Optimal Supply & Demand Balance In Service Environments

Eylem Koca, Ph.D., Fairleigh Dickinson University, USA
Mohammad Sedaghat, Ph.D., Fairleigh Dickinson University, USA
K. Paul Yoon, Ph.D., Fairleigh Dickinson University, USA

ABSTRACT

We study service environments that can be modeled as stochastic finite-capacity double-ended queues, where supply and demand arrive in independent Poisson processes to be instantly paired-off. In the case where throughput (output rate) is not a significant metric of system performance (as typically studied in the literature), we derive analytical results to gain managerial insights. We find that the operational decision on optimal supply/demand balance and the strategic decision on how to achieve that optimal balance can be decoupled and stratified. With the purpose of providing a managerial guide, we identify conditions for when to manipulate demand rather than supply, and vice versa. For the first time in the literature, we study throughput considerations in this context, and we analytically characterize the optimal strategy. Specifically, we show that it is optimal to manipulate either demand, or supply (and not both), and that the optimal system balance and the strategy on how to achieve it are strongly tied. Our findings can shed light on the managerial decision making process in these environments, and they can be used to revisit any governing strategies dictating management of demand (or supply) as a first course of action.

Keywords: Service Environment; Optimal Supply & Demand Balance; Double-Ended Queue

INTRODUCTION

Firms, especially in the service industry, are vulnerable to mismatch between demand and available supply due to the stochastic nature of one or both; excess demand means lost revenue, excess supply means waste of resources. What differentiates service environments is the requirement of perfect pairing between a demand arrival and supply availability for a service delivery to occur. Without a taxi cab, you have waiting customers; without customers, you have a line of waiting cabs. In other environments, such as telecommunication networks, computing (Parthasarathy et al., 1999), stock exchange, just-in-time assembly systems (Som et al., 1994), organ allocation (Zenios, 1999) etc., imbalances between demand and supply degrade the system performance through operational inefficiencies, rather than revenues lost.

In this study, we investigate certain types of service environments and systems, where "requests for supply" and "supply availability" arrive in two separate stochastic streams, to be paired for the service request to be fulfilled. We model these systems using a synchronizing queue, where service requests and supply arrive from separate ends, and they leave the system in FIFO (first in, first out) fashion upon instant pairing-off. This type of queuing model, traditionally called "double-ended queues," was first studied by Kendall (1951) using the taxi stand example, which was also studied using double-ended queues by and Kashyap (1966). A similar queue structure arises when pairs (or bulks) of customers are to be serviced at a time (Latouche, 1981; Hsu et al., 1993).

In synchronization queues as studied in this paper, the system is unstable in the steady-state unless a mechanism for control is introduced. For example, Sasieni (1961), Perry & Stadje (1999), Frutos & Gallego (1999), Conolly et al. (2002), and Li & Jiang (2013) study customer impatience. Prabhakar et al. (2000) focus on the output process of double-ended queues with infinite queue capacities. The other common approach is to assume finite queue capacities. Bollapragada & Rao (2006) investigate the non-stationary behavior of capacitated double-ended queues in the inventory management context. Lander (1986) assumes equal demand and supply queue capacities (k)

and reports asymptotic findings as *k* approaches infinity. Takahashi et al. (2000) study the finite capacity case with phase-type supply process. While these papers give technical solutions to extensions of double-ended queuing systems, they do not address the managerial decision making process of matching supply and demand.

We study the decision-making process for a non-symmetric, finite-queue system where there is supply/demand imbalance (mismatch). Mendoza et al. (2009) and Mendoza et al. (2014) study a similar case, addressing excess supply and excess demand cases, respectively. However, neither study provides insights on *how to manage* the supply/demand imbalance, nor do they address systems where throughput[1] is a crucial aspect of system performance and systems that are profit centers (rather than cost centers). This paper fills the gap in the literature by 1) developing a unified framework of supply/demand imbalance, 2) providing managerial insights on strategies and policies in manipulating demand and supply in case of undesired imbalance, 3) addressing supply/demand systems where throughput is an important consideration, and those that are profit centers as well as those that are cost centers.

The rest of the paper is organized as follows. Section 2 presents the double-ended queue model, its steady state solutions, and the system parameters that we address in this paper. In Section 3, we look at the case where the system is a cost center with no throughput considerations. In Section 4, we study supply/demand environments where throughput is considered an important performance metric. Section 5 draws conclusions and shows pointers for future research.

THE MODEL

We conceptualize a system where demand for a single type of supply (service, goods, etc.), and the supply (servers, items, etc.) to meet this type of demand both arrive one unit at a time in separate and independent stochastic processes. At the arrival of a demand (supply), it is paired with a supply (demand) waiting in line, after which both leave the system instantaneously; if there is no supply (demand), then it joins the queue on the demand (supply) side unless the queue is full, at which time it leaves the system. We assume finite demand and supply queue capacities of d and s, respectively. The system operates according to FIFO principle. It is easily seen that if the demand (supply) queue is non-empty, supply (demand) queue must be empty; they can be both empty, but they cannot be both non-empty.

With the above description of the system, it is also seen that the definition of supply and demand is arbitrary. In fact, in many environments that operate in this fashion, there is either no notion of demand and supply, or both streams can be seen as demand (or supply). Consider assembly systems where components come together to form ``kits'' to go into the main assembly line, the stock exchange where sell requests and buy requests are matched, or the taxi stand where customers are demanding taxi cabs while taxi drivers are demanding customers. In the light of this observation, we adopt a unified approach where *we call the arrival process with higher rate, the demand process.*

We assume that both demand and supply arrival processes are stationary Poisson with known rates of λ and μ, respectively. According to the convention mentioned above, we assume, without loss of generality, that $\lambda \geq \mu$. We define the state of the system, t, as demand queue length minus supply queue length. Since both of them cannot be positive at the same time, and given the queue capacities, we have $t \in \{-s, \dots, -2, -1, 0, +1, +2, \dots, d\}$. As a result, if the system is at a positive (negative) t state, then the demand (supply) queue is non-empty and the supply (demand) queue is empty. We denote the steady-state probability that system is in state t by P_t.

Suppose that each unit of unmet demand costs the system c_d, and each unit of idle supply costs c_s. Suppose it costs the system c_λ^+ and c_λ^- (c_μ^+ and c_μ^-) on the average per unit of demand (supply) rate increased and decreased, respectively. Therefore, the expected the cost of supply/demand imbalance, C^I, is given by

[1] In queuing systems, throughput is the rate of output from the system. In our context, it refers to the rate of demand paired with supply.

$$C^I = c_s \sum_{t=-s}^{-1} (-t) P_t + c_d \sum_{t=1}^{d} t\, P_t\,,$$

and the total cost of efforts in balancing the system is a function of λ and μ:

$$C^B(\lambda,\mu) = c_\lambda^+ (\lambda-\lambda_0)^+ + c_\lambda^- (\lambda-\lambda_0)^- + c_\mu^+ (\mu-\mu_0)^+ + c_\mu^- (\mu-\mu_0)^-, \quad (1)$$

where $(\cdot)^+ = \max\ \{0,(\cdot)\}$, $(\cdot)^- = \max\ \{0,-(\cdot)\}$, and λ_0 and μ_0 indicate the *current* demand and supply rates, respectively.

The steady-state balance equations for the system under investigation are straight-forward:

$$\mu\, P_d = \lambda\, P_{(d-1)},$$

$$(\lambda+\mu)\, P_t = \mu\, P_{(t+1)} + \lambda\, P_{(t-1)}, \quad \forall\ -s < t < d,$$

$$\mu\, P_{(-s+1)} = \lambda\, P_{(-s)}.$$

These equations lead to $P_t = \rho^{(d-t)}\, P_d$, $-s \le t \le d$, where $\rho = \mu/\lambda < 1$ is a measure of supply/demand imbalance; the smaller the ρ, the deeper the imbalance. Since steady-state probabilities must add up to 1, we find that

$$P_t = \begin{cases} \rho^{(d-t)}\, \dfrac{(1-\rho)}{(1-\rho^{(1+d+s)})}, & \rho \neq 1 \\ \dfrac{1}{1+d+s}, & \rho = 1 \end{cases}, \qquad \forall\ -s \le t \le d.$$

After substituting, we get

$$C^I(\rho) = \begin{cases} \dfrac{c_d\big(d(1-\rho)-\rho(1-\rho^d)\big) + c_s\, \rho^{(1+d)}\Big(1-\rho^s\big(1+s(1-\rho)\big)\Big)}{(1-\rho)\,(1-\rho^{(1+d+s)})}, & \rho \neq 1 \\ \dfrac{\big(c_s\, s\,(1+s) + c_d\, d\,(1+d)\big)}{2(1+d+s)}, & \rho = 1 \end{cases}. \quad (2)$$

In the rest of the paper, we first look at service environments where the main objective is to minimize the system cost due to supply/demand imbalance (Case I); then, we investigate systems where throughput is a crucial system performance metric (Case II).

CASE I: COST CENTER WITHOUT THROUGHPUT CONSIDERATIONS

When the system is a cost center and there are no throughput considerations, the goal is to minimize the total operational costs that arise due to mismatch between demand and supply:

$$C^T = C^I(\rho) + C^B(\lambda,\mu).$$

One may argue that throughput is always of concern in any queuing system. However, there are cases where throughput does not play a significant role in the success of the queuing system (relative to implications of supply/demand mismatch), and it can be ignored. Good examples include any non-profit service center where customers wait for servers to be available, organ allocation to donors (Zenios, 1999), and tenant assignment to

public housing (Kaplan, 1987). In this section, we study such environments, which are what the current literature typically addresses.

Observe from Eq. (2) that the expected cost of imbalance, $C^I(\rho)$, is a function of ratio of demand and supply arrival rates (ρ) only, and not of individual arrival rates. Therefore, the problem of minimizing C^T can be stratified: First, find the optimal ρ (denoted ρ^*) that minimizes $C^I(\rho)$; then, find the values of λ and μ that minimizes $C^B(\lambda, \mu)$, subject to $\mu/\lambda = \rho^*$, where $C^B(\lambda, \mu)$ is as given in Eq. (1). We present this two-stage procedure below.

Stage 1: Determine ρ^* that satisfies

$$C^I(\rho^*) = C^{I*} = \min_{\rho} C^I(\rho)$$

Stage 2: Determine λ^* and μ^* that satisfies

$$C^B(\lambda^*, \mu^*) = C^{B*} = \min_{\lambda,\mu \ s.t. \ \mu/\lambda=\rho^*} C^B(\lambda, \mu)$$

If $\rho^* > \rho_0 = \mu_0/\lambda_0$, then the solution of Stage 1 suggests that the system is currently unable to meet the demand at a desired level. If $\rho^* < \rho_0$, then the optimal solution states that the system is being underutilized (relative to optimal) even though the demand rate is more than supply rate; note that this is possible due to the stochastic nature of demand and supply. If $\rho^* = \rho_0$, the system is already at optimal balance, and no balancing is needed.

In the case of $\rho^* > \rho_0$, the optimal balance can be achieved either by increasing the supply rate, or by decreasing the demand rate, or by a combination. Recall that the unit cost of decreasing the demand rate (while keeping supply rate constant) is c_λ^-. With the $\mu/\lambda = \rho^*$ constraint, it is easy to verify that each unit decrease of demand rate can be compensated by increasing the supply rate by ρ^*. Since each unit of supply rate increased costs c_μ^+, the decision process is simplified: either decrease the demand rate by one unit with a cost of c_λ^-, or increase the supply rate by ρ^* units with a cost of $\rho^* c_\mu^+$; a combination of decreasing the demand rate and increasing the supply rate at the same time can be optimal only if $c_\lambda^- = \rho^* c_\mu^+$. To summarize, we find that

- If $\rho^* > c_\lambda^-/c_\mu^+$, then the optimal strategy is to decrease the demand rate to $\lambda^* = \mu_0/\rho^*$, keeping the supply rate at μ_0;
- If $\rho_0 < \rho^* < c_\lambda^-/c_\mu^+$, then the optimal strategy is to increase the supply rate to $\mu^* = \lambda_0 \rho^*$, keeping the demand rate at λ_0;
- If $\rho^* = c_\lambda^-/c_\mu^+ > \rho_0$, then the system is indifferent between decreasing demand and increasing supply.

Remark: Note that decreasing the demand rate beyond $\lambda^* = \mu_0/\rho^*$ would require decreasing the supply rate proportionately (since the optimal balance must be satisfied) as well, resulting in higher cost; similar argument goes for increasing the supply rate beyond $\mu^* = \lambda_0 \rho^*$.

Corollary: Consider now a special case where $c_\lambda^- = c_\mu^+$, for which the optimal strategy translates to:

- If $\rho^* > 1$, decrease demand rate to $\lambda^* = \mu_0/\rho^* < \mu_0$, rather than increasing the supply rate to $\mu^* = \lambda_0 \rho^* > \lambda_0$.
- If $\rho_0 < \rho^* < 1$, increase supply rate to $\mu^* = \lambda_0 \rho^* < \lambda_0$, rather than decreasing the demand rate to $\lambda^* = \mu_0/\rho^* > \mu_0$.

Note that if $\rho^* > 1$, the optimal balance calls for having a larger supply rate than demand rate, shifting the direction of imbalance; recall that currently, the demand rate is larger than the supply rate. As a result, if the unit cost of decreasing demand is equal to that of increasing supply, and $\rho^* > 1$, then the optimal strategy is to decrease

the demand rate to less than the current supply rate ($\lambda^* = \mu_0/\rho^* < \mu_0$), rather than increasing the supply rate to larger than the current demand rate ($\mu^* = \lambda_0\ \rho^* > \lambda_0$).

In case $\rho^* < \rho_0$, we use the same argument as above to reach the following decision process: either increase the demand rate by one unit with a cost of c_λ^+, or decrease the supply rate by ρ^* units with a cost of $\rho^* c_\mu^-$; we are indifferent between the two options if $c_\lambda^+ = \rho^* c_\mu^-$. In summary, we find that

- If $\rho_0 > \rho^* > c_\lambda^+/c_\mu^-$, then the optimal strategy is to increase the demand rate to $\lambda^* = \mu_0/\rho^*$, keeping the supply rate at μ_0;
- If $\rho^* < c_\lambda^+/c_\mu^-$, then the optimal strategy is to decrease the supply rate to $\mu^* = \lambda_0\ \rho^*$, keeping the demand rate at λ_0;
- If $\rho^* = c_\lambda^+/c_\mu^- < \rho_0$, then there is indifference between increasing demand and decreasing supply.

Remark: With the same argument as above, increasing the demand rate beyond $\lambda^* = \mu_0/\rho^*$, or decreasing the supply rate beyond $\mu^* = \lambda_0\ \rho^*$ results in higher cost, and is therefore non-optimal.

Corollary: Suppose $c_\lambda^-/c_\mu^+ = c_\lambda^+/c_\mu^- = \alpha$. If $\rho^* > \alpha$, then the optimal strategy is to manipulate the demand rate only to achieve the optimal balance; if $\rho^* < \alpha$, then the optimal strategy calls for adjusting only the supply rate to attain ρ^*. This result is reached by combining the findings for $\rho^* > \rho_0$ and $\rho^* < \rho_0$ cases.

To summarize this section, we show that when the system is a cost center with no throughput considerations, the minimization of cost of imbalance and cost of balancing are decoupled, and the minimum total system cost ($C^{T*} = C^{I*} + C^{B*}$) can be achieved in a stratified decision process. While CI^* is achieved by finding the optimal supply/demand balance (ρ^*), C^{B*} is achieved by figuring out how to achieve that optimal balance, either by manipulating demand, or by manipulating supply. From a decision-making point of view, (ρ^*) is an operational decision, whereas the choice of how to achieve it is a strategic decision. In certain environments, industries and businesses, there may be a governing strategy that suggests manipulating demand (or, supply) as a first course of action. Our study can be used as a framework to revisit such strategies.

In relation to the existing literature, this section complements and extends the work by Mendoza et al. (2009) and Mendoza et al. (2014) by providing a unified framework of queue imbalance in supply/demand systems, and by finding conditions on the optimal strategy of how to achieve the optimal queue balance.

CASE II: SYSTEMS WITH THROUGHPUT CONSIDERATIONS

In many supply/demand systems that can be modeled as a synchronization queue as studied in this paper, throughput is a very crucial aspect of system performance. For example, consider parallel processing and data synchronization (Parthasarathy et al., 1999) in computing, communication protocols in telecommunication networks, and other systems where efficient operation of the system as well as high volume of output is greatly desired, such as stock exchange, and just-in-time assembly systems (Som et al., 1994).

In order to incorporate the role of throughput in system performance, we assume in this section that there is a *fixed* unit net reward of r for each pairing of demand and supply. Therefore, the goal is to maximize total expected net rewards minus total expected costs.

Remark: The modeling framework presented in this section also captures supply/demand systems where there is a fee (or, price) collected for each service provided (or, goods sold), and the demand rate is not sensitive to fee amount, and fee (or, price) is virtually fixed (for example, due to intense competition). Brokerage firms, bike sharing systems (Raviv & Kolka, 2013) and job placement services and agencies in various sectors can be given as example.

The steady-state probability calculations carry over from Section 2, since the queuing system is modeled in the same way. We now calculate the throughput rate. Note that a successful pairing can occur only in one of two ways: 1) At a demand arrival when there is supply available (when $t < 0$), or 2) at a supply arrival when there is

demand waiting (when $t > 0$); since demand and supply arrival processes are independent Poisson, there is zero probability that a demand and supply arrival occurs at the same time. Thus, the mean throughput rate, θ, is given by

$$\theta = \lambda\, P(t < 0) + \mu\, P(t > 0)$$

$$= \begin{cases} \frac{\lambda\, \rho^{(1+d)}\,(1-\rho^{s})+\mu\,(1-\rho^{d})}{1-\rho^{(1+d+s)}}, & \rho \neq 1 \\ \\ \frac{\lambda\, s+\mu\, d}{1+d+s}, & \rho = 1 \end{cases}. \tag{3}$$

Total expected net rewards is equal to

$$R = r\, \theta, \tag{4}$$

and the objective is to maximize total net rewards less total system cost, given by the following objective function:

$$\Pi = R - (C^{I} + C^{B}) \tag{5}$$

Observe that now the decision process cannot be readily decoupled, since R includes the rates of demand and supply individually (not just their ratio, ρ). Furthermore, observe that there seems to be a trivial solution to the problem where both λ and μ are increased infinitely large, resulting in infinite objective function value. We show below how we avoid this triviality.

First, combining Eq.'s (3) and (4), we show that the total expected net rewards can be written in two equivalent forms:

Form 1: $$R_{\lambda} = \begin{cases} r\,\lambda\, \dfrac{\rho - \rho^{(1+d+s)}}{1-\rho^{(1+d+s)}}, & \rho \neq 1 \\ \\ r\,\lambda\, \dfrac{d+s}{1+d+s}, & \rho = 1 \end{cases} \tag{6}$$

Form 2: $$R_{\mu} = \begin{cases} r\,\mu\, \dfrac{1 - \rho^{(d+s)}}{1-\rho^{(1+d+s)}}, & \rho \neq 1 \\ \\ r\,\mu\, \dfrac{d+s}{1+d+s}, & \rho = 1 \end{cases} \tag{7}$$

In the first form, we eliminate μ from the expression; in the second, we eliminate λ. This enables us to look at the problem from two perspectives: in Form 1, the problem can be expressed as "finding the optimal ρ while keeping λ constant;" in form 2, it can be viewed as "finding the optimal ρ while keeping μ constant."

In Form 1, for a constant λ to be optimal at any given ρ value, we need $\partial\, \Pi / \partial\, \lambda < 0$ at any fixed ρ. Using Eq. (2), and the fact that $\partial\, C^{I} / \partial\, \lambda = 0$ from Eq. (2), we need $\partial\, R_{\lambda} / \partial\, \lambda < \partial\, C^{B} / \partial\, \lambda$ for any fixed ρ. Note from Eq. (6) that

$$\frac{\partial\, R_{\lambda}}{\partial\, \lambda} = \begin{cases} r\, \dfrac{\rho - \rho^{(1+d+s)}}{1-\rho^{(1+d+s)}}, & \rho \neq 1 \\ \\ r\, \dfrac{d+s}{1+d+s}, & \rho = 1 \end{cases}. \tag{8}$$

Note also that, since $\rho = \mu/\lambda$, increasing λ by one unit while fixing ρ requires increasing μ by ρ. Therefore,

$$\frac{\partial C^B}{\partial \lambda} = c_\lambda^+ + \rho\, c_\mu^+. \tag{9}$$

Similarly, in Form 2, for a constant μ to be optimal for any given ρ, we need $\partial \Pi/\partial \lambda < 0$, or equivalently $\partial R_\mu/\partial \mu < \partial C^B/\partial \mu$, since $\partial C^I/\partial \mu = 0$ from Eq. (2). From Eq. (7), we have

$$\frac{\partial R_\mu}{\partial \mu} = \begin{cases} r\,\dfrac{1-\rho^{(d+s)}}{1-\rho^{(1+d+s)}}, & \rho \neq 1 \\ \\ r\,\dfrac{d+s}{1+d+s}, & \rho = 1 \end{cases}, \tag{10}$$

and since $\rho = \mu/\lambda$, we get

$$\frac{\partial C^B}{\partial \mu} = \frac{1}{\rho}\, c_\lambda^+ + c_\mu^+. \tag{11}$$

Utilizing Eq.'s (8) through (11), we find that the conditions $\partial R_\lambda/\partial \lambda < \partial C^B/\partial \lambda$ and $\partial R_\mu/\partial \mu < \partial C^B/\partial \mu$ have equivalent requirements:

$$\left\{\frac{\partial R_\lambda}{\partial \lambda} < \frac{\partial C^B}{\partial \lambda}\right\} \equiv \left\{\frac{\partial R_\mu}{\partial \mu} < \frac{\partial C^B}{\partial \mu}\right\} \equiv \begin{cases} r < \left(c_\lambda^+ + \rho\, c_\mu^+\right)\dfrac{1-\rho^{(1+d+s)}}{\rho-\rho^{(1+d+s)}}, & \rho \neq 1 \\ \\ r < \left(c_\lambda^+ + c_\mu^+\right)\dfrac{1+d+s}{d+s}, & \rho = 1 \end{cases}.$$

From above, it is easy to verify that the right-hand side of the inequality for the $\rho = 1$ case greater than $\left(c_\lambda^+ + c_\mu^+\right)$. For the $\rho \neq 1$ case, we show the same as follows:

$$\begin{aligned} \left(c_\lambda^+ + \rho\, c_\mu^+\right)\frac{1-\rho^{(1+d+s)}}{\rho-\rho^{(1+d+s)}} &= \left(c_\lambda^+ + \rho\, c_\mu^+\right)\frac{1-\rho^{(1+d+s)}}{\rho-\rho^{(1+d+s)}} + \left(c_\lambda^+ + c_\mu^+\right) - \left(c_\lambda^+ + c_\mu^+\right) \\ &= \left(c_\lambda^+ + c_\mu^+\right) + c_\lambda^+\left(\frac{1-\rho}{\rho-\rho^{(1+d+s)}}\right) + c_\mu^+\left(\frac{(1-\rho)\,\rho^{(d+s)}}{1-\rho^{(d+s)}}\right) \\ &\geq \left(c_\lambda^+ + c_\mu^+\right), \quad \forall\, \rho. \end{aligned}$$

As a result, to guarantee for any given ρ value that there is no incentive to increase λ (in Form 1) and μ (in Form 2) unilaterally, we need the following assumption, without which we have an unbounded problem:

$$r < c_\lambda^+ + c_\mu^+. \tag{12}$$

Using Form 1, and therefore keeping the demand rate constant at the current value λ_0, suppose $\rho_{\lambda_0}^*$ is the optimal balance maximizing the objective function given in Eq. (5)**Error! Reference source not found.**. Recall that C^B is not a function of ρ. Therefore, $\rho_{\lambda_0}^*$ satisfies

$$\left(R_{\lambda_0} - C^I\right)^* = \left(R_{\lambda_0}\left(\rho_{\lambda_0}^*\right) - C^I\left(\rho_{\lambda_0}^*\right)\right) = \max_\rho \left(R_{\lambda_0} - C^I\right),$$

where R_{λ_0} follows from Eq. (6) by substituting λ_0 for λ. Further, since we are keeping the demand rate at λ_0, the optimal supply rate (μ^*) and the resulting cost of balancing ($C_{\lambda_0}^{B*}$) is realized directly from Eq. (1):

$$\mu^* = \lambda_0 \, \rho_{\lambda_0}^*,$$

and

$$C_{\lambda_0}^{B*} = c_\mu^+ \left(\lambda_0 \, \rho_{\lambda_0}^* - \mu_0\right)^+ + c_\mu^- \left(\lambda_0 \, \rho_{\lambda_0}^* - \mu_0\right)^-.$$

As a result, the maximum objective function value using Form 1 (i.e. keeping the demand rate at λ_0) is equal to

$$\Pi_{\lambda_0}^* = \left(R_{\lambda_0} - C^I\right)^* - C_{\lambda_0}^{B*}.$$

Similarly for Form 2, where we keep the supply rate at μ_0, let $\rho_{\mu_0}^*$ be the optimal balance such that

$$\left(R_{\mu_0} - C^I\right)^* = \left(R_{\mu_0}\left(\rho_{\mu_0}^*\right) - C^I\left(\rho_{\mu_0}^*\right)\right) = \max_\rho \left(R_{\mu_0} - C^I\right),$$

where R_{μ_0} is found by substituting μ_0 for μ in Eq. (7). Then, the optimal demand rate (λ^*) and the resulting cost of balancing ($C_{\mu_0}^{B*}$) follow directly from Eq. (1):

$$\lambda^* = \frac{\mu_0}{\rho_{\mu_0}^*},$$

and

$$C_{\mu_0}^{B*} = c_\lambda^+ \left(\frac{\mu_0}{\rho_{\mu_0}^*} - \lambda_0\right)^+ + c_\lambda^- \left(\frac{\mu_0}{\rho_{\mu_0}^*} - \lambda_0\right)^-.$$

Therefore, using Form 2 (i.e. keeping the supply rate at μ_0), the maximum objective function value is given by

$$\Pi_{\mu_0}^* = \left(R_{\mu_0} - C^I\right)^* - C_{\mu_0}^{B*}.$$

Consequently, we determine the following solution framework:

- If $\Pi_{\lambda_0}^* > \Pi_{\mu_0}^*$, then it is optimal to keep the demand rate constant and adjust the supply rate accordingly to achieve the optimal balance. Hence, the optimal solution is the (ρ, λ, μ) triplet given by $(\rho, \lambda, \mu)^* = \left(\rho_{\lambda_0}^*, \ \lambda_0, \ \lambda_0 \, \rho_{\lambda_0}^*\right)$.

- If $\Pi_{\lambda_0}^* < \Pi_{\mu_0}^*$, then it is optimal to keep the supply rate constant and manipulate the demand rate accordingly to attain the optimal balance. Thus, the optimal solution is $(\rho, \lambda, \mu)^* = \left(\rho_{\mu_0}^*, \ \frac{\mu_0}{\rho_{\mu_0}^*}, \ \mu_0\right)$.

The direct corollary from this result is that given the description of the supply/demand problem addressed, where throughput is an important consideration, and given a fairly mild assumption as in Eq. (12), the operational decision of coming up with the best supply/demand balance is linked with the strategy of attaining that optimal balance. Furthermore, we ascertain that the optimal strategy is either to manipulate demand while keeping supply constant, or to adjust the supply while keeping demand constant.

CONCLUSIONS

In this paper, we study service environments that are stochastic supply/demand systems where demand and supply arrive in separate, independent streams according to Poisson distribution. Each demand (supply) arrival leave the system either when paired with a supply (demand), or when the associated queue is full; otherwise, they wait for a supply (demand) arrival in the demand (supply) queue. This type of queuing system has been referred to as double-ended queue, synchronization queue, or synchronizing queue in the literature.

Many real-life environments, ranging from a taxi stand to data synchronization in computing, function in a similar manner. However, these supply/demand systems have different characteristics and objectives. In some systems, performance is measured almost exclusively by looking at operational inefficiencies that result from mismatch between demand and supply arrivals. One can give organ allocation to donors (Zenios, 1999), and tenant assignment to public housing (Kaplan, 1987) as examples. In many others, throughput of the system (the rate of demand paired with supply) is a crucial measure of system performance; e.g., data synchronization (Parthasarathy et al., 1999), and just-in-time assembly systems (Som et al., 1994). Interestingly, literature has not addressed throughput considerations in our context.

We contribute to the existing literature in a number of ways. Firstly, building on the canonical double-ended queuing model, we provide a unified treatment of supply/demand mismatch in cost centers, and show that the decision-making process in minimizing system cost is decoupled. Specifically, we show that the operational decision of finding the best system balance (ρ^*) and the strategic decision of how to arrive at that optimal balance—either by manipulating demand, or by manipulating supply—can be stratified. Further, as a first in the literature, we determine analytical results for when it is best to manipulate demand instead of supply, and vice versa. We also draw interesting corollaries for special cases.

Furthermore, also first time in the literature, we investigate systems where throughput is a significant metric of system performance. In order to reflect the impact of this metric, we conceptualize a system where a constant reward is earned for each successful demand-supply pairing. We note that this generic framework also captures supply/demand systems where there is a fee (or, price) collected for each service provided (or, goods sold), and the demand rate is not sensitive to fee amount, and fee (or, price) is virtually fixed (for example, due to intense competition). Analytically studying the model, we provide a characterization of the optimal solution which states that the operational decision on optimal system balance and the strategic decision on how to attain the optimal balance (through either demand, or supply decisions) are strongly tied. In other words, a "manage only demand" or a "manage only supply" strategy may lead to suboptimal results if not used wisely.

The framework and the findings we present in this paper can shed light on the managerial decision making process in these environments, and they can be used to revisit certain policies—dictating management of demand or supply as first action—that are held as part of firm strategy.

We foresee two direct extensions of the modeling framework provided in this paper. First, in many service environments there are multiple interfaces between demand and supply, and it would be interesting to investigate whether our findings extend to such environments and if other valuable managerial insights can be drawn. Second, supply/demand systems where the demand is sensitive to price is of particular interest. In such environments, price can be used to regulate demand, and the relationship between price-elasticity and the resulting optimal strategies can provide an in-depth understanding on usage of pricing as a tactical means to attain desired system efficiency.

AUTHOR INFORMATION

Eylem Koca is an Assistant Professor of Operations Management at the Information Systems and Decision Sciences department at the Silberman College of Business at Fairleigh Dickinson University. He holds a Ph.D. in Operations Management/Management Science from Smith Business School at the University of Maryland, College Park. His M.S. in Industrial Engineering and B.S. in Mechanical Engineering degrees are from Bogazici University, Istanbul, Turkey. His research interests include operations management, operations/marketing interfaces, closed-loop supply

chains, and applications of decision theory. *Mailing Address:* Fairleigh Dickinson University, 1000 River Road, H-DH2-06, Teaneck, NJ 07666, USA. *E-mail*: koca@fdu.edu (Corresponding author)

Mohammad Sedaghat is a full-time faculty at the Information Systems and Decision Sciences department at the Silberman College of Business at Fairleigh Dickinson University. He obtained his Ph.D. in Operations Research from Polytechnic University, New York. His research interests include queuing theory and stochastic processes. *Mailing Address:* Fairleigh Dickinson University, 1000 River Road, H-DH2-06, Teaneck, NJ 07666, USA. *E-mail*: sedaghat@fdu.edu

K. Paul Yoon is Professor and Chair of Information Systems and Decision Sciences department at the Silberman College of Business at Fairleigh Dickinson University. His M.S. and Ph.D. degrees in Operations Research are from Kansas State University. His research areas include multiple criteria decision-making (MCDM) and its applications to service and production systems. He is the co-author of three books on MCDM. *Mailing Address:* Fairleigh Dickinson University, 1000 River Road, H-DH2-06, Teaneck, NJ 07666, USA. *E-mail*: yoon@fdu.edu

REFERENCES

1. Bollapragada, R., & Rao, U. S. (2006). Replenishment planning in discrete-time, capacitated, non-stationary, stochastic inventory systems. *IIE Transactions, 38*(7), 605–615.
2. Conolly, B. W., Parthasarathy, P. R., & Selvaraju, N. (2002). Double-ended queues with impatience. *Computers & Operations Research, 29*(14), 2053–2072.
3. Frutos, I. P., & Gallego, J. A. (1999). Multiproduct monopoly: a queuing approach. *Applied Economics, 31*(5), 565–576.
4. Hsu, G. H., He, Q. M., & Liu, X. S. (1993). Matched queuing systems with a double input. *Acta Mathematicae Applicatae Sinica, 9*(1), 50–62.
5. Kaplan, E. H. (1987). Analyzing tenant assignment policies. *Management Science, 33*(3), 395–408.
6. Kashyap, B. R. K. (1966). The double-ended queue with bulk service and limited waiting space. *Operations Research, 14*(5), 822–834.
7. Kendall, D. G. (1951). Some problems in the theory of queues. *Journal of the Royal Statistical Society. Series B (Methodological), 13*(2), 151–185.
8. Lander, G. H. (1986). A queuing model approach to stochastic supply/demand systems. *Journal of Information and Optimization Sciences, 7*(2), 137–144.
9. Latouche, G. (1981). Queues with paired customers. *Journal of Applied Probability, 18*(3), 684–696.
10. Li, N., & Jiang, Z. (2013). Modeling and optimization of a product-service system with additional service capacity and impatient customers. *Computers & Operations Research, 40*(8), 1923–1937.
11. Parthasarathy, P. R., Selvaraju, N., & Manimaran, G. (1999). A paired queuing system arising in multimedia synchronization. *Mathematical and Computer Modelling, 30*, 133–140.
12. Perry, D., & Stadje, W. (1999). Perishable inventory systems with impatient demands. *Mathematical Methods of Operations Research, 50*(1), 77–90.
13. Prabhakar, B., Bambos, N., & Mountford, T. S. (2000). The synchronization of Poisson processes and queuing networks with service and synchronization nodes. *Advances in Applied Probability, 32*(3), 824–843.
14. Raviv, T., & Kolka, O. (2013). Optimal inventory management of a bike-sharing station. *IIE Transactions, 45*(10), 1077–1093.
15. Sasieni, M. W. (1961). Double queues and impatient customers with an application to inventory theory. *Operations Research, 9*(6), 771–781.
16. Som, P., Wilhelm, W. E., & Disney, R. L. (1994). Kitting process in a stochastic assembly system. *Queueing Systems, 17*, 471–490.
17. Takahashi, M., Osawa H., & Fujisawa T. (2000). On a synchronization queue with two finite buffers. *Queueing Systems, 36*, 107–123.
18. Zenios, S. A. (1999). Modeling the transplant waiting list: a queuing model with reneging. *Queueing Systems, 31*, 239–251.