(\lambda)e^{-\lambda t}\frac{(\lambda t)^j}{j!}\,d\lambda}. \tag{3}$$

The denominator of the preceding can further be computed as follows:

$$\begin{aligned}\int_0^\infty f_{\Lambda_i}(\lambda)e^{-\lambda t}\frac{(\lambda t)^j}{j!}\,d\lambda &= \frac{a_i^{k_i}\, t^j}{j!\Gamma(k_i)}\int_0^\infty \lambda^{k_i+j-1}\ e^{-\lambda(a_i+t)}\,d\lambda\\&= \frac{a_i^{k_i}\, t^j}{j!\Gamma(k_i)}\,\frac{\Gamma(k_i+j)}{(a_i+t)^{k_i+j}},\end{aligned}$$

where the last equality follows from the definition of the gamma function, given by $\Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}\,dx$, for $\alpha > 0$. Substituting the preceding into Eq. (3) yields

$$f_{\Lambda_i|N_i(t)=j}(\lambda) = \frac{(a_i+t)e^{-(a_i+t)\lambda}((a_i+t)\lambda)^{k_i+j-1}}{\Gamma(k_i+j)}, \tag{4}$$

which is a gamma density function with parameters $(k_i + j, a_i + t)$. For $i = 1, 2$, define

$$M_i(t,j) \equiv N_i(T) - N_i(t)|N_i(t) = j$$

as the additional number of type i customers that will arrive in $(t, T]$ conditional on that j type i customers have arrived in $[0, t]$. Using Eq. (4), we can calculate the probability distribution of $M_i(t,j)$. For $n = 0, 1, \ldots,$

[5] It is straightforward to compute the expected sales in Eq. (2) because $N_i(T)$, $i = 1, 2$, follows a negative binomial distribution, as will be shown in Eq. (5) below (its probability mass function can be obtained by substituting $j = 0$ and $t = 0$ in Eq. (5)).

$$\begin{aligned}Pr\{M_i(t,j)=n\} &= Pr\{N_i(T)-N_i(t)=n|N_i(t)=j\}\\ &= \int_0^\infty Pr\{N_i(T)-N_i(t)\\ &\quad =n|N_i(t)=j,\Lambda_i=\lambda\}f_{\Lambda_i|N_i(t)=j}(\lambda)\,\mathrm{d}\lambda\\ &= \int_0^\infty \frac{\mathrm{e}^{-(T-t)\lambda}((T-t)\lambda)^n}{n!}\\ &\quad\times\frac{(a_i+t)\mathrm{e}^{-(a_i+t)\lambda}((a_i+t)\lambda)^{k_i+j-1}}{\Gamma(k_i+j)}\,\mathrm{d}\lambda\\ &= \frac{\Gamma(n+k_i+j)}{n!\Gamma(k_i+j)}\left(\frac{a_i+t}{a_i+T}\right)^{k_i+j}\left(\frac{T-t}{a_i+T}\right)^n,\end{aligned} \tag{5}$$

where the last equality follows by the definition of a gamma function. In other words, $M_i(t,j)$ follows a negative binomial distribution having parameters $k_i + j$ and $(a_i + t)/(a_i + T)$.

3. Single-decision setting

A single-decision policy is a policy under which the manager must make a capacity allocation decision when $\max\{N_1(t), N_2(t)\}$ reaches $(C_1 + 1)$ for the first time (as long as this event occurs before time T), and cannot reverse this decision later. (Observe that as long as $N_1(t)$ and $N_2(t)$, $t \in [0, T]$, are both smaller than or equal to C_1, it is clearly optimal to accept any arriving customer without having to commit a resource type to a customer type.) In this case, the manager either accepts this arriving customer and assigns type 2 resource ($C_2 > C_1$) to this type of customers, or rejects the arriving customer and assigns type 2 resource to the other type of customers. In either case, no further decision needs to be made, and the manager simply accepts all subsequent customers of each type until the assigned capacity limit is reached. That is, the single-decision class of policies are subject to the constraint:

(**A1**) Once a customer is rejected, no future customers of the same type can be accepted.

Theorem 1 shows that we can find the optimal policy in this class of policies, which we call the *single-decision optimal* (SDO) policy.

A single-decision policy is suitable when customers line up in front of the store to purchase their services, and thus, are aware of the status of the sale. A typical example for this situation is the movie theater business. While early sales may take place over the phone or on the Internet, as the show times approach, people will line up and buy their tickets from the box office. In this setting, a rejected customer may feel being unfairly treated if the manager accepts a later customer of the same type. Obviously, there are systems where the relaxation of this assumption is appropriate, and we will discuss such systems in detail in Section 4.

If $E[\Lambda_1] \geqslant E[\Lambda_2]$, then a seemingly intuitive policy is to assign type 2 resource ($C_2 > C_1$) to type 1 customers if, for some $t < T$, $N_1(t)$ reaches $C_1 + 1$ before $N_2(t)$ does. However, the next example shows that this is not necessarily optimal, as the variability of the arrival rate also plays an important role in the optimal policy.

Example. Suppose $C_1 = 1$, $C_2 = 100$, and $T = 10$. Let $k_1 = 101$, $a_1 = 100$, $k_2 = a_2 = 1$ so that $E[\Lambda_1] = 1.01 > 1 = E[\Lambda_2]$. On the other hand, $\mathrm{Var}[\Lambda_1] = 0.0101 < 1 = \mathrm{Var}[\Lambda_2]$. Suppose that the second type 1 customer arrives at time $t = 0.1$ – by which only one type 2 customer has arrived. According to Eq. (5), $M_1(0.1, 2)$ follows a negative binomial distribution with parameters $(103, \frac{100.1}{110})$, whose expectation is equal to $E[M_1(0.1, 2)] = 103(110/100.1 - 1) \approx 10.19$. Similarly, $M_2(0.1, 1)$ follows a negative binomial distribution with parameters $(2, 0.1)$, with $E[M_2(0.1, 1)] = 2(1/0.1 - 1) = 18$. Therefore, it is better for the manager to reject the second type 1 customer and assign type 2 resource to type 2 customers. Because the variance of Λ_2 is much larger than that of Λ_1, assigning type 2 resource ($C_2 = 100$) to type 2 customers gives the manager a better chance to use more of type 2 resource.

This example illustrates how variability in the arrival rate (in addition to its expected value) plays an important role in determining an optimal resource allocation. The optimal allocation is determined by considering the trade-off between the risk of unsatisfied demand versus underutilized capacity, both of which depend on the mean and the variance of the arrival rate. Thus, a greedy approach that assigns capacities based on an ordering of means is not necessarily optimal. Nevertheless, we can still characterize the structure of an optimal solution and show that it is of a threshold type. For this purpose, we briefly review the definition of stochastic orders; see, for instance, pages 404–405 in Ross (1996). For two random variables X and Y, we say X is stochastically larger than Y, written $X \geqslant_{\mathrm{st}} Y$, if

$$Pr\{X > t\} \geqslant Pr\{Y > t\} \quad \text{for all } t.$$

It is straightforward to check that a negative binomial random variable increases stochastically in its first parameter, and decreases stochastically in its second parameter. Then, from Eq. (5), it follows that $M_i(t,j)$ increases stochastically in j and decreases stochastically in t.

Theorem 1. *Let $\{N_i(t),\ 0 \leqslant t \leqslant T\}$ denote the arrival process of type i customers, with arrival rate Λ_i, $i = 1,2$, and assume that each Λ_i, $i = 1,2$, follows an independent gamma distribution with parameters $k_i > 0$ and $a_i > 0$. In the single-decision setting, a threshold-type policy suffices to be optimal. That is, if the $(C_1 + 1)$st type 1 (or 2) customer arrives before the $(C_1 + 1)$st type 2 (or 1) customer, say at time t, then it is optimal to accept the arriving customer if $N_2(t) \leqslant h_2(t)$ (or $N_1(t) \leqslant h_1(t)$), where $h_i(t)$, $i = 1,2$, are the threshold functions.*

Proof. Suppose that the $(C_1 + 1)$st customer of type 1 arrives at time t, at which point the number of type 2 customers, denoted by b_2, is smaller than or equal to C_1. The expected total number of sales if the manager accepts the arriving type 1 customer and allocates the larger capacity C_2 to type 1 customers is equal to

$$C_1 + 1 + E[\min(M_1(t, C_1 + 1), C_2 - C_1 - 1)] + b_2 + E[\min(M_2(t, b_2), C_1 - b_2)], \tag{6}$$

while the expected total number of sales if the manager rejects the arriving type 1 customer and allocates the larger capacity C_2 to type 2 customers is equal to

$$C_1 + b_2 + E[\min(M_2(t, b_2), C_2 - b_2)]. \tag{7}$$

The optimal expected sales in the single-decision setting is given by the maximum of Eqs. (6) and (7). To show that the optimal policy is of a threshold type, it is sufficient to show that the difference in expected number of sales between rejecting and accepting increases in b_2. In other words, we need to show that the difference between Eqs. (7) and (6)

$$E[\min(M_2(t, b_2), C_2 - b_2)] - E[\min(1 + M_1(t, C_1 + 1), C_2 - C_1)] - E[\min(M_2(t, b_2), C_1 - b_2)] \tag{8}$$

increases in b_2. To do so, consider a sequence of inequalities

$$\begin{aligned} &E[\min(M_2(t, b_2), C_2 - b_2) - \min(M_2(t, b_2), C_1 - b_2)] \\ &\quad \leqslant E[\min(M_2(t, b_2 + 1), C_2 - b_2) - \min(M_2(t, b_2 + 1), C_1 - b_2)] \\ &\quad \leqslant E[\min(M_2(t, b_2 + 1), C_2 - b_2 - 1) - \min(M_2(t, b_2 + 1), C_1 - b_2 - 1)]. \end{aligned}$$

The first inequality follows because for $y > z$, $\min(x,y) - \min(x,z)$ increases in x and that $M_2(t, b_2 + 1)$ is stochastically larger than $M_2(t,b_2)$. The second inequality follows because for $y > z$, $\min(x,y) - \min(x,z) \leqslant \min(x,y-1) - \min(x,z-1)$. Therefore, Eq. (8) increases in b_2 and the proof is completed. □

Because $M_i(\cdot,\cdot)$, $i = 1,2$, are negative binomial random variables whose probability mass functions can be calculated according to Eq. (5), Eqs. (6) and (7) can be straightforwardly evaluated. Consequently, the optimal policy in the single-decision setting can be calculated explicitly, as will be done numerically in Section 5.

Proposition 1. *Let $\{N_i(t), 0 \leqslant t \leqslant T\}$ denote the arrival process of type i customers, with arrival rate Λ_i, $i = 1,2$, and assume that each Λ_i, $i = 1,2$, follows an independent gamma distribution with parameters $k_i > 0$ and $a_i > 0$. If $k_1 \geqslant k_2$ and $a_1 \leqslant a_2$, then it is always optimal to accept the $(C_1 + 1)$st type 1 customer as long as type 2 resource has not been assigned to type 2 customers. In other words, $h_2(t) = C_1$ for $t \in [0,T]$.*

Proof. Suppose that the $(C_1 + 1)$st customer of type 1 arrives at time t, at which point the number of type 2 customers, denoted by b_2, is smaller than or equal to C_1. Because a negative binomial distribution increases stochastically in its first parameter and decreases stochastically in its second parameter, we have that

$$M_1(t, C_1 + 1) \geqslant_{\mathrm{st}} M_1(t, b_2) \geqslant_{\mathrm{st}} M_2(t, b_2),$$

where the first inequality follows because $C_1 + 1 > b_2$, and the second follows because $k_1 \geqslant k_2$ and $a_1 \leqslant a_2$.

The expected number of sales if the manager accepts the $(C_1 + 1)$st type 1 customer is equal to

$$\begin{aligned} &C_1 + E[\min(1 + M_1(t, C_1 + 1), C_2 - C_1)] + b_2 + E[\min(M_2(t, b_2), C_1 - b_2)] \\ &\quad \geqslant C_1 + E[\min(M_2(t, b_2), C_2 - C_1)] + b_2 + E[\min(M_2(t, b_2), C_1 - b_2) \\ &\quad \geqslant C_1 + b_2 + E[\min(M_2(t, b_2), C_2 - b_2)], \end{aligned}$$

where the first inequality follows because $M_1(t, C_1 + 1)$ is stochastically larger than $M_2(t, b_2)$, and the second follows from the identity that $\min(x,y) + \min(x,z) \geqslant \min(x,y + z)$ for any three numbers x, y, and z. The proposition then follows because the last equation represents the expected number of sales if the manager rejects the arriving type 1 customer and assigns the larger capacity C_2 to type 2 customers. □

Corollary 1. *Let $\{N_i(t), 0 \leqslant t \leqslant T\}$ denote the arrival process of type i customers, with arrival rate Λ_i, $i = 1, 2$, and assume that each Λ_i, $i = 1, 2$, follows an independent gamma distribution with parameters $k_i > 0$ and $a_i > 0$. Suppose that Λ_1 and Λ_2 have the same coefficient of variation, then if $E[\Lambda_1] \geqslant E[\Lambda_2]$, it is always optimal to accept the $(C_1 + 1)$st type 1 customer as long as type 2 resource has not been assigned to type 2 customers.*

Proof. The condition is equivalent to

$$\frac{\sqrt{k_1/a_1^2}}{k_1/a_1} = \frac{\sqrt{k_2/a_2^2}}{k_2/a_2} \text{ and } \frac{k_1}{a_1} \geqslant \frac{k_2}{a_2},$$

which is equivalent to $k_1 = k_2$ and $a_1 \leqslant a_2$. The result then follows from Proposition 1. □

While our numerical example demonstrates that, when $E[\Lambda_1] > E[\Lambda_2]$, it is not always optimal to assign the larger capacity to type 1 customers even if its number reaches $C_1 + 1$ first, Proposition 1 establishes a dominance relationship between Λ_1 and Λ_2 when the preceding is indeed the case. In practice, coefficients of variation of the arrival rate measure the quality of the demand forecast, so it is often reasonable to assume that they are the same (or close) for different demand types. In this case, Corollary 1 guarantees that if $E[\Lambda_1] > E[\Lambda_2]$, then it is always optimal to accept the $(C_1 + 1)$st type 1 customer as long as the larger capacity has not been assigned to type 2 customers.

4. Repeated-decision setting

In this section, we consider a setting in which an earlier rejection of a customer does not prevent the manager from accepting a later customer of the same type (i.e., we relax Assumption **A1**). Consequently, this setting allows the manager to postpone the capacity commitment time beyond that in the single-decision setting (which happens if the manager chooses to reject the first $(C_1 + 1)$st customer of either type). Observe that the allocation decision can no longer be postponed when $(C_1 + 1)$ customers of a type is admitted. This policy is suitable when the sale takes place exclusively over the phone or on the Internet, because a rejected customer will not learn about whether a later customer of the same type is accepted for service.

Because the manager can postpone the commitment decision when he rejects a customer, it seems suitable to use continuous-time dynamic programming to formulate the problem. However, such a dynamic programming formulation is not mathematically tractable because customer arrival processes do not have independent increments – customer arrival rates Λ_1 and Λ_2 are random variables whose posterior probability distributions depend not only on the number of customers arrived but also on the time elapsed. For that reason, the state transition density function becomes too complicated, and it is not likely that we can derive the optimal policy from the optimality equation. Hence, we develop a heuristic policy in which the manager *repeatedly* applies the capacity allocation rule of the single-decision setting (given in Theorem 1) until he admits the $(C_1 + 1)$st customer of either type (hence determining the capacity allocation decision).

More specifically, as in the single-decision setting, if both $N_1(t)$ and $N_2(t)$ are smaller than or equal to C_1, then it is obviously optimal to accept either type of customer. The manager needs to make a nontrivial decision for a type i customer arriving at time t only if the current number of bookings of type i customers is equal to C_1, while that of the other type is smaller than or equal to C_1, for $i = 1, 2$. As stated above, at these decision epochs, we propose the following heuristic policy, which accepts an arriving type i customer, if assigning the larger capacity to type i customers yields a higher expected total sales than assigning it to the other type of customers. In other words, with this heuristic policy, the manager accepts the arriving type $i, i = 1, 2$, customer at time t if

$$\begin{aligned}
1 &+ E[\min(M_i(t, N_i(t)), C_2 - C_1 - 1)] \\
&+ E[\min(M_{3-i}(t, N_{3-i}(t)), C_1 - N_{3-i}(t))] \\
&\geqslant E[\min(M_{3-i}(t, N_{3-i}(t)), C_2 - N_{3-i}(t))];
\end{aligned}$$

otherwise, the manager rejects the arriving type i customer. In practice, the manager may need to apply this comparison several times before finally accepting a new customer to complete the capacity allocation decision.

In what follows, we refer to this heuristic policy as the *repeated-decision heuristic* (RDH) policy. To evaluate the performance of the RDH policy, which we will do numerically in Section 5, we next derive an upper bound on the expected total number of sales in any policy.

Proposition 2. *Let* $\{N_i(t), 0 \leqslant t \leqslant T\}$ *denote the arrival process of type i customers, with arrival rate* Λ_i, $i = 1, 2$, *and assume that each* Λ_i, $i = 1, 2$, *follows an independent gamma distribution with parameters* $k_i > 0$ *and* $a_i > 0$. *An upper bound on the expected sales in any policy is given by*

$$\begin{aligned}&E[\min\{\min\{N_1(T), N_2(T)\}, C_1\}\\&\quad + \min\{\max\{N_1(T), N_2(T)\}, C_2\}]\\&\quad = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\{(\min(C_1, \min(j,k))\\&\quad + \min(C_2, \max(j,k)))Pr(N_1(T) = j)\\&\quad \times Pr(N_2(T) = k)\}.\end{aligned}$$

Proof. The above expression is an upper bound on the expected sales, because it is what can be achieved if the capacity allocation can be decided *after* the realization of the random demands. □

5. Numerical experiments

In this section, we present the numerical experiments. For each numerical experiment, we determine the average number of sales in four strategies: (1) no decision postponement, (2) the SDO policy in the single-decision setting, (3) the RDH policy in the repeated-decision setting, and (4) the upper bound in Proposition 2. Our numerical experiments are designed to answer the following questions.

1. What is the value of postponing the capacity allocation decision?
2. How effective is the heuristic policy in the repeated-decision setting?
3. What is the additional value of moving from the single-decision setting to the repeated-decision setting?

In our numerical experiments, we consider $C_1 = 50$, $a_1 = a_2 = 1$, $k_1 = 10$, $T = 5$, and vary the values of k_2 and C_2, over 10–20 and 60–150, respectively. Recall that for a given a_2, as k_2 increases, both the expected value and variance of Λ_2 increase (see Eq. (1)). In each setting, we simulate 10,000 replications so that the standard errors are within ±0.001% of the estimators. Specifically, for each problem instance, we generate a sample path of demand arrival times and dynamically update the arrival rate forecasts based on the demand realized over time. When considering the single-decision setting, upon arrival of the first $(C_1 + 1)$st customer of either type, we calculate the required threshold function, $h_i(t)$, $i = 1$ or 2, $t \in [0, \ldots, T]$, and use the optimal policy in Theorem 1 to decide whether or not to accept the current arriving customer. Our numerical results are reported in Tables 1–3 and Figs. 1–4; all tables and figures are relegated to the Appendix.

Our numerical study leads to the following insights.

1. *The value of postponement.* We compare the average sales of no decision postponement and postponement in the single-decision setting (see Table 1). The benefit of the single-decision setting varies depending on the parameters, and reaches its highest when Λ_1 and Λ_2 have the same distribution and when $C_2 \gg C_1$. Intuitively, when Λ_1 and Λ_2 have the same distribution, the manager who makes the allocation decision at time 0 will make the "right" decision only half of the time. Allowing the allocation decision to be postponed can significantly increase this probability. If Λ_1 and Λ_2 have much different distributions, say $E[\Lambda_1] \gg E[\Lambda_2]$, then assigning the larger capacity to type 1 customers at time 0 may turn out to be the right decision most of the time, so postponing does not provide much benefit. Similarly, when C_1 and C_2 are close, it does not make much difference to postpone the decision. In our numerical experiments, the decision postponement can bring as much as 5% increase in the number of sales. In industries where profit margins are thin, an increase of even 1% in revenue can translate into a large increase in profit.
2. *The effectiveness of the RDH policy in the repeated-decision setting.* In all numerical experiments tested, we find that the expected sales from the RDH policy is at least 99.88% of the upper bound (see Table 2). Not only does this observation imply that our heuristic policy is very effective, it also suggests that the optimal policy in

the repeated-decision setting – even if we could determine it – would not improve much beyond our heuristic policy.

3. *The difference between the single- and repeated-decision settings.* Our numerical experiments suggest that the performance of the SDO policy and the RDH policy are very similar. In all cases tested, the SDO policy generates at least 99.97% of the revenue of the RDH policy (see Table 3). This observation suggests that the simple single-decision rule contains most of the benefits from decision postponement, and can be very effective to hedge against forecast error and demand variability.

In addition, our numerical study indicates the following.

- For a given capacity C_2, the average sales for all policies is concave increasing in k_2 (see Fig. 1). This is because as k_2 increases, not only the expected value of Λ_2 increases, but also its variability increases. In addition, since capacity is fixed, demand in excess of capacity cannot be accepted and does not affect sales. As a result, the average sales function exhibits diminishing returns as k_2 increases.
- For a given k_2, the average sales for all policies is concave increasing in C_2 (see Fig. 2). Intuitively, a very high C_2 (compared to expected demands) does not necessarily translate into a higher sales, since it will not find enough demand to satisfy.
- For a given capacity C_2, the percent increase in average sales in both the single- and repeated-decision settings and the upper bound (over no postponement) is convex decreasing in k_2 (see Fig. 3). As k_2 increases, the expected values of the two demand streams become more different (i.e., $E[\Lambda_2] - E[\Lambda_1]$ increases). However, this comes at the expense of an increased variability ($\mathrm{Var}(\Lambda_2) > \mathrm{Var}(\Lambda_1)$), which explains the convex decreasing part.
- For a given k_2, the percent increase in average sales in both the single- and repeated-decision settings and the upper bound (over no postponement) is concave increasing in C_2 (see Fig. 4). This again is because a very high C_2 does not necessarily translate into a higher sales. In addition, as C_2 increases, the average sales in the no postponement policy increases further, and the same deviation from it will give a smaller percent change, and hence the concavity.

6. Discussion, conclusions, and future research directions

In this paper, we study the benefit of an indivisible resource flexibility structure under demand forecast uncertainty and demand variability. Resource flexibility allows the revenue manager to delay the resource allocation decision to a time when more information on the demand distributions is gathered and demand uncertainty is reduced. Considering a simple two-resource two-demand-type model with forecast error, we characterize the structure of the optimal delayed resource allocation policy. Our findings suggest that a simple threshold policy, which consists of at most one decision epoch, can be quite effective in hedging against demand forecast error and variability. Its revenue benefit can be significant, especially when demand rate forecasts are close and resource capacities are much different.

Our model is rather stylized and generic, with only two resources and two demand streams, and our results come with some limitations, as we make some simplifying assumptions such as no cancelation, no consumer-driven substitution, and exogenous pricing. Below we discuss the implications of these three main assumptions.

1. *No cancelation.* We assume that all customers who have purchased their tickets will show up at the time of service. In reality, cancelations do occur occasionally over time, and canceling customers may either get no refund, partial refund, or full refund, depending on when they cancel. In any case, including cancelations in our model will only increase the value of capacity postponement, because the firm can switch its capacity allocation back and forth based on the real-time demand taking into account cancelations.
2. *No consumer-driven substitution.* We assume that a consumer does not switch to a different service type if her desired service type is unavailable. In reality, this assumption need not always be true. For instance, a person may be willing to see another movie if her top choice is sold out. With consumer-driven substitution, the no postponement strategy would be able to recapture more demand (that would have been lost otherwise) than the two postponement strategies, because the postponement strategies are designed to recapture some lost demand in the first place. The postponement strategies would still generate

more revenue, but the difference would be less. Nevertheless, the postponement strategies would allow more customers to receive their most preferred service than the no postponement strategies, whose value is more difficult to quantify.

3. *Exogenous pricing.* In our model, we assume that the prices are exogenously determined. In reality, the firm may have some control over the price, and can use it to its advantage to better fit demands to capacities. If the demand is not elastic in its price, then our model can be a good approximation; otherwise, the relation between price and demand needs to be specifically modeled, which is beyond the scope of this paper. Another interesting extension is to allow a firm to change its product in real time. Such practice has the potential to increase the revenue even more.

As a future research direction, it would be interesting to relax the above three assumptions as well as analyze the case with multiple (>2) resources and demand streams.

We next discuss the implications of our findings to other service industries that have different types of resource flexibility structure. For example, consider the assignment of tourist guides to tour groups, nurses to patients, lawyers or consultants to clients, operating rooms in a hospital to patients requiring surgery, etc. In all these examples, resources (tourist guides, nurses, etc.) have different specialties, and each customer (tour groups, patients, etc.) has different requirements for resources. For instance, a Japanese tour group may require a tourist guide who is fluent in Japanese. Thus, some resources may be *flexible* in that they may possess skills required by different types of customers. However, resources are *not indivisible* as discussed here, since all tour guides of a given type (e.g., who are fluent in Japanese) need not be assigned to the same tour group. These situations have been studied in the literature as discussed in Section 1, and are different from our model. Our approach provides a different perspective on resource flexibility and contributes to this research area in a broad sense.

Finally, in this paper, we consider only the revenue side of the decision postponement. In practice, however, there are several costs associated with postponing the allocation decision, such as the loss of goodwill, the monetary cost of changing the current assignment, and the risk of not completing the swap on time. Taking into account these costs would provide a comprehensive cost-benefit analysis. However, these costs are difficult to quantify in practice, and additional research is needed to understand these issues.

Acknowledgments

We thank four anonymous referees and the Editor for their insightful and constructive comments that have improved our discussion and the presentation of the results.

Appendix

The appendix contains all tables and figures (see Tables 1–3 and Figs. 1–4).

Table 1
Average percent deviation of the upper bound, the RDH policy, and the SDO policy over the no decision postponement case (with two decimal-point accuracy)

$E[\Lambda_2]$	10	11	12	14	16	18	20
$C_2\vert k_2$	10	11	12	14	16	18	20
60	2.51, 2.49, 2.49	2.06, 2.04, 2.04	1.55, 1.54, 1.54	0.86, 0.84, 0.84	0.39, 0.38, 0.38	0.16, 0.15, 0.15	0.05, 0.05, 0.05
70	4.10, 4.05 , 4.04	3.41, 3.37, 3.36	2.70, 2.65, 2.65	1.53, 1.48, 1.48	0.78, 0.74, 0.74	0.36, 0.33, 0.33	0.14, 0.12, 0.12
80	4.86, 4.81, 4.79	4.11, 4.06, 4.04	3.34, 3.29, 3.26	1.97, 1.89, 1.89	1.06, 0.97, 0.97	0.55, 0.49, 0.48	0.24, 0.20, 0.20
90	5.19, 5.13, 5.12	4.44, 4.36, 4.36	3.63, 3.58, 3.54	2.19, 2.09, 2.08	1.22, 1.11, 1.11	0.66, 0.58, 0.58	0.30, 0.24, 0.24
100	5.35, 5.28, 5.27	4.58, 4.51, 4.49	3.74, 3.66, 3.64	2.28, 2.16, 2.16	1.28, 1.17, 1.17	0.71, 0.64, 0.62	0.33, 0.27, 0.26
110	5.41, 5.35, 5.33	4.63, 4.54, 4.54	3.77, 3.68, 3.68	2.31, 2.19, 2.19	1.30, 1.19, 1.18	0.73, 0.66, 0.63	0.34, 0.29, 0.27
120	5.43, 5.35, 5.35	4.64, 4.56, 4.56	3.79, 3.69, 3.69	2.32, 2.20, 2.20	1.31, 1.19, 1.18	0.73, 0.66, 0.64	0.34, 0.29, 0.27
130	5.44, 5.36, 5.36	4.65, 4.56, 4.56	3.79, 3.69, 3.69	2.32, 2.21, 2.21	1.31, 1.19, 1.19	0.73, 0.67, 0.64	0.34, 0.29, 0.27
140	5.44, 5.36, 5.36	4.66, 4.57, 4.57	3.79, 3.69, 3.69	2.32, 2.21, 2.21	1.31, 1.19, 1.19	0.73, 0.67, 0.64	0.34, 0.29, 0.27
150	5.44, 5.36, 5.36	4.66, 4.57, 4.57	3.79, 3.69, 3.69	2.32, 2.21, 2.21	1.31, 1.19, 1.19	0.73, 0.67, 0.64	0.34, 0.29, 0.27

$C_1 = 50$, $a_1 = a_2 = 1$, $k_1 = 10$, $T = 5$.

Table 2
Ratios of the average sales of (i) the RDH policy to the upper bound, and (ii) the SDO policy to the upper bound (with four decimal-point accuracy)

$E[\Lambda_2]$ $C_2\|k_2$	10 10	11 11	12 12	14 14	16 16	18 18	20 20
60	0.9998	0.9998	0.9998	0.9998	0.9999	0.9999	1.0000
70	0.9995	0.9995	0.9995	0.9995	0.9995	0.9997	0.9998
80	0.9993	0.9993	0.9992	0.9991	0.9991	0.9994	0.9996
90	0.9993	0.9992	0.9991	0.9990	0.9989	0.9992/0.9991	0.9994
100	0.9992	0.9992	0.9991	0.9989	0.9988	0.9992/0.9991	0.9993
110	0.9992	0.9992	0.9991	0.9989	0.9988	0.9993/0.9991	0.9994/0.9993
120	0.9992	0.9991	0.9990	0.9989	0.9988	0.9993/0.9991	0.9995/0.9993
130	0.9992	0.9991	0.9991	0.9989	0.9988	0.9994/0.9991	0.9995/0.9993
140	0.9992	0.9991	0.9990	0.9989	0.9988	0.9994/0.9991	0.9995/0.9993
150	0.9992	0.9991	0.9990	0.9989	0.9988	0.9994/0.9991	0.9995/0.9993

When the two ratios are the same, they are reported by one number; when they are different, they are reported in the form of x/y, where x corresponds to the first ratio, and y to the second ratio; $C_1 = 50$, $a_1 = a_2 = 1$, $k_1 = 10$, $T = 5$.

Table 3
Ratio of the average sales of the SDO policy to that of the RDH policy (with four decimal-point accuracy)

$E[\Lambda_2]$ $C_2\|k_2$	10 10	11 11	12 12	14 14	16 16	18 18	20 20
60	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
70	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
80	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
90	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	1.0000
100	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
110	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9998
120	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9998
130	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9998
140	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9998
150	1.0000	1.0000	1.0000	1.0000	1.0000	0.9997	0.9998

$C_1 = 50$, $a_1 = a_2 = 1$, $k_1 = 10$, $T = 5$.

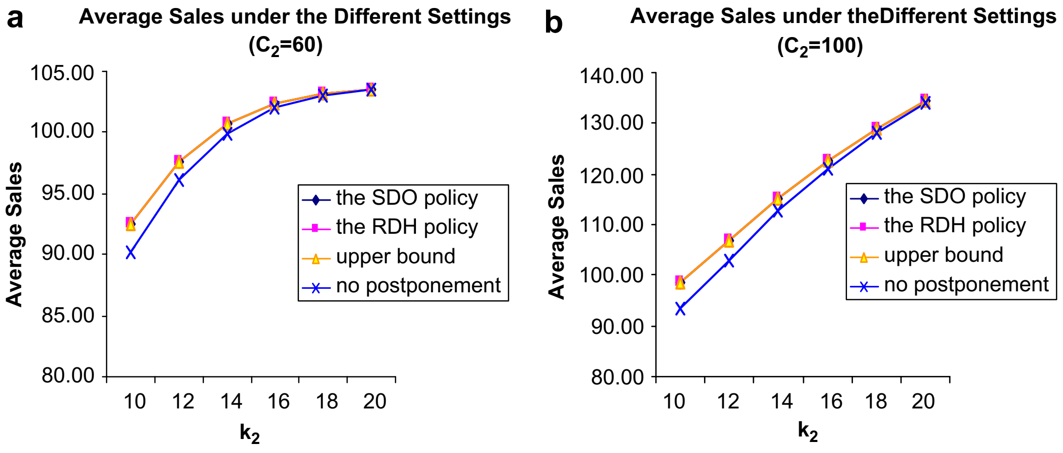

Fig. 1. Average sales under different values of C_2: (a) $C_2 = 60$; (b) $C_2 = 100$.

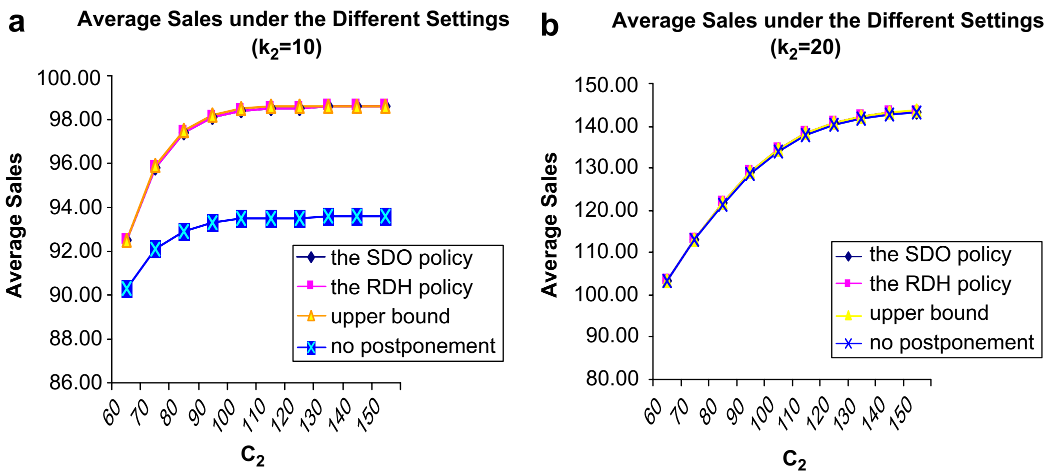

Fig. 2. Average sales under different values of k_2: (a) $k_2 = 10$; (b) $k_2 = 20$.

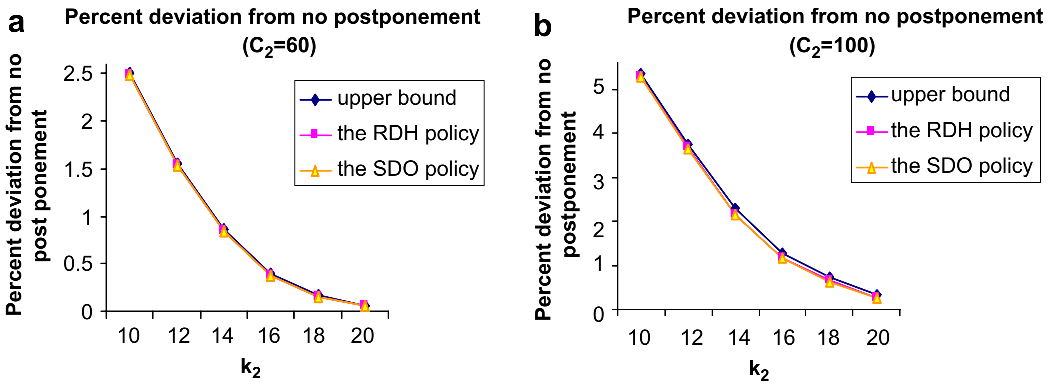

Fig. 3. Percent deviation of the SDO policy, the RDH policy, and the upper bound from no postponement under different values of C_2: (a) $C_2 = 60$; (b) $C_2 = 100$.

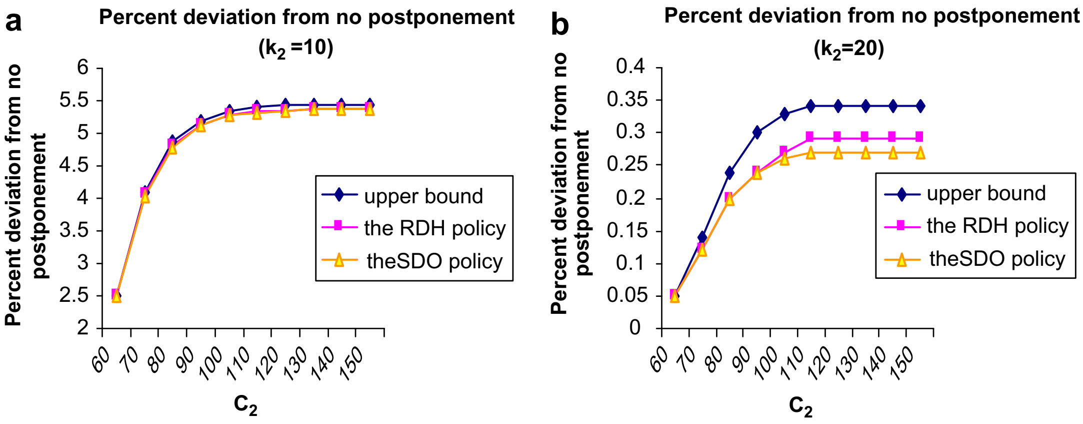

Fig. 4. Percent deviation of the SDO policy, the RDH policy, and the upper bound from no postponement under different values of k_2: (a) $k_2 = 10$; (b) $k_2 = 20$.

References

Bish, E.K., Suwandechochai, R., Bish, D.R., 2004. Strategies for managing the flexible capacity in the airline industry. Naval Research Logistics 51, 654–685.

Bish, E.K., Wang, Q., 2004. Optimal investment strategies for flexible resources, considering pricing and correlated demands. Operations Research 52 (6), 954–964.

Chod, J., Rudi, N., 2005. Resource flexibility with responsive pricing. Operations Research 53, 532–548.

de Kok, A.G., Graves, S.C. (Eds.), 2003. Supply chain management: design, coordination and operation. Handbooks in Operations Research and Management Science 11.

Fine, C.H., Freund, R.M., 1990. Optimal investment in product-flexible manufacturing capacity. Management Science 36, 449–466.

Netessine, S., Dobson, G., Shumsky, R.A., 2002. Flexible service capacity: Optimal investment and the impact of demand correlation. Operations Research 50, 375–388.

Ross, S.M., 1996. Stochastic Processes, second ed. Wiley, New York, NY.

Sethi, A.K., Sethi, S.P., 1990. Flexibility in manufacturing: A survey. The International Journal of Flexible Manufacturing Systems 2, 289–328.

Sherali, H.D., Bish, E.K., Zhu, X., 2006. Airline fleet assignment concepts, models, and algorithms. European Journal of Operational Research 172 (1), 1–30.

Talluri, T.K., 1996. Swapping applications in a daily airline fleet assignment. Transportation Science 30, 237–248.

Tayur, S., Ganeshan, R., Magazine, M., 1999. Quantitative Models for Supply Chain Management. Kluwer Academic Publishers, Boston.

Van Mieghem, J.A., 1998. Investment strategies for flexible resources. Management Science 44, 1071–1078.

Van Mieghem, J.A., 2003. Capacity management, investment and hedging: Review and recent developments. Manufacturing & Service Operations Management 5, 269–302.