2011

# Consolidating or Non-consolidating Queues: A Game Theoretic Queuing Model with Holding Costs

Kwan Eng WEE
*Singapore Management University*, kewee@smu.edu.sg

Ananth Iyer
*Purdue University*



## Citation

WEE, Kwan Eng and Iyer, Ananth. Consolidating or Non-consolidating Queues: A Game Theoretic Queuing Model with Holding Costs. (2011). *Operations Research Letters*. 39, (1), 4-12. Research Collection Lee Kong Chian School Of Business.
**Available at:** https://ink.library.smu.edu.sg/lkcsb_research/3208

Operations Research Letters 39 (2011) 4–10

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

# Consolidating or non-consolidating queues: A game theoretic queueing model with holding costs

Kwan-Eng Wee [a,*], Ananth Iyer [b]

[a] Lee Kong Chian School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore
[b] Krannert School of Management, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

*Article history:*
Received 20 September 2009
Accepted 10 September 2010
Available online 18 December 2010

*Keywords:*
Holding cost allocation
Nash equilibrium
Queueing
Service rate
Incentives
Game theory

ABSTRACT

We consider a two-server queueing system in which the servers choose their service rate based on the demand and holding cost allocation scheme offered by the demand generating entity. We provide an optimal holding cost allocation scheme that leads to the maximum possible service rate for each of a pooled and a split system. Our results suggest that careful allocation of holding costs can create incentives that enable minimum turnaround times using a common queue.



## 1. Introduction

We first provide a motivating context for the model. Consider the repair process for aircraft components (such as engines and gearboxes) at the US Coast Guard (USCG) Aviation Division. When such components fail, they are frequently sent to outside vendors (such as GE, Pratt and Whitney, Rockwell or others) for repair (see [4]). The first step is to diagnose the good parts within these components and provide credit to the Coast Guard for these good parts. Here credit refers to the payment from the vendors to the Coast Guard as "security" for these good parts while the broken component waits to be repaired. Upon repair, the working component (with the original good parts installed) is returned to the Coast Guard, the security for the good parts is reimbursed, and the vendors get paid for the repaired functioning product based on predetermined contract terms. Holding cost in this context thus refers to the financing cost for the good parts and the storage cost of components while they are waiting for repair and has to be borne by the vendors. The goal of the Coast Guard is to manage the repair task allocation and holding cost allocation to vendors (i.e., deciding the workload and the portion of the financing and storage costs each vendor is responsible for) in order to incentivize fast turnaround of components.

The details of the flows are described next. The broken components may be stored in a common pool and allocated to vendors based on a demand allocation scheme or they may be allocated to vendors upon arrival. They can thus be part of a common queue or a separate queue of components waiting to be repaired. In both cases, the realized holding cost has to be allocated to vendors. The vendor who does the repair is paid the revenue associated with the repair. It is intuitive that vendors will choose a rate of repair that maximizes their profits over time, i.e., the net revenue of holding costs. The goal of this paper is to examine how holding cost allocation and repair task allocation to vendors affect vendor repair rates and thus component turnaround time, i.e., the sum of waiting time plus repair time for components.

We now describe a model that abstracts the context and provides associated notation. We thus consider a two-server (where the vendor is a server) queueing system in which the servers are independent companies who choose their service rate, given a customer allocation scheme and an associated holding cost allocation scheme offered by the demand generating entity. The service times are modeled as exponentially distributed with the individual servers determining their own rates $\mu_1$ and $\mu_2$. Let $c(\mu)$ denote the cost of serving at a rate $\mu$ for each server; we assume $c(\mu)$ to be convex increasing in $\mu$. Demand for the service (broken components) arrives according to a Poisson process of rate $\Lambda$. A server receives $R$ for each completed service. The holding cost associated with each unit of demand is based on a positive marginal holding cost $h$ times, the time between demand arrival and demand service completion. Note that all demands have to be served in a finite amount of time; otherwise the servers would incur infinitely large holding costs. Since we are interested in the

* Corresponding author.
*E-mail addresses:* kewee@smu.edu.sg (K.-E. Wee), aiyer@purdue.edu (A. Iyer).

0167-6377/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2010.09.011

system's performance under a competitive environment, we shall assume that the parameters are such that an exit strategy is never part of an equilibrium; that is, it is never optimal for a server to exit the market completely.

Let $i, j = 1, 2, i \neq j$ denote the indices for the servers. Each server chooses a service rate to maximize profits per unit time. Though each repair completion will garner a specific holding cost that will be allocated to the servers, we are interested in the holding cost per unit time. The holding cost per unit time can be expressed (using Little's Law) as the holding cost associated with the random variable that denotes the time average number of the outstanding orders charged to server $i$, denoted by $l_i$. Let $L_i$ denote the expected value of $l_i$. Given a demand and holding cost allocation rule, server $i$ would choose $\mu_i$ so as to maximize its profit per unit time denoted as:

$$\pi_i(\mu_i, \mu_j) = R\lambda_i - hL_i - c(\mu_i), \tag{1}$$

where $\lambda_i$ depends on the demand allocation rule, and $L_i$ depends on the holding cost allocation rule.

Our main results are as follows: (i) We propose an incentivized holding cost allocation (IHCA) scheme that leads to the maximum possible service rate for any given demand allocation and demonstrate its effect for split systems and pooled systems in equilibrium. (ii) We develop conditions under which IHCA generates monotonic decreasing holding cost allocations with service rate increase. (iii) We show that the optimal pooled system always dominates the optimal split system in terms of system expected waiting time.

The rest of the paper is organized as follows. Section 2 provides a literature review and positions this paper with respect to other existing work. Section 3 identifies the maximum possible service rate for the split and pooled systems and examines their relationship. Section 4 describes IHCA in general and applies it to split and pooled systems. It also provides insights into the link between problem parameters and the associated holding cost allocation scheme. Finally Section 5 provides a summary of conclusions.

## 2. Literature review

The operational management literature in the area of incentive effects on queue service rates is nascent but growing. Using demand allocation as a mechanism to induce faster service, in the absence of holding costs, has been analyzed in the literature. For example, [5] consider a two-server system with two types of demand allocation schemes: (1) balanced allocation under which jobs, upon arrival, are immediately assigned to one of the two servers with the goal of balancing the expected waiting time at each server based on the service rates; and, (2) common queue allocation (first studied in [7]) under which jobs are only allocated to idle servers, with each idle server equally likely to be allocated a job; jobs form a common queue if both servers are busy. Conventional wisdom suggests that common queue allocation, by pooling capacity and thus risks, is more efficient in utilizing system resources and hence typically leads to better system performance than would balanced allocation (see for example [8]). However, in the presence of strategic servers i.e., servers who are allowed to choose their service rates depending on the demand allocation process, [5] show that the demand allocation mechanism in the balanced allocation provides an incentive effect on the servers (since the balanced allocation allocates more demand to the faster server than would common queue allocation) giving rise to higher service rates that could lead to shorter expected system waiting time (including service time) than would common queue allocation.

Bell and Stidham [1] discuss social versus individual optimization and describe a balanced allocation scheme that we will analyze later. Cachon and Zhang, [3] classify balanced allocation and common queue allocation as a state-independent system and a state-dependent system respectively. They do so because balanced allocation allocates demand to servers based only on their capacities but not on the current state of the system (e.g., information regarding which server is idle) unlike common queue allocation. Cachon and Zhang [3] then propose and analyze an optimal state-independent system, which they label linear allocation.

Lin and Kumar [10] show that threshold allocation scheme is an optimal state-dependent allocation scheme. Under threshold allocation, demand is allocated to the busy faster server instead of the idle slower server as long as the demand in queue at the faster server is less than a threshold. Unfortunately, as pointed out in [3], while a numerical method to evaluate the system's performance under threshold allocation given non-strategic servers is available (see [11]), there is neither explicit expression nor analytical characterization for the optimal threshold for both cases of strategic and non-strategic servers.

Both [3,5] focus on demand allocation mechanisms and do not consider holding cost allocation mechanisms. In fact, their models assume holding costs to be zero. As in the motivating example, while the total system holding costs for any given vendor capacity may be fixed, the allocation schemes that attribute the system holding costs to the vendors may impact the vendor capacity choices much like the incentive effect of the demand allocation mechanisms.

## 3. Model formulation

Following the convention in the literature, we divide the set of demand allocation policies into two broad classes: (1) split (or state-independent) systems in which arriving jobs are immediately assigned to one of the two servers based on their capacities; and, (2) pooled (or state-dependent) systems in which jobs are allocated to the servers based on the current state of the system. We will first identify the maximum service rate for the split and pooled systems. We will then devise a holding cost allocation scheme that attains this service rate. Finally we will explore the nature of the holding cost allocation process.

Our first goal is to determine, for each of the split and pooled systems, the optimal demand and holding cost allocation scheme. Our proposed scheme for each system is optimal in the sense that it leads to the maximum possible *symmetric* service rates at equilibrium that can be offered by the servers for each system. We emphasize that while our goal is to attain the maximum possible symmetric service rates at equilibrium, we do not restrict our analysis to just symmetric equilibria when we are determining the Nash equilibria for the game; a symmetric Nash equilibrium solution emerges as the unique (and hence symmetric) above-mentioned maximum possible symmetric service rate. Our rationale for the optimality of a unique symmetric maximum possible service rate at equilibrium, if attained, is two-fold. First, our proposed allocation schemes, as well as those considered in the literature all culminate in both servers splitting the total demand rate equally. From a central planner's perspective, given that both servers split the total demand rate equally, it is straightforward to show that for any asymmetric service rate pair $(\mu_1, \mu_2)$ offered by the servers, the expected waiting time for each customer can be improved when both servers offer identical service rate $\mu$, where $\mu = \frac{\mu_1+\mu_2}{2}$. Put it differently, given that both servers split the total demand rate equally, then our proposed scheme leads to the maximum possible service rate that each server could offer at equilibrium. Second, consistent with the literature, a symmetric equilibrium has a natural appeal and is hence desirable (see for

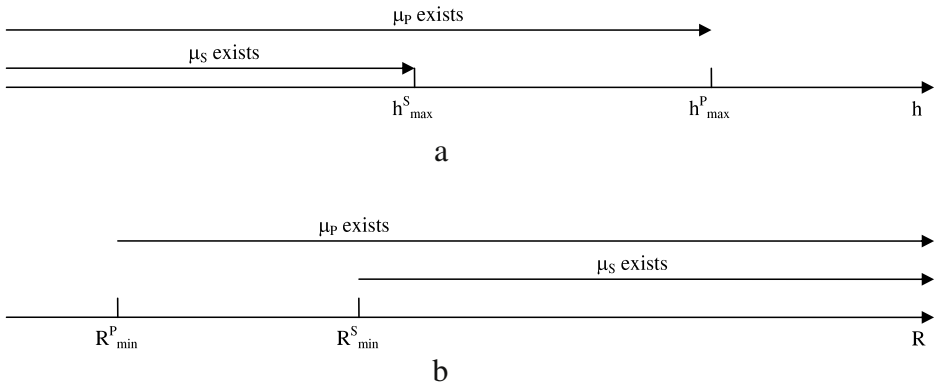



**Fig. 1.**

example [2]). Indeed, as stated in [6]: "it is more natural for firms to focus on a symmetric equilibrium, so that it is more likely to occur than an asymmetric one". (See more discussion of the focal principle in [9].)

We first derive the maximum possible symmetric service rates at equilibrium, $\mu_s$ and $\mu_p$, for a split system and a pooled system respectively. Suppose that each server has an outside option of value $v$; i.e., a server will not provide any service unless its expected profit for offering its service is at least $v$. For simplicity of the presentation, we shall henceforth assume that $v$ equals 0, although the analysis can be extended readily to the case of positive $v$. Notice that the maximum possible symmetric service rates must be such that the servers would each earn zero profit given that they are allocated half of the repair tasks (since a larger service rate is feasible if each server earns positive profit). Hence, applying standard $M/M/1$ and $M/M/2$ queueing results to the server's profit function in (1), $\mu_s$ and $\mu_p$ must be each respectively a root of $\pi^s$ and $\pi^p$ presented as follows:

$$\pi^s(\mu) = R\Lambda/2 - h\frac{\Lambda}{2\mu - \Lambda} - c(\mu). \tag{2}$$

$$\pi^p(\mu) = R\Lambda/2 - h\frac{2\mu\Lambda}{4(\mu)^2 - \Lambda^2} - c(\mu). \tag{3}$$

In fact, since our goal is to attain the maximum possible symmetric service rates at equilibrium, $\mu_s$ and $\mu_p$ are each respectively the largest root of $\pi^s$ and $\pi^p$. We note that in the event of positive $v$, we only need to add the term $-v$ to both Eqs. (2) and (3) respectively.

**Lemma 1.** *Both $\pi^s$ and $\pi^p$ are concave in $\mu$ in the range $(\Lambda/2, \infty)$ and hence each has at most two real roots larger than $\Lambda/2$.*

Lemma 1 implies that $\mu_s$ and $\mu_p$, if they exist, are each the larger of the two roots of Eqs. (2) and (3). Let $w_s(\mu_s)$ and $w_p(\mu_p)$ denote the expected system waiting time corresponding to $\mu_s$ and $\mu_p$ for the split and pooled system respectively. Note also that $\pi^p \geq \pi^s$. Consequently, we have

**Corollary 1.** (i) $\mu_p > \mu_s$. (ii) $w_p(\mu_p) < w_s(\mu_s)$.

Hence, consistent with the conventional wisdom, an optimal pooled system (that induces $\mu_p$) always outperforms an optimal split system (that induces $\mu_s$). A natural question is: *What are the conditions under which $\mu_s$ and $\mu_p$ exist*? Such conditions can be derived from Lemma 1, together with the observations that both $\pi^s$ and $\pi^p$ are increasing in $R$ and decreasing $h$.

**Corollary 2.** (i) *For any given $h$, there exist $R^s_{\min}(h)$ and $R^p_{\min}(h)$, $R^p_{\min}(h) \leq R^s_{\min}(h)$, such that $\mu_s$ ($\mu_p$) exists if and only if $R \geq R^s_{\min}$ ($R \geq R^p_{\min}$).* (ii) *For any given $R$, there exist $h^s_{\max}(R)$ and $h^p_{\max}(R)$, $h^s_{\max}(R) \leq h^p_{\max}(R)$, such that $\mu_s$ ($\mu_p$) exists if and only if $h \leq h^s_{\max}$ ($h \leq h^p_{\max}$).*

Corollary 2 shows that there exists a lower bound for $R$ (for a given $h$) and a lower bound for $h$ (for a given $R$) for each of split systems and pooled systems, beyond which $\mu_s$ and $\mu_p$ fail to exist; in this case, it is never economically viable for the servers to offer any service rate. Since $\pi^p \geq \pi^s$, it is not surprising to observe from Corollary 2 that $\mu_p$ exists over a larger parameter range than does $\mu_s$. This is depicted in Fig. 1.

To avoid the trivial case, we shall assume that the conditions with respect to the bounds in Corollary 2 hold so that both $\mu_s$ and $\mu_p$ are well-defined. To facilitate our analysis in the next section, we present the following comparative static results of $\mu_s$ and $\mu_p$ with respect to $R$ and $h$.

**Lemma 2.** *Both $\mu_s$ and $\mu_p$ are concave increasing in $R$ and concave decreasing in $h$.*

So far in this section, our results concerning $\mu_s$ and $\mu_p$ are applicable to the general split and pooled systems. Next, we focus on a specific pooled system and split system, namely, common queue allocation and balanced allocation. Consequently, whenever applicable, we shall identify variables associated with each of the two allocations using their respective letter superscripts: CQ, B.

## 4. Incentivized holding cost allocation (IHCA) policy

In this section, we provide holding cost allocation policies that will induce $\mu_s$ and $\mu_p$ and thus generate the optimal service rate at the servers. Before we introduce our optimal holding cost allocation policy, IHCA, we first state the following desirable properties of a holding cost allocation policy. Let $l$ denote the total number of outstanding orders in the system and let $L$ denote the expected value of $l$.

(P1) The holding cost allocation induces $\mu_s$ for the split system and $\mu_p$ for the pooled system.

(P2) The holding costs allocated to the servers should add up to $h \times l$ i.e., the realized average holding cost per unit time; that is, $l_1 + l_2 = l$.

Both (P1) and (P2) are natural. Another desirable property is a monotonically decreasing holding cost allocation i.e., the policy would decrease the holding cost charged as a server increases its service rate. This is stated as follows:

(P3) $\frac{\mathrm{d}l_i}{\mathrm{d}\mu_i} \leq 0$.

We now present IHCA, under which the holding costs per unit time charged to each server are $hl_i$ or $hl_j$ for $i, j = 1, 2,\ i \neq j$, as follows:

$$l_i = \begin{cases} 0 & \text{if } \kappa(\mu_j - \mu_i) < -l/2 \\ l/2 + \kappa(\mu_j - \mu_i) & \text{if } -l/2 \leq \kappa(\mu_j - \mu_i) \leq l/2 \\ l & \text{if } \kappa(\mu_j - \mu_i) > l/2. \end{cases} \tag{4}$$

Note that $0 \le l_i \le l$ as desired. Since $l$ depends on the service rates $\mu_i$ and $\mu_j$, the associated holding cost allocation is nonlinear in the service rates. The allocation starts with each server being allocated half the system holding costs with an adjustment (via the parameter $\kappa$) proportional to the difference between the servers' service rates. As we shall see later, by choosing an appropriate value for the parameter $\kappa$, one could influence the servers so that a desirable service rate at equilibrium can be achieved. Note also that clearly Eq. (4) satisfies (P2).

We now derive for any demand allocation schemes $(\lambda_1, \lambda_2)$ and holding cost allocation schemes $(l_1, l_2)|_l$ such that $l_1 + l_2 = l$ (be it split or pooled systems), conditions under which there is a unique $\kappa$ that leads to the maximum possible service rate ($\mu_s$ for a split system and $\mu_p$ for a pooled system) being the unique Nash equilibrium.

**Theorem 1.** *For any given demand allocation scheme, $(\lambda_1, \lambda_2)$ and associated system holding cost hL, there is an IHCA holding cost allocation $(l_1, l_2)|_l$ with*

$$\kappa = \kappa^{IHCA} = \frac{1}{h}\left[c'(\mu_\bullet) - R\frac{\mathrm{d}\lambda_1}{\mathrm{d}\mu_1}\Big|_{\mu_1=\mu_2=\mu_\bullet}\right] - \frac{1}{2}\frac{\mathrm{d}L}{\mathrm{d}\mu_1}\Big|_{\mu_1=\mu_2=\mu_\bullet} \quad (5)$$

*where $\bullet = s(p)$ if it is a split (pooled) system such that if $\lambda_i$ and $L_i$ satisfy the following conditions:*

(C1) $\dfrac{\mathrm{d}\lambda_i}{\mathrm{d}\mu_i} \le \dfrac{\mathrm{d}\lambda_j}{\mathrm{d}\mu_j}$ *for any* $\mu_i \ge \mu_j$

(C2) $\dfrac{\mathrm{d}L}{\mathrm{d}\mu_i} \ge \dfrac{\mathrm{d}L}{\mathrm{d}\mu_j}$ *for any* $\mu_i \ge \mu_j$

*then any Nash equilibrium must be symmetric and $\mu_i = \mu_\bullet$ is one of the Nash equilibria.*

**Corollary 3.** *In addition to conditions* (C1) *and* (C2)*, if $\frac{\mathrm{d}\pi_i}{\mathrm{d}\mu_i}|_{\mu_1=\mu_2=\mu} = 0$ has exactly one solution in $\mu$, then $\mu_i = \mu_\bullet$ is the unique Nash equilibrium.*

Conditions (C1) and (C2) demonstrate the phenomenon of diminishing returns. Condition (C1) states that the rate of increase of demand allocation, $\lambda_i$ with respect to $\mu_i$ is decreasing in $\mu_i$ while condition (C2) states that the rate of decrease of $L$ with respect to $\mu_i$ is decreasing in $\mu_i$. In general, $\kappa^{IHCA}$ can be one of the Nash equilibria for any systems; conditions (C1) and (C2) are merely sufficient conditions (but not necessary) that ensure that asymmetric equilibria do not exist, while the condition "$\frac{\mathrm{d}\pi_i}{\mathrm{d}\mu_i}|_{\mu_1=\mu_2=\mu} = 0$ has exactly one solution in $\mu$" ensures that there is at most one symmetric Nash equilibrium so that $\kappa^{IHCA}$ is indeed the unique Nash equilibrium. It turns out that conditions (C1) and (C2) and the condition in Corollary 3 are quite general and are satisfied by pooled systems such as common queue system and split systems such as linear allocation (with $\theta \ge 1$) in [3], as well as balanced allocation which is presented next.

### 4.1. Split system — balanced allocation with IHCA policy

If we consider a split system and demand allocation following [1] (i.e., under balanced allocation), demand is allocated as follows:

$$\lambda_i = (\mu_i - \mu_j + \Lambda)/2, \quad i, j = 1, 2,\ i \ne j. \quad (6)$$

Note that $\lambda_i \ge 0$ implies that

$$\mu_i \ge \mu_j - \Lambda. \quad (7)$$

Applying Corollary 3 to the balanced allocation, we get the following corollary.

**Corollary 4.** *Under balanced allocation and IHCA, if $\kappa = \kappa^{B-IHCA} = \frac{1}{h}[c'(\mu_s) - \frac{R}{2}] - \frac{\Lambda}{(2\mu_s-\Lambda)^2}$, then $\mu_i = \mu_s$ is the unique Nash equilibrium.*

Note that $\kappa^{B-IHCA}$ can be uniquely determined once the exogenous parameters such as $R$ and $h$ ($>0$) are given. Henceforth, we shall label the above optimal split system B — IHCA system.

We will examine the impact of problem parameters on the holding cost allocation policy in Section 4.3.

### 4.2. Pooled system—common queue allocation with IHCA policy

For the pooled system, we consider the common queue allocation, which is first studied in [7] for the case without holding costs. Following [7], for any given service rates $\mu_1$ and $\mu_2$, the demand allocated to server $i$ is:

$$\lambda_i = \Lambda\frac{\Lambda\mu_i^2 + \mu_1\mu_2(\mu_1+\mu_2)}{\Lambda(\mu_1+\mu_2)^2 + 2\mu_1\mu_2(\mu_1+\mu_2-\Lambda)}. \quad (8)$$

Analogous to Corollary 4, we have

**Corollary 5.** *Under common queue allocation and IHCA, if $\kappa = \kappa^{CQ-IHCA} = \frac{1}{h}[c'(\mu_p) - R\frac{\Lambda^2}{2\mu_p(2\mu_p+\Lambda)}] - \frac{\Lambda(4\mu_p^2+\Lambda^2)}{(4\mu_p^2-\Lambda^2)^2}$, then $\mu_i = \mu_p$ is the unique Nash equilibrium.*

As before, $\kappa^{CQ-IHCA}$ can be uniquely determined once the exogenous parameters are given. Henceforth, we shall label the above optimal pooled system CQ — IHCA system.

### 4.3. Sufficient conditions for the monotonic nonincreasing holding cost allocation policy P3

So far, we have shown that IHCA policies can satisfy both properties (P1) and (P2). In this section, we present conditions under which property (P3) is also satisfied. Note from Eq. (4) that, since $l$ is decreasing in $\mu_i$, property (P3) is satisfied if $\kappa \ge 0$. Hence it suffices to determine sufficient conditions under which $\kappa^{B-IHCA} \ge 0$ and $\kappa^{CQ-IHCA} \ge 0$. It turns out that $\kappa^{B-IHCA}$ is concave in $R$ under a very general condition, and is decreasing in $h$ while it remains positive. This is presented next.

**Theorem 2.** (i) *If $\frac{\mathrm{d}^3c(\mu)}{\mathrm{d}\mu^3}\frac{\mathrm{d}c(\mu)}{\mathrm{d}\mu} \le (\frac{\mathrm{d}^2c(\mu)}{\mathrm{d}\mu^2})^2$, then $\kappa^{B-IHCA}$ is concave in $R$.* (ii) *While $\kappa^{B-IHCA}$ remains positive, it is decreasing in $h$.*

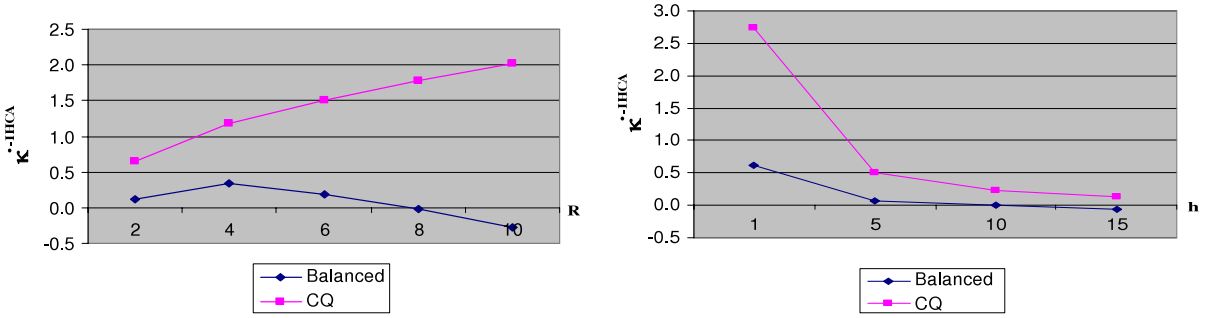
Consequently, we have,

**Corollary 6.** (i) *Suppose $\frac{\mathrm{d}^3c(\mu)}{\mathrm{d}\mu^3}\frac{\mathrm{d}c(\mu)}{\mathrm{d}\mu} \le (\frac{\mathrm{d}^2c(\mu)}{\mathrm{d}\mu^2})^2$. Then for any given $h$, there exist $R_l^s(h)(\ge R_{\min}^s)$ and $R_u^s(h)$ such that $\kappa^{B-IHCA} > 0$ if and only if $R_l^s \le R \le R_u^s$.* (ii) *For any given $R$, there exists $h_u^s(R)$ such that $\kappa^{B-IHCA} > 0$ if and only if $h \le h_u^s$.*

The condition $\frac{\mathrm{d}^3c(\mu)}{\mathrm{d}\mu^3}\frac{\mathrm{d}c(\mu)}{\mathrm{d}\mu} \le (\frac{\mathrm{d}^2c(\mu)}{\mathrm{d}\mu^2})^2$ is rather general and is satisfied, for example, when $c(\mu) = a\mu^n$ for any positive real number $a$ and positive integer $n$. We also note that in all our numerical experiments, $R_l^s$ is either equal or very close to $R_{\min}^s$ so that this lower bound is not as restrictive as it may appear most of the time. For example, consider a representative case with $c(\mu) = 0.1\mu^2$, $\Lambda = 10$ and $h = 2$. Then $R_{\min}^s = 1.924$ whereas $R_l^s = 1.935$.

In contrast to $\kappa^{B-IHCA}$, while $\kappa^{CQ-IHCA}$ is also decreasing in $h$ as long as $\kappa^{CQ-IHCA} > 0$, it is increasing in $R$ under a very general condition.

**Theorem 3.** (i) *If $\frac{\mathrm{d}^2c(\mu)}{rmd\mu^2} - \frac{\Lambda}{\mu(2\mu+\Lambda)}\frac{\mathrm{d}c(\mu)}{\mathrm{d}\mu} + \frac{\Lambda(4\mu+\Lambda)}{\mu^2(2\mu+\Lambda)^2}c(\mu) \ge 0$, then $\kappa^{CQ-IHCA}$ is increasing in $R$.* (ii) *While $\kappa^{CQ-IHCA}$ remains positive, it is decreasing in $h$.*




(a) The behavior of $\kappa^{B-IHCA}$ and $\kappa^{CQ-IHCA}$ as $R$ increases ($\Lambda = 10$, $c(\mu) = 0.1\mu^2$ and $h = 2$).

(b) The behavior of $\kappa^{B-IHCA}$ and $\kappa^{CQ-IHCA}$ as $h$ increases ($\Lambda = 10$, $c(\mu) = 0.1\mu^2$ and $R = 5$).

**Fig. 2.**

Consequently, we have,

**Corollary 7.** (i) *Suppose* $\frac{d^2c(\mu)}{d\mu^2} - \frac{\Lambda}{\mu(2\mu+\Lambda)}\frac{dc(\mu)}{d\mu} + \frac{\Lambda(4\mu+\Lambda)}{\mu^2(2\mu+\Lambda)^2}c(\mu) \geq 0$. *Then for any given h, there exists* $R_l^p(h)$ *such that* $\kappa^{CQ-IHCA} > 0$ *if and only if* $R \geq R_l^p$. (ii) *For any given R, there exists* $h_u^p(R)$ *such that* $\kappa^{CQ-IHCA} > 0$ *if and only if* $h \leq h_u^p$.

Just as in Corollary 6(i), the condition $\frac{d^2c(\mu)}{d\mu^2} - \frac{\Lambda}{\mu(2\mu+\Lambda)}\frac{dc(\mu)}{d\mu} + \frac{\Lambda(4\mu+\Lambda)}{\mu^2(2\mu+\Lambda)^2}c(\mu) \geq 0$ is rather general and is again satisfied, for example, when $c(\mu) = a\mu^n$ for any positive real number $a$ and positive integer $n$. It turns out that for a very general class of $c(\mu)$, $R_l^p \leq R_{\min}^p$ and $h_u^p \geq h_{\max}^p$ so that $\kappa^{CQ-IHCA} > 0$.

**Theorem 4.** *Let* $c(\mu) = a\mu^n$, *where* $a > 0$ *and n is an integer larger than one. Then* $\kappa^{CQ-IHCA} > 0$.

In Fig. 2 we present an example that shows the behavior of $\kappa^{B-IHCA}$ and $\kappa^{CQ-IHCA}$ as $R$ and $h$ increase. In this example, we fix $c(\mu) = 0.1\mu^2$ and $\Lambda = 10$. In Fig. 2(a), we set $h$ equal to 2 and evaluate $\kappa^{B-IHCA}$ and $\kappa^{CQ-IHCA}$ for $R = 2, 4, 6, 8, 10$. In Fig. 2(b), we set $R$ equal to 5 and evaluate $\kappa^{B-IHCA}$ and $\kappa^{CQ-IHCA}$ for $h = 1, 5, 10, 15$. Fig. 2(a) shows that $\kappa^{CQ-IHCA}$ increases with $R$, i.e., it is optimal to charge a lower holding cost to the faster server as the revenue increases when a common queue is used. Fig. 2(b) shows that the impact of increasing holding costs is to decrease the holding cost allocation rate. Note that the graphs are consistent with Theorems 2–4 as well as Corollaries 6 and 7 and that $\kappa^{CQ-IHCA} > 0$ for the range of parameters.

### 4.4. Managerial insights

We have thus provided a holding cost allocation scheme that permits an entity like the Coast Guard to choose a holding cost allocation and a demand allocation scheme that permits the minimum turnaround time for components. To put our holding cost allocation scheme in perspective, note that for balanced allocation, Corollary 6 asserts that for moderate $R$ and $h$, $\kappa^{B-IHCA} > 0$ and thus $\frac{dl_i}{d\mu_i} \leq 0$, while for high $R$ and $h$, $\kappa^{B-IHCA} < 0$ and thus $\frac{dl_i}{d\mu_i} \geq 0$. An interpretation is that for balanced allocation, when $R$ and $h$ are moderate, it is not profitable for servers to provide a high service rate. In this case, to create the appropriate incentive to offer high service rates, B − IHCA induces high service rate by assessing a faster server a smaller portion of the overall system holding costs. However, when $R$ and $h$ are high, the servers are already motivated to offer high service rate. In this case, to induce a Nash equilibrium, B − IHCA adjusts incentives by assessing a faster server, who serves a greater fraction more demand, a higher portion of the overall system holding costs. We emphasize that this last observation does not imply that the equilibrium does not favor high service rate. On the contrary, the high $R$ and $h$ offer enough incentives for the servers to provide high service rate. However, to achieve equilibrium, it takes a $\kappa^{B-IHCA}$ (qualitatively different than $\kappa^{CQ-IHCA}$) that results in the faster server being assessed a higher portion of the overall system holding costs. In contrast to balanced allocation, for a pooled common queue allocation, Theorem 4 asserts that under a very general class of $c(\mu)$, $\kappa^{CQ-IHCA} > 0$ and thus $\frac{dl_i}{d\mu_i} \leq 0$ for all $R$ and $h$. Hence, for a pooled common queue allocation, under a very general class of $c(\mu)$, CQ − IHCA exhibits monotonicity in that it always assess a faster server a smaller portion of the overall system holding costs. We attribute such difference in the behaviors of $\kappa^{B-IHCA}$ and $\kappa^{CQ-IHCA}$ to the higher efficiency in utilizing system resources under the pooled queue over the split queue system. Thus, the holding cost allocation adjusts incentives, by decreasing or increasing holding cost allocations to the faster server, to enable the system to operate under the maximum possible service rate.

The insights above also suggest that results with no holding costs should be used with caution when adapted to systems with holding costs. We note that $\kappa^{B-IHCA}$ is not well-defined if $h = 0$, in which case $\pi_i$ in Eq. (1), and subsequently the model, reduce to those considered in [3,5]. They present conditions under which the unique Nash equilibrium, which generates nonzero profits for the servers, exists. The corresponding rates are lower than those we consider because their incentive scheme cannot induce the maximum service rate. However, the presence of holding cost and its appropriately chosen allocation scheme can help realize $\mu_s$. Thus, in the presence of holding costs, the system reverts back to the optimality of pooled service systems. This helps in reconciling the counterintuitive result in [5] (that a split system could outperform a pooled system) and the traditional findings in queueing theory. One additional note of caution is when the linear allocation scheme in G&Z (considered for $h = 0$) is extended to a linear holding cost allocation when $h > 0$, it is possible to generate multiple Nash equilibria, thus making attainment of the maximum service rate difficult.

The key takeaway is that (a) holding cost allocations enable a large enough set of incentive schemes that can permit maximum service rate to be generated from servers and (b) the associated allocations enable the traditional queueing results to be resurrected as the best option for the system.

## 5. Conclusions

While our proposed demand and holding cost allocation schemes lead naturally to a unique and symmetric Nash equilibrium, it is possible that other allocation schemes may lead to asymmetric equilibria. In this case, in order to search for the optimal

allocation scheme, one may have to extend the notion of maximum possible symmetric service rates at equilibrium to the case of asymmetric equilibria. We leave such exploration of optimal allocation schemes to future research.

## Appendix

**Proof of Lemma 1.** The proof follows by taking the second-order derivatives of $\pi^s$ and $\pi^p$ respectively. □

**Proof of Lemma 2.** The proofs for $\mu_s$ and $\mu_p$ are similar. Setting $\pi^s(\mu_s) = 0$, differentiating both sides with respect to $R$ and $h$, and with further manipulation, we can verify that $\frac{d\mu_s}{dR} \geq 0$, $\frac{d^2\mu_s}{dR^2} \leq 0$, $\frac{d\mu_s}{dh} \leq 0$ and $\frac{d^2\mu_s}{dh^2} \leq 0$. The above inequalities follow since $\frac{d\pi^s}{d\mu}|_{\mu_s} = h\frac{2\Lambda}{(2\mu_s-\Lambda)^2} - c'(\mu_s) \leq 0$ (because $\mu_s$ is the largest root and $\pi^s$ is concave). □

**Proof of Theorem 1.** From Eq. (1), $\frac{d\pi_i}{d\mu_i} = R\frac{d\lambda_i}{d\mu_i} - h\frac{dL_i}{d\mu_i} - c'(\mu_i)$. We first show that asymmetric equilibria do not exist. Suppose on the contrary $(\mu_1, \mu_2)$ is an asymmetric equilibrium and without loss of generality, let $\mu_1 > \mu_2$. Note that from Eq. (4) and conditions (C1) and (C2),

$$0 = \frac{d\pi_1}{d\mu_1} - \frac{d\pi_2}{d\mu_2} = \begin{cases} R\left(\frac{d\lambda_1}{d\mu_1} - \frac{d\lambda_2}{d\mu_2}\right) + c'(\mu_2) - c'(\mu_1) < 0 \\ \quad \text{if } \kappa(\mu_j - \mu_i) < -L/2; \\ R\left(\frac{d\lambda_1}{d\mu_1} - \frac{d\lambda_2}{d\mu_2}\right) + \frac{h}{2}\left(\frac{dL}{d\mu_2} - \frac{dL}{d\mu_1}\right) \\ \quad + c'(\mu_2) - c'(\mu_1) < 0 \quad \text{if } -L/2 \leq \kappa(\mu_j - \mu_i) \leq L/2; \\ R\left(\frac{d\lambda_1}{d\mu_1} - \frac{d\lambda_2}{d\mu_2}\right) + h\left(\frac{dL}{d\mu_2} - \frac{dL}{d\mu_1}\right) \\ \quad + c'(\mu_2) - c'(\mu_1) < 0 \quad \text{if } \kappa(\mu_j - \mu_i) > L/2, \end{cases}$$

a contradiction. Finally, setting $\frac{d\pi_1}{d\mu_1}|_{\mu_1=\mu_\bullet=\mu_2}$ equal zero, we get $\kappa = \kappa^{IHCA}$. □

**Proof of Corollary 4.** From standard $M/M/1$ queueing results, $L = \frac{2\Lambda}{\mu_1+\mu_2-\Lambda}$. It is straightforward to show that Eqs. (4) and (6) satisfy conditions (C1) and (C2). Next, let $G^{B-IHCA}(\mu) = \frac{d\pi_1}{d\mu_1}|_{(\mu_1=\mu=\mu_2)} = R/2 + h(\kappa + \frac{2\Lambda}{(2\mu-\Lambda)^2}) - c'(\mu)$. Then

$$\frac{dG^{B-IHCA}}{d\mu} = -\frac{4h\Lambda}{(2\mu-\Lambda)^3} - c''(\mu) < 0,$$

so that $G^{B-IHCA}$ is decreasing in $\mu$ and thus a symmetric equilibrium is unique. Finally, set $G^{B-IHCA}$ equal zero with $\mu$ replaced by $\mu_s$, we get $\kappa = \kappa^{B-IHCA}$. □

**Proof of Corollary 5.** From standard $M/M/2$ queueing results, we have

$$L = \frac{\Lambda(\mu_1+\mu_2)^3}{(\mu_1+\mu_2-\Lambda)[2\mu_1\mu_2(\mu_1+\mu_2) + \Lambda(\mu_1^2+\mu_2^2)]}.$$

It is straightforward to show that Eqs. (4) and (8) satisfy conditions (C1) and (C2). Next, let $G^{CQ-IHCA}(\mu) = \frac{d\pi_1}{d\mu_1}|_{\mu_1=\mu=\mu_2} = R\frac{\Lambda^2}{2\mu(2\mu+\Lambda)} + h(\kappa + \frac{\Lambda}{4\mu^2-\Lambda^2}) - c'(\mu)$. Then

$$\frac{dG^{CQ-IHCA}}{d\mu} = -R\frac{\Lambda^2(4\mu+\Lambda)}{\mu^2(2\mu+\Lambda)^2} - h\frac{8\mu\Lambda}{(4\mu^2-\Lambda^2)^2} - c''(\mu) < 0,$$

so that $G^{CQ-IHCA}$ is decreasing in $\mu$ and thus a symmetric equilibrium is unique. Finally, set $G^{CQ-IHCA}$ equal zero with $\mu$ replaced by $\mu_p$, we get $\kappa = \kappa^{CQ-IHCA}$. □

**Proof of Theorem 2.** (i) The following is straightforward to derive.

$$\frac{d\kappa^{B-IHCA}}{dR} = \frac{1}{h}\left\{\left[c''(\mu_s) + \frac{4h\Lambda}{(2\mu_s-\Lambda)^3}\right]\frac{d\mu_s}{dR} - \frac{1}{2}\right\} \tag{A.1}$$

$$\frac{d^2\kappa^{B-IHCA}}{dR^2} = \frac{1}{h}\left\{\left[\frac{d^3c(\mu)}{d\mu^3}\Big|_{\mu_s} - \frac{24h\Lambda}{(2\mu_s-\Lambda)^4}\right]\left(\frac{d\mu_s}{dR}\right)^2 + \left[c''(\mu_s) + \frac{4h\Lambda}{(2\mu_s-\Lambda)^3}\right]\frac{d^2\mu_s}{dR^2}\right\}. \tag{A.2}$$

From (2), setting $\pi^s(\mu_s)$ equal zero and differentiating twice, we get

$$\frac{d^2\mu_s}{dR^2} = \frac{h\frac{8\Lambda}{(2\mu_s-\Lambda)^3} + c''(\mu_s)}{h\frac{2\Lambda}{(2\mu_s-\Lambda)^2} - c'(\mu_s)}\left(\frac{d\mu_s}{dR}\right)^2. \tag{A.3}$$

Substituting (A.3) into (A.2), we have

$$\begin{aligned} h\frac{d^2\kappa^{B-IHCA}}{dR^2} &= -\frac{\left(\frac{d\mu_s}{dR}\right)^2}{h\frac{2\Lambda}{(2\mu_s-\Lambda)^2} - c'(\mu_s)} \\ &\quad \times \left\{\left[\frac{d^3c(\mu)}{d\mu^3}\Big|_{\mu_s} - \frac{24h\Lambda}{(2\mu_s-\Lambda)^4}\right]\left[c'(\mu_s) - h\frac{2\Lambda}{(2\mu_s-\Lambda)^2}\right]\right. \\ &\quad \left. - \left[c''(\mu_s) + \frac{4h\Lambda}{(2\mu_s-\Lambda)^3}\right]\left[c''(\mu_s) + \frac{8h\Lambda}{(2\mu_s-\Lambda)^3}\right]\right\} \\ &< -\frac{\left(\frac{d\mu_s}{dR}\right)^2}{h\frac{2\Lambda}{(2\mu_s-\Lambda)^2} - c'(\mu_s)}\left\{\left(\frac{d^3c(\mu)}{d\mu^3}\frac{dc(\mu)}{d\mu}\right)\Big|_{\mu_s} - c''(\mu_s)^2\right\} \\ &< 0 \end{aligned}$$

where both inequalities make use of $\frac{d\pi^s}{d\mu}|_{\mu_s} = h\frac{2\Lambda}{(2\mu_s-\Lambda)^2} - c'(\mu_s) \leq 0$.

(ii)

$$\frac{d\kappa^{B-IHCA}}{dh} = -\frac{1}{h^2}[c'(\mu_s) - R/2] + \left[\frac{c''(\mu_s)}{h} + \frac{4h\Lambda}{(2\mu_s-\Lambda)^3}\right]\frac{d\mu_s}{dh}. \tag{A.4}$$

Part (ii) follows by observing that $[\frac{c''(\mu_s)}{h} + \frac{4h\Lambda}{(2\mu_s-\Lambda)^3}]\frac{d\mu_s}{dh}$ is negative and so is $-\frac{1}{h^2}[c'(\mu_s) - R/2]$ if $\kappa^{B-IHCA}$ is positive. □

**Proof of Theorem 3.** (i) The following are straightforward to derive.

$$\frac{d\kappa^{CQ-IHCA}}{dR} = \frac{1}{h}\left\{\left[c''(\mu_p) + h\frac{8\Lambda\mu_p(3\Lambda^2+4\mu_p^2)}{(4\mu_p^2-\Lambda^2)^3} + R\frac{\Lambda^2(4\mu_p+\Lambda)}{2\mu_p^2(2\mu_p+\Lambda)^2}\right]\frac{d\mu_p}{dR} - \frac{\Lambda^2}{2\mu_p(2\mu_p+\Lambda)}\right\} \tag{A.5}$$

$$\begin{aligned} \frac{d\mu_p}{dR} &= -\left(\frac{d\pi^p}{dR} \Big/ \frac{d\pi^p}{d\mu}\right)\Big|_{\mu_p} \\ &= -\frac{\Lambda/2}{h\frac{2\Lambda(4\mu_p^2+\Lambda^2)}{4\mu_p^2-\Lambda^2} - c'(\mu_p)} \geq 0. \end{aligned} \tag{A.6}$$

Note from Eq. (3) that

$$R = \frac{2}{\Lambda}\left[h\frac{2\mu_p\Lambda}{4(\mu_p)^2-\Lambda^2} + c(\mu_p)\right] > \frac{2}{\Lambda}c(\mu_p) \tag{A.7}$$

Using (A.5)–(A.7), we have

$$
\begin{aligned}
&\frac{\mathrm{d}\kappa^{CQ-IHCA}}{\mathrm{d}R}\\
&> \frac{1}{h}\frac{\mathrm{d}\mu_p}{\mathrm{d}R}\left\{c''(\mu_p) + R\frac{\Lambda^2(4\mu_p+\Lambda)}{2\mu_p^2(2\mu_p+\Lambda)^2} - \frac{\Lambda^2}{2\mu_p(2\mu_p+\Lambda)\frac{\mathrm{d}\mu_p}{\mathrm{d}R}}\right\}\\
&> \frac{1}{h}\frac{\mathrm{d}\mu_p}{\mathrm{d}R}\left\{c''(\mu_p) + \frac{\Lambda(4\mu_p+\Lambda)}{\mu_p^2(2\mu_p+\Lambda)^2}c(\mu_p)\right.\\
&\quad\left. - \frac{\Lambda}{\mu_p(2\mu_p+\Lambda)}\left[c'(\mu_p) - h\frac{2\Lambda(4\mu_p^2+\Lambda^2)}{4\mu_p^2-\Lambda^2}\right]\right\}\\
&> \frac{1}{h}\frac{\mathrm{d}\mu_p}{\mathrm{d}R}\left\{c''(\mu_p) + \frac{\Lambda(4\mu_p+\Lambda)}{\mu_p^2(2\mu_p+\Lambda)^2}c(\mu_p)\right.\\
&\quad\left. - \frac{\Lambda}{\mu_p(2\mu_p+\Lambda)}c'(\mu_p)\right\}. \quad \square
\end{aligned}
$$

(ii) The proof is similar to that of Theorem 2(ii).

**Proof of Theorem 4.** We shall show that $\kappa^{CQ-IHCA} > 0$ for $R = R^p_{\min}$; for $R > R^p_{\min}$, the result then follows from Theorem 3 (i). Recall that $\pi^p(\mu)$ is concave and note that at $R = R^p_{\min}$, $\pi^p(\mu)$ touches the $x$-axis at $\mu_p$ and is negative at other values of $\mu$; that is $\mu_p$ satisfies $\pi^p(\mu_p) = 0$ and $\frac{\mathrm{d}\pi^p(\mu)}{\mathrm{d}\mu}|_{\mu_p} = 0$. (For other $R$ larger than $R^p_{\min}$, $\pi^p(\mu_p) = 0$ but $\frac{\mathrm{d}\pi^p(\mu)}{\mathrm{d}\mu}|_{\mu_p} < 0$ since $\mu_p$ is the larger root of $\pi^p$.) Solving these two equations for $R$ and $h$, we have

$$R = \frac{2a\mu_p^n[4(n+1)\mu_p^2 - (n-1)\Lambda^2]}{\Lambda(4\mu_p^2+\Lambda^2)} \tag{A.8}$$

$$h = \frac{na\mu_p^{n-1}(4\mu_p^2-\Lambda^2)^2}{2\Lambda(4\mu_p^2+\Lambda^2)}. \tag{A.9}$$

Substitute (A.8) and (A.9) into $\kappa^{CQ-IHCA}$, we have

$$\kappa^{CQ-IHCA} = \frac{\Lambda}{n(2\mu_p-\Lambda)^2(2\mu_p+\Lambda)^3}f(\mu_p)$$

where

$$
\begin{aligned}
f(\mu_p) &= 8n\mu_p^3 - 4(n+2)\Lambda\mu_p^2 + 2n\Lambda^2\mu_p + (3n-2)\Lambda^3\\
&= n\mu_p(4\mu_p^2 - 4\Lambda\mu_p + \Lambda^2) + 4n\mu_p^3 - 8\Lambda\mu_p^2\\
&\quad + n\Lambda^2\mu_p + (3n-2)\Lambda^3\\
&> n\mu_p(4\mu_p^2 - 4\Lambda\mu_p + \Lambda^2) + 2\mu_p(4\mu_p^3 - 4\Lambda\mu_p + \Lambda^2)\\
&\quad + (3n-2)\Lambda^3 \quad (\text{since } n \geq 2)\\
&> 0.
\end{aligned}
$$

Hence $\kappa^{CQ-IHCA} > 0$ at $R = R^p_{\min}$. $\square$

## References


[1] C. Bell, S. Stidham, Individual versus social optimization in the allocation of customers to alternative servers, Mgmt Sci. 29 (1983) 831–839.
[2] G. Cachon, S. Netessine, Game theory in supply chain analysis, in: D. Simchi-Levi, S.D. Wu, M. Shen (Eds.), Handbook of Supply Chain Analysis in the E-Business Era, Kluwer Academic Publishers, Amsterdam, 2004, pp. 13–66.
[3] G. Cachon, F. Zhang, Obtaining fast service in a queueing system via performance-based allocation of demand, Mgmt Sci. 53 (3) (2007) 408–420.
[4] K. Everingham, G. Polaski, F. Riedlin, V Deshpande, A. Iyer, Operations research enhances supply chain management at the US coast guard aircraft repair and supply center, Interfaces 38 (1) (2008) 61–75.
[5] S.M. Gilbert, Z.K. Weng, Incentive effects favor nonconsolidating queues in a service system: the principle-agent perspective, Mgmt Sci. 44 (12) (1998) 1662–1669.
[6] A.Y. Ha, S. Tong, Contracting and information sharing under supply chain competition, Mgmt Sci. 54 (4) (2008) 701–715.
[7] E. Kalai, M.I. Kamien, M. Rubinovitch, Optimal service speeds in a competitive environment, Mgmt Sci. 38 (8) (1992) 1154–1163.
[8] Kleinrock, Queueing Systems Volume I: Theory, John Wiley & Sons, 1975.
[9] D. Kreps, A Course in Microeconomics Theory, Princeton University Press, Princeton, NJ, 1990.
[10] W.P. Lin, R. Kumar, Optimal control of a queuing system with two heterogeneous servers, IEEE Trans. Automat. Control 29 (1984) 696–703.
[11] M. Rubinovitch, The slow server problem: a queue with stalling, J. Appl. Probab. 22 (1985) 879–892.