Optimal Order Processing Policies for E-commerce Servers

Yong Tan • Vijay S Mookerjee • Kamran Moinzadeh

University of Washington Business School, Box 353200,
Seattle, Washington 98195-3200, USA

University of Texas at Dallas, School of Management, PO Box 830688, JO44
Richardson, TX 75803-0688, USA

University of Washington Business School, Box 353200,
Seattle, Washington 98195-3200, USA

ytan@u.washington.edu • *vijaym@utdallas.edu* • *kamran@u.washington.edu*

The explosive growth in online shopping has provided online retailers impressive opportunities for revenue and profit. At the same time, retailers may lose considerable online business from slow response times at electronic shopping sites. Although increasing server capacity may improve response time, the resources and capital needed to do so are clearly not free. In this study, we propose a scheme that can improve a server's performance under its current capacity. This scheme is based on priority order processing, where the priority of an order depends on the potential revenue that would be generated from the order. The results for single-period analysis show that the benefit from priority processing increases as the server becomes busier.

We have also modeled a multi-period version of the problem, where the demand in a period depends on the Quality of Service (QoS) that buyers receive in the previous period. In multi-period problem, both the server capacity and the order processing policy in each period are determined optimally. Here, it may be optimal for the retailer to sacrifice profit and increase QoS in initial periods in order to increase demand (and revenue) in later periods. Thus, the order processing policy of the server evolves from an emphasis on QoS in initial periods, to one on profit in subsequent periods.

(*Electronic Commerce; Priority Order Processing; Queues; Reneging*)

1. Introduction

The last few years have observed an explosive growth in online business. This growth is likely to sustain since the number of Internet users around the world is continues to grow. The Computer Industry Almanac has reported that by the year 2002, 490 million people around the world will have Internet access, that is, 79.4 per 1,000 people worldwide, and 118 people per 1,000 by year-end 2005.

More significantly, Jupiter Communications estimates that online shopping will grow to $78 billion by Year 2003. According to Jupiter Communications, retailers seeking to benefit from the huge online market must design the online experience with a view to establish long term relationships with new online buyers. This is because meeting and exceeding the expectations of new buyers has an effect far beyond the current season; a positive online experience with a specific retailer could go a long way toward securing a future stream of revenues.

In this study, we consider the problem of congestion in e-commerce servers. Our concern is specifically to do with slow response time. A significant statistic here is one provided by Forrester Research Inc. – 66% of all electronic shopping carts are abandoned before a sale is completed. While the abandoning of a cart may occur from reasons that are not related to response time, it is quite clear that slow response time is indeed a significant cause. One study entitled "The Web Is No Shopper's Paradise," published in the November 1999 issue of *Fortune* magazine states that (Mardesich, 1999):

> *"Among the 200 Web consumers surveyed for Cognitiative's (a Net strategy consulting firm based in San Francisco) quarterly "Pulse of the Customer" report, the No. 1 reason that customers got fed up and took their business*

> *elsewhere was technical problems, including unacceptably slow response times."*

On similar lines, Green (1999) conducted a more detailed survey using 1000 online shoppers to study congestion problems in e-commerce servers. The survey revealed that 28% of the respondents found sites too slow and left, never to return.

An obvious solution to the above congestion problems is to increase server capacity. An interesting case in point here is the e-merchant "800.com," that sells consumer electronics. After its launch in October 1998, the company tried launched a promotion that offered three movies or three CDs for $1. The promotion worked, perhaps too well. On November 27, 1998 the company's site was swamped and virtually shut down as hundreds of thousands of customers flocked to the site to obtain cheap movies and CDs. Based on this experience, 800.com now relies on 50 rather than 16 main servers to handle site traffic! However, while adding more servers should clearly alleviate congestion problems; there is no easy formula to determine the optimal number of computers. Furthermore, adding capacity is obviously not free and should be done only if the additional capacity can be economically justified.

A resource neutral solution to the server congestion problem is to improve performance using priority order processing. There could be many dimensions of performance (e.g., average response time, throughput, maximum response time, revenue generated, etc.); our focus here is on two of these measures: profit per unit time and percentage of buyers lost. While profit is an obvious performance measure, the percentage of buyers lost is used to measure Quality of Service (QoS). A buyer's satisfaction with the electronic shopping process should depend, in part, on whether or not the purchase transaction is successfully completed in a reasonable amount of time. Therefore, the percentage of buyers lost can be used as a measure of the quality of service achieved by a certain electronic site. When profit is the goal of the processing policy, the amount of processing power assigned to an order depends on two considerations: (a) the potential revenue that would be generated from the order if it is successfully executed, and (b) the probability that the order will be lost because the customer's delay tolerance is exceeded. On the other hand, when QoS is the goal, the amount of processing power assigned to an order depends only on the probability that the order will be lost.

In deriving the optimal policies, our study spans three research areas: time-shared systems, reneging in queuing theory, and differentiated (or prioritized) services. The concept of time-sharing has been widely adopted in scheduling computer processing (Kleinrock, 1964, 1967, 1976; Stallings, 1997). In time-sharing, processing power can be divided in two ways. The first method, often called "Round-Robin", is a true time-sharing policy. Each process gets a slice (quantum) of time during which the full capacity of the processor is dedicated to the process. The second method is to divide processing power on a full-time but part-capacity basis. This method, referred to as "Processor-Sharing," represents the continuous limit of time-sharing where the time slice approaches zero. Kleinrock (1964, 1967) obtained the expected response time (time spent in the system, waiting time plus required processing time) for exponentially distributed arrival and departure (time-shared M/M/1). The waiting time is proportional to the processing time attained; suggesting that round robin systems are more "fair" than sequential queues since the waiting time in sequential systems does not depend on the actual amount of processing incurred on a job. Coffman *et al* (1970) derived the expression for the waiting time distribution in time-shared systems that was found to deviate from the exponential form that describes the sequential M/M/1 queuing model. Morrison (1985) later found the response time (waiting plus required process time) distribution for M/M/1 processor-sharing systems. Sakata, Noguchi, and Oizumi (1971), and O'Donovan (1974) have extended the results of expected response time to

M/G/1 processor-sharing models.

Customer impatience is a direct source for loss of sales in online retailing; customers quit shopping if the waiting exceeds their tolerance. This phenomenon, termed as queuing with reneging (and balking), has been studied for about 40 years and is still an active research area. Barrer (1957) obtained the results for the queuing problem with impatient customers based on heuristics. Gnedenko and Kovalenko (1989) provided a more rigorous analysis that verifies Barrer's results. In their studies, they assume a M/M/1 system with either a constant or exponentially distributed reneging time. More detailed studies on M/M/1 system with an exponentially distributed reneging time have been presented by Ancker and Gafarian (1962, 1963). Rao (1968) extended the results to the reneging in M/G/1 systems. In both studies, an exponentially distributed reneging time was assumed. Recent studies are focused on extending to more general queuing systems, for example a GI/G/1 model (Stanford, 1979). In addition, Whitt (1999) studied the reneging problem when customers are informed about anticipated delays. There has been no research to our knowledge that includes the reneging in the time-shared queuing systems. Assaf and Haviv (1990) have studied the strategy of reneging from processor sharing systems. However, the impact of reneging on the queues is not explicitly modeled.

Kleinrock (1967) introduced the notion of priority processing for computer processors and derived the expected waiting time for different priority classes with exponentially distributed processing time. O'Donovan (1974) showed that the result holds even for M/G/1 processor-sharing systems. Similarly in networking, the vision of providing differentiated services has been around for well over a decade (Turner, 1986). Future packet networks will likely support Quality of Service in order to provide a full array of services (Aiello et al, 2000). Research in this area is becoming increasingly popular, as new technology makes implementation possible. Other studies have been focused on economic side, namely pricing for providing priority services in communication networks as well as in other settings (Marchand, 1974).

In this study, we propose and analyze optimal order processing policies that can be used by e-commerce servers to improve performance. We also provide a variety of analytical and numerical results that prescribe how E-retailers should best run their servers to achieve their performance goals. We also study a multi-period version of the server congestion problem, where both the server capacity and the processing policy in each period are optimally determined. In the multi-period problem, we consider QoS externalities; demand in a given period depends on the QoS achieved in the previous period. The multi-period problem is solved using dynamic programming. The results from the multi-period analysis can provide guidelines for e-retailers to adjust the order processing policy over time. For example, the analysis reveals that e-retailers should initially pay more attention to QoS. However, as the customer base of an e-retailer matures, profit-oriented order processing policies should be pursued.

The rest of the paper is organized as follows. Sections 2 and 3 are dedicated to the analysis of optimal server operating policies. A single period model is presented in Section 2 whereas Section 3 extends the single period model to a multi-period one. Numerical results are presented and practical implications are discussed. Section 4 provides a summary of results and offers directions for future research.

2. The Model

In this section, we set up the basic model where the demand is assumed to be stationary. This is a single period problem. E-retailers first determine and acquire the optimal capacity that remains unchanged for the rest of the planning horizon.

2.1 Assumptions and Preliminaries

We characterize orders arriving at an E-commerce server by their values. The predicted final purchase value of an order, h, follows a value distribution, $f(h)$. One way to predict the final purchase value is to use the customer's historic data, for example, an average value (or a moving average). After each purchase, the data can be modified and a new average value can be calculated for use for the next visit. For a first time buyer, a typical value, such as the market average for all new buyers can be used. This study assumes a "static" policy, i.e., each customer is assigned an estimated value that remains unchanged throughout the course of shopping. In a more sophisticated model, the predicted final value can vary "dynamically," for example, by utilizing information about the value of items in the shopping cart. These and other methods (e.g., data mining techniques) can be applied. A dynamical policy offers more accurate value prediction, however with the cost of extra computational overhead. A static policy estimates the value "offline", therefore does not consume online resources but may suffer from inaccuracy. We will attempt to address this tradeoff question in subsequent studies.

The total arrival rate of orders is l, this is the number of orders per unit time. We assume that orders arrive according to a Poisson process, which represents real situation quite well (Moe and Fader, 2001). Each order requires certain amount of time to be processed. A typical online shopping process involves various activities, for example, browsing, searching, adding items to shopping cart, and checkout. For a static policy, we ignore the details of involved shopping activities (as these information are not used to update the predicted value), and instead define the required processing time as the total time that an order receives processing from the server including all shopping activities. The required processing time for orders with value h is assumed to be exponentially distributed with an average time $t(h)$, which is a function of the value.

Therefore the overall processing time distribution for all orders is described by the hyper-exponential distribution, or a linear combination of exponential distributions with different means. The assumption of exponential processing time simplifies policy analysis. We have not attempted to generalize this assumption; however, it is a good approximation for time-shared systems where it has been shown (Sakata, Noguchi, and Oizumi, 1971; O'Donovan, 1974) that several important results (such as mean delays) depend only on the mean processing time and not on the distribution. In general, average required processing time is a non-decreasing function of the value. In our model, any function form of $t(h)$ can be used.

Impatient customers may quit shopping if the time spent exceeds their tolerance for waiting. We define the tolerance level of a shopper in terms of average time in the server, $w(h)$, that is, buyers with the value h are willing to wait for a random time that is exponentially distributed (Ancker and Gafarian, 1962) with a mean $w(h)$. It is likely that customer impatience varies during the course of shopping. Similarly for a static policy, we suppress this variation and adopt an average level over various points of the shopping process. The impatience of an on-line shopper is modeled using an analogy from the "express lane" in a traditional grocery store. Typically store express lanes provide faster service to customers with relatively fewer items. The implicit assumption here is that customers with fewer items may be less tolerant of delays; more generally, the tolerance level of a customer is assumed to depend in some way on the number of items that the customer intends to purchase. In our model, the delay tolerance $w(h)$ can be any function of h; however, it is plausible that $w(h)$ is increasing in h.

Similar to the scheduling of a computer processor, server capacity is shared by orders in the server. The sharing scheme in a typical e-commerce server is based upon "Round-Robin" time-sharing. Specifically, an order is given a slice of processing time, say Q, when it enters the processing unit of the server. It exits from the system if it finishes the desired processing during this allotted time. Otherwise, it goes back to the end of the queue and waits for its next turn. Round-Robin scheduling is better than a sequential scheduling (in which customers are served one at a time) because sequential scheduling may waste capacity while waiting for a client's response.

We introduce the prioritized processing by assigning a weight g_k ($g_k \geq 0$) for orders in priority class-k so they receive processing time, g_kQ. This scheme was first proposed by Kleinrock (1967). The priority of an order is determined by the potential final value of the purchase that a buyer makes. In this study, we limit ourselves to the case where there are an infinite number of classes. This allows us to focus on the properties of priority scheme without involving complicated problem of assigning orders to classes. The processing weight g_k becomes continuous and is a function of value h, $g(h)$. Offering discrete classes may be optimal if the cost of implementing the policy is considered. A model for discrete classes has been published elsewhere (Tan, 2000).

As mentioned earlier, customers will abandon their shopping cart (or renege) if they have been made to wait for too long. We have solved a queuing problem that incorporates reneging in a priority-based time-sharing system. The result is summarized in the following proposition.

Proposition 1 *The loss function density (defined as number of customers lost per unit time per unit value),* $l(h)$, *for orders with value h, is*

$$l(h) = \frac{t(h)\big(g(h)+1\big)}{t(h)\big(g(h)+1\big)+w(h)g(h)}\, lf(h),$$

where $g(h)$ *must satisfy,*

$$\int_{h\in H} \frac{w(h)t(h)\big(g(h)+1\big)g(h)}{t(h)\big(g(h)+1\big)+w(h)g(h)}\, lf(h)dh = 1, \tag{1}$$

and H is the set of all possible values.

The proof is presented in the Appendix, where the proofs for all the propositions, corollaries and lemmas can be found.

In the next section, we introduce a multi-period version of this model where capacity choices can be made at the beginning of each period. In this section, however, we have assumed that capacity can be chosen once, at the beginning of the period. The effect of increasing server capacity is to reduce the required processing time of an order:

$$t(h) = t^0(h)/C\,, \tag{2}$$

where $t^0(h)$ is the processing time for a standard unit of capacity, and C is the server capacity. A more powerful server has a larger value of C, and therefore orders can be processed faster. There is a cost associated with acquiring capacity. We assume a linear cost, gC, where g is the unit cost (normalized to per unit time). The linear form can be justified as in most cases capacity can be additive, for example, if more servers are added. We ignore the fixed cost, as it does not affect our results.

2.2 Profit-focused Policy

The total expected revenue (per unit time) can be written as,

$$S = \int_{h\in H} h\big(lf(h) - l(h)\big)dh\,. \tag{3}$$

A retailer's objective is to choose capacity and processing weight such that the expected profit per unit time is maximized. Explicitly, we define an E-retailer's problem:

$$\max_{g(h),C}\ p\big(g(h),C\big) = \int_{h\in H} \frac{w(h)Cg(h)h}{t^0(h)\big(g(h)+1\big) + w(h)Cg(h)}\, lf(h)dh - gC\,,$$

subject to

$$1 = \int_{h\in H} \frac{w(h)t^{0}(h)\big(g(h)+1\big)g(h)}{t^{0}(h)\big(g(h)+1\big)+w(h)Cg(h)}\, l f(h)dh, \qquad (4)$$

$$g(h) \geq 0, \quad \forall h \in H .$$

Equation 4 is identical to Equation 1 with capacity C explicitly expressed. The solution to this problem is given in the following proposition:

Proposition 2 *The optimal processing weight allocation is*

$$g^{*}(h) = \begin{cases} \dfrac{1}{1+w(h)/t(h)}\left(\left(1+\left(\dfrac{1}{y}\dfrac{h}{t(h)}-1\right)\left(1+\dfrac{w(h)}{t(h)}\right)\right)^{1/2}-1\right) & h\in H^{S}; \\ 0, & h\notin H^{S}; \end{cases}$$

where y is the Lagrange multiplier associated with Equation 4, and the serviceable set H^S is defined as,

$$H^{S} = \left\{ h : \frac{h}{t(h)} \geq y\, , h\in H \right\}.$$

The optimal capacity is,

$$C^{*} = \frac{l\,y}{g} \int_{h\in H} \frac{w(h)t(h)\big(g^{*}(h)+1\big)^{2} g^{*}(h)}{t(h)\big(g^{*}(h)+1\big)+w(h)g^{*}(h)}\, f(h)dh .$$

Proposition 2 indicates that there is a revenue realization threshold defined by $h/t(h) \geq y$; only orders with revenue realization above this threshold receive processing capacity. The revenue realization threshold is rate at which an order's revenue is realized. The value of y can be obtained by substituting the expressions of $g^{*}(h)$ and C^{*} in Equation 4. The problem is reduced to solving for y and C^{*}. The following corollary provides bounds for numerical search.

Corollary 1 *The optimal capacity C^{*} satisfies the following inequalities:*

$$y \leq gC^{*} \leq S^{*}.$$

S^{*} is the revenue when the optimal policy is adopted. The first inequality indicates that an order will get processed if it can recover the cost of capacity. The system can also afford to admit some orders whose values are below the cost of capacity. The second inequality guarantees a non-negative profit; representing the individual rationality condition for the retailer.

We have shown the following comparative static results hold.

Corollary 2 *For the optimal processing weight allocation $g(h)$,*

i. $$\frac{\partial g^*(h)}{\partial \left(h / t(h)\right)} > 0;$$

ii. $$\frac{\partial g^*(h)}{\partial \left(w(h) / t(h)\right)} < 0.$$

Orders with higher rate of revenue realization (i.e., the $h/t(h)$ ratio) receive more processing time. The ratio $w(h)/t(h)$ measures a customer's patience level. Orders with a higher value of this ratio (more patient buyers) can tolerate more delay and hence receive less processing time.

Figure 1 shows the optimal processing weight, $g^*(h)$, for linear processing time and waiting tolerance: $t(h) = 0.7 + 0.3h$, and $w(h) = 1 + 0.5h$. Here we assume that capacity is fixed, so only the threshold y needs to be calculated. The value distribution is uniform, namely, $f(h) = 1$ for $h \in [0, 1]$. $g^*(h)$ increases with value h as the ratio $h/t(h)$ is an increasing function of h while $w(h)/t(h)$ decreases with h. However in Figure 2, with nonlinear forms of $t(h) = 0.3 + 0.7h^3$, and $w(h) = 0.5 + 0.8h^2$, $g^*(h)$ is not monotonic. It turns out that the ratio $h/t(h)$ peaks around the value $h \approx 0.6$. This example shows the importance of gathering information on customer purchasing behavior (e.g., the delay tolerance); a higher value (h) alone does not warrant higher priority.

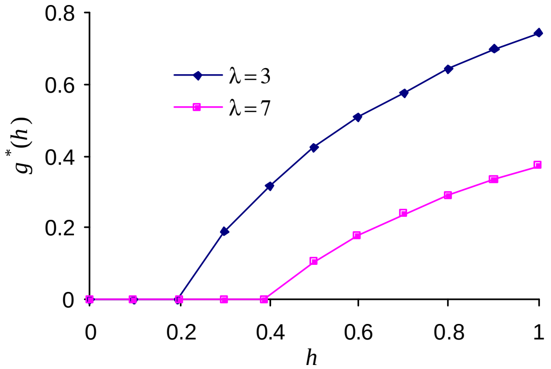

Figure 1. Optimal processing weight, $g^*(h)$, plotted as a function of value, h, for linear processing time and waiting tolerance.

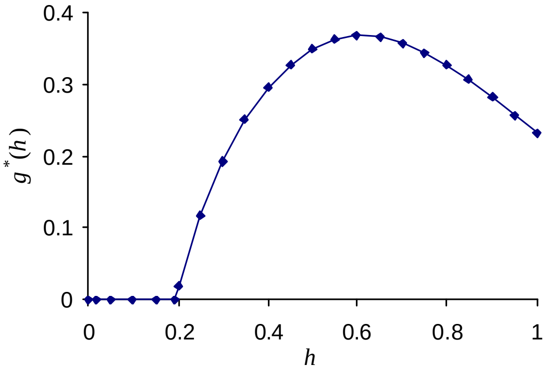

Figure 2. Optimal processing weight, $g^*(h)$, plotted as a function of value, h, for non-linear processing time and waiting tolerance.

Corollary 3 *With the change of demand,* l*, the following results hold:*

i. $$\frac{\partial p^*}{\partial l} > 0;$$

ii. $$\frac{\partial C^*}{\partial l} > 0;$$

iii. $$\frac{\partial}{\partial l}\left(\frac{y}{C^*}\right) > 0, \; if \; \frac{y}{C^*} > a_0,$$

where the expression for the coefficient a_0 *is given in the proof.*

When the total demand increases, the expected profit (i) will also increase. This can be achieved by adding more capacity (ii). The server becomes more discriminating by raising the threshold (iii) when the threshold exceeds a certain level. The threshold y/C^* translates to value threshold h_c,

$$\frac{h_c}{t^0(h_c)} = \frac{y}{C^*}.$$

Figure 3 shows the value threshold h_c as a function of demand, with the same parameter setting of Figure 1. If the capacity is fixed, an E-retailer has to drop more low value orders so that higher priority orders are more likely to complete. Otherwise, the capacity will be adjusted optimally with demand, without much increase in the degree of differentiation.

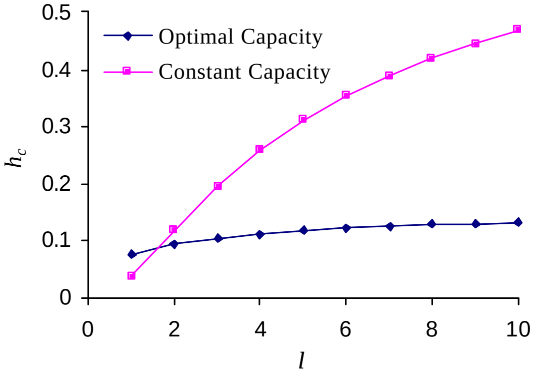

Figure 3. Value threshold for serviceable orders, h_c, plotted as a function of demand, l for fixed capacity and capacity determined optimally.

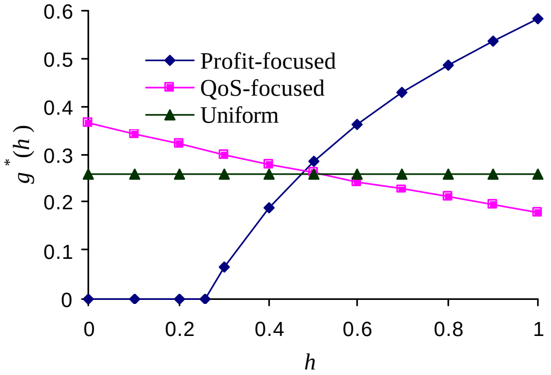

Figure 4. Optimal processing weight, $g^*(h)$, plotted as a function of value, h, for profit-focused, QoS-oriented, and uniform (non-differentiated) policies.

Corollary 4 *With the change of capacity cost, g, the following results hold:*

i. $$\frac{\partial p^*}{\partial g} < 0;$$

ii. $$\frac{\partial C^*}{\partial g} < 0;$$

iii. $$\frac{\partial}{\partial g}\left(\frac{y}{C^*}\right) > 0.$$

It is intuitive that when the cost of capacity increases, the expected profit (i) and capacity (ii) will drop. When the cost of capacity increases, the server becomes more discriminating (iii) by raising the threshold.

2.3 QoS-focused Policy

In this subsection, we consider a policy that focuses on the Quality of Service (QoS). We define the QoS as the number of lost orders per unit time, regardless of their order value. The E-retailers' problem becomes,

$$\min_{g(h)} \; L \equiv \int_{h \in H} l(h) dh \,, \tag{5}$$

where $l(h)$ is the loss function density, given in Proposition 1.

The optimal processing allocation can be obtained similarly,

$$g^*(h) = \begin{cases} \dfrac{1}{1 + w(h)/t(h)} \left(\left(\left(1 + \left(\dfrac{t_c}{t(h)} - 1 \right) \left(1 + \dfrac{w(h)}{t(h)} \right) \right) \right)^{1/2} - 1 \right), & t(h) < t_c; \\ 0, & t(h) \geq t_c. \end{cases}$$

The processing time threshold t_c can be found by substituting the above expression in Equation 1. It is obvious from the above equation that orders more processing will be assigned less processing time. More patient buyers (with higher $w(h)/t(h)$ ratio) also receive less processing time. The value threshold h_c is given by $t(h_c) = t_c$. Assuming $t(h)$ increases with the value h, orders with value *above* the threshold h_c will not receive any processing capacity.

We next compare three policies: profit-focused, QoS-focused, and uniform. The uniform policy is a non-discriminating policy where $g(h)$ is a constant. Here, linear forms of $t(h) = 0.7 + 0.3h$ and $w(h) = 1 + 0.5h$ are used, and the capacity C is set to be 1. Figure 4 clearly shows the difference between profit-focused and QoS-oriented policies. QoS policies favor low value customers, as orders from such buyers consume less processing time. Figure 5 plots the expected revenue (or profit as the capacity is fixed). The QoS-focused policy is the worst performer as it favors low value orders (with shorter processing time). For the Profit-focused policy, the improvement in revenue increases when server becomes busier (higher l). Figure 6 plots the loss ratio of customers, L/l, a measure for the quality of service. The QoS-focused policy is most effective in the middle range of demand, since L/l converges to 0 as $l \to 0$, and 1 as $l \to \infty$, for all policies.

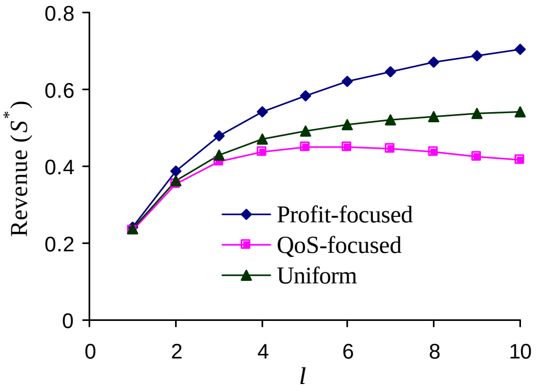

Figure 5. Sales, plotted against demand, l, for profit-focused, QoS-oriented, and uniform (non-differentiated) policies.

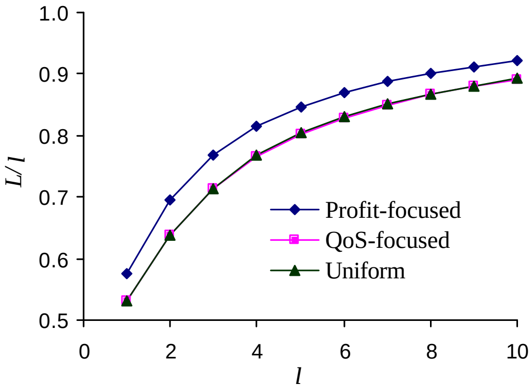

Figure 6. QoS (percent lost customers, L/l), against demand, l, for profit-focused, QoS-oriented, and uniform policies.

3. Multi-Period Model

In Section 2, we presented various polices that attempt to optimize an E-retailer's profit in a single period. However, retailers often value their customer base more than immediate profit. This consideration is not without merit since lost customers rarely come back. This suggests that during initial periods, the value of orders should be paid less importance so as to build a solid customer base. In this section, we propose a multi-period model that includes feedback on quality of service. Specifically, the quality of service that customers receive in a period affects the demand in the next period. E-retailers are allowed to vary capacity to match the demand.

We add an index, *j* for the period, to the notation used in the previous section. We start with the loss function density $l_j(h)$ in the *j*-th period. The expression for $l_j(h)$ is the same as the one shown in Proposition 1 with subscripts for the period *j* wherever necessary. The expected number of buyers lost per unit time in the *j*-th period, L_j, is calculated using Equation 5 and $l_j(h)$.

We model the demand in the $(j + 1)$-th period in the following way:

$$l_{j+1} = L_j \cdot p_j + \left(l_j - L_j \right) \cdot r_j , \qquad (6)$$

where, p_j is the probability that unsatisfied buyers return and r_j is the rate of growth in demand due to successful purchases, for example through “word of mouth”. Equation 6 can also be expressed in terms of the expected percentage of buyers lost, L_j/ l_j, a measure of QoS.

E-retailers are allowed to adjust capacity, both upwards and downwards; implying that we accommodate situations in which capacity can be rented. We assume that the cost of acquiring capacity is proportional to the capacity added, namely, $g_j(C_j - C_{j-1})$, where C_j is the capacity in the j-th period. Similarly as in Equation 2, we have $t_j(h) = t_j^0(h)/C_j$. Since hardware cost typically decreases with time, the coefficient g_j is assumed to decrease from period to period.

Without loss of generality, we assume that the demand becomes stationary in the N-th period. E-retailers have N opportunities to adjust capacity, after which the capacity remains unchanged from the N-th period onwards. The discounted profit is,

$$\max_{g(h),C} \sum_{j=1}^{\infty} d^{j-1} p_j = \sum_{j=1}^{N} d^{j-1}\left(S_j - g_j(C_j - C_{j-1})\right) + \frac{d^N}{1-d} S_N ,$$

where d is the discount factor; and $g(h)$ and C are short hands for the processing weight and capacity vectors. The profit, $p_j = S_N$, for $j \geq N + 1$, because the operating policy and capacity remain unchanged from the N-th period onwards. The discounted profit can be represented by a finite horizon formulation, $\sum_{j=1}^{N} d^{j-1} p_j$, where p_N is redefined as,

$$p_N = \frac{S_N}{1-d} - g_N(C_N - C_{N-1}) .$$

We can solve this multi-period problem using the method of dynamic programming (Dreyfus and Law, 1977), more specifically, backward induction. Let us define

$$\Pi_j(l_j, C_{j-1}) \equiv \max_{g_l(h),C_l} \sum_{l=j}^{N} d^{l-j} p_l\left(l_l, C_{l-1}; g_l(h), C_l\right),$$

where $g_l(h)$ and C_l, for $l \geq j$, are the decision variables. l_l is the demand for the l-th period, which can be obtained recursively using the demand generation model as specified by Equation 6. $\Pi_j(l_j, C_{j-1})$ is the maximum discounted profit starting from the j-th period. It depends on the demand arriving in this period, l_j, and the capacity carried over, C_{j-1}. This allows us to write the recursive relation,

$$\Pi_j(l_j, C_{j-1}) \equiv \max_{g_j(h), C_j} \; p_j\left(l_j, C_{j-1}; g_j(h), C_j\right) + d \cdot \Pi_{j+1}(l_{j+1}, C_j), \tag{7}$$

where $\Pi_0(l_0)$ is the objective function.

We first introduce the following lemma.

Lemma 1 *The effect of carry-over capacity on the maximum discounted profit is described as below,*

$$\frac{\partial \Pi_j}{\partial C_{j-1}} = g_j.$$

This result is intuitive as more existing capacity reduces the cost of capacity in the current period, and consequently increases the discounted profit. Now we are in a position to characterize the optimal policies.

Proposition 3 *The optimal processing weight allocation in the j-th period ($j < N$) is*

$$g_j^*(h) = \begin{cases} \dfrac{1}{1 + w_j(h)/t_j(h)} \left(\left(1 + \left(\dfrac{k_j(h)}{y_j} - 1 \right) \left(1 + \dfrac{w_j(h)}{t_j(h)} \right) \right)^{1/2} - 1 \right), & h \in H_j^S; \\ 0, & h \notin H_j^S; \end{cases}$$

where the serviceable value set H_j^S is defined as,

$$H_j^S = \left\{ h : k_j(h) \geq y_j, h \in H_j \right\},$$

and

$$k_j(h) = \frac{h}{t_j(h)} + d \frac{\partial \Pi_{j+1}}{\partial l_{j+1}} (r_j - p_j) \frac{1}{t_j(h)}. \tag{8}$$

The optimal capacity is,

$$C_j^* = \frac{l_j y_j}{g_j - dg_{j+1}} \int_{h \in H_j} \frac{w_j(h) t_j(h) \left(g_j^*(h)+1\right)^2 g_j^*(h)}{t_j(h)\left(g_j^*(h)+1\right) + w_j(h) g_j^*(h)} f_j(h) dh \, .$$

For periods j ≥ N, the processing weight and capacity are given by Proposition 2 with the effective unit capacity cost, $(1 - d)g_N$*.*

As displayed in Equation 8, this policy seems to combine the profit-focused and QoS-oriented policies for the single period problem; the factor $k_j(h)$ is a linear combination of two ratios, $h/t(h)$ and $1/t(h)$. The relative weight of this combination is determined by the discount factor d, and two parameters r_j and p_j that determine the demand growth. This controls the evolution of the operating policy from a QoS-focus at the beginning to a profit focus when the demand is steady.

In the following, we present some numerical demonstrations. Given the results described in Proposition 3, in each period, we numerically evaluate y_j and C_j^*, and make use of Corollary 5.

Corollary 5 *The following recursive relation holds,*

$$\frac{\partial \Pi_j}{\partial l_j} = y_j \int_{h \in H_j} w_j(h) g_j^*(h)^2 f_j(h) dh + dp_j \frac{\partial \Pi_{j+1}}{\partial l_{j+1}} .$$

The computation procedure is as follows. We start from the N-th period where the policy in this period is described by Proposition 2. The recursion in Corollary 5 is then applied in the previous period to obtain the optimal policy in the (N-1)-th period, using Proposition 3.

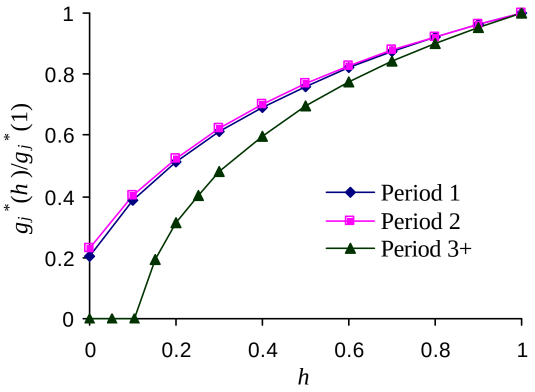

Figure 7. Optimal processing weights (normalized), $g_j^*(h)/g_j^*(1)$, plotted as a function of value, h, for discounting factor $d = 0.3$.

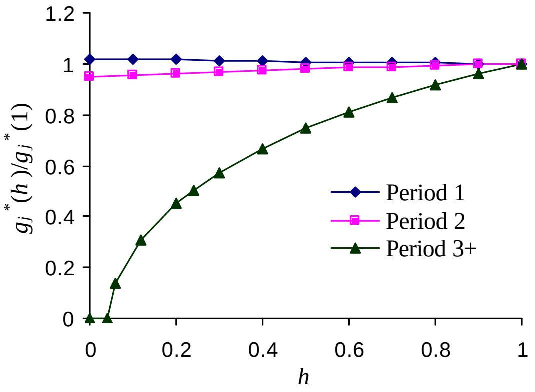

Figure 8. Optimal processing weights (normalized), $g_j^*(h)/g_j^*(1)$, plotted as a function of value, h, for discounting factor $d = 0.6$.

Figures 7 and 8 plot the optimal processing allocations for $N = 3$; $r_j = 2$, $p_j = 0.2$; $t_j^0(h) = 0.7 + 0.3h$, $w_j(h) = 1 + 0.5h$; $g_1 = 0.4$, $g_2 = 0.35$, $g_3 = 0.3$; and uniform value distribution for $0 \le h \le 1$. The processing allocation is purely profit-focused from the third period onwards. E-retailers with higher d are less discriminating as is evident from a lower value threshold h_c. A higher value of discount factor d indicates that the E-retailer is more concerned about the long-term. There is heavier capacity investment in the initial period to accommodate more customers. As d increases, the policies in the initial periods becoming less discriminating, or more QoS-oriented, due to the influence of future demand growth.

Figure 9 shows the optimal capacity increments in three periods. We assume the capacity remains unchanged from the third period onwards. E-retailers with low value of d invest in capacity in the initial period and maintain a more or less a constant capacity level, even with the decreasing cost of capacity. They focus on profit making even in the initial periods. This results in poor QoS, and therefore there is no growth in demand. E-retailers with higher d invest heavily in the initial period and adopt a QoS-oriented policy to build future demand. The second period sees a small increase or even a decrease in capacity. Partly, the demand growth is yet to take full effect so the slightly increased or even decreased capacity can sustain the same level of QoS. Also, the acquiring of capacity can be postponed to the third period that has a lower capacity cost. Recall that the effective unit capacity cost is $(1 - d)g_N$. It decreases with discount factor *d*, therefore we expect the optimal capacity increment in the third period to increase with *d*.

Figure 10 shows that the profits in the second, third, and fourth periods increase with discount factor. This is because of the increased customer base (or demand). In the second period, releasing of capacity gives a slightly faster increase of profit with higher *d*. The difference between the third period and periods from the fourth onwards is the cost of capacity. It is apparent that E-retailers with longer horizons (higher *d*) suffer more in the initial periods, from heavy investments in capacity that are aimed at improving QoS.

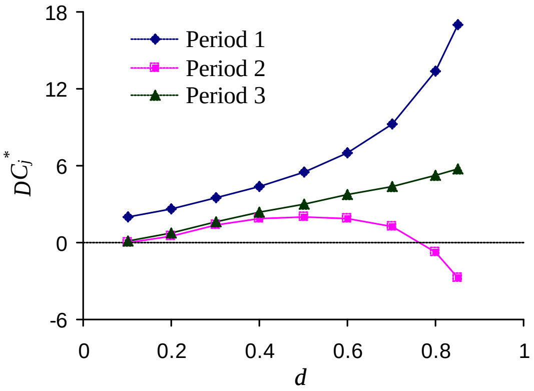

Figure 9. Optimal capacity increments, DC_j^*, plotted against discounting factor, *d*.

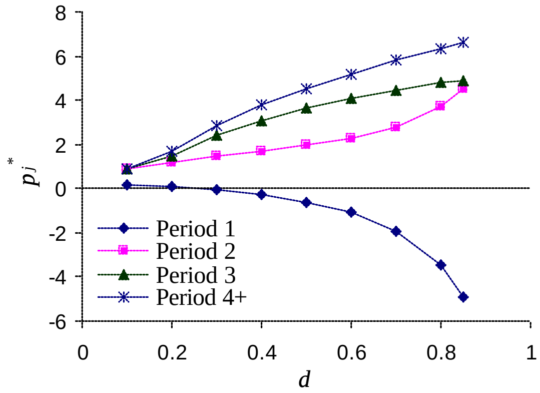

Figure 10. Profit in each period, p_j^*, plotted as a function of discounting factor, *d*.

4. Conclusions and Future Research

In this study, we propose an order processing scheme that improves the performance of an E-commerce site. This scheme is based on priority order processing, where the priority of an order depends on the potential value of the order. We present the model and the results for a single-period. It is shown that the benefit from priority processing increases as the server becomes busier.

We have also modeled a multi-period problem, where the demand in a period depends on the Quality of Service (QoS) that buyers receive in the previous period. It is observed that the retailer usually loses money in the first period in order to provide better service and growth in future demand. The operating policy of the server evolves from a QoS-focused one in initial periods, to a profit-focused one in subsequent periods. In multi-period problem, the server capacity planning in each period is determined optimally.

The current study involves one server (or treats multiple servers as a single server). It is of practical importance to study the load-balancing problem. This is a two-level problem. Each server will be optimized using priority scheme, conditional on an optimal amount of traffic directed to the server. Another interesting extension is to control or influence the value distribution, $f(h)$. This can be achieved using appropriate incentive schemes, such offering on the spot discount to buyers that may otherwise leave. Finally, we are working on the analysis of dynamic order processing policies that allow the priority of an order to vary during the course of shopping, especially in response to a more accurate prediction of the final value, and the delay tolerance.

References

Aiello, W. A., Mansour, Y., Rajagopolan, S., and Rosen, A., "Competitive Queue Policies for Differentiated Services," *Proceedings of INFOCOM*, 2000.

Ancker, C. J. and Gafarian, A. V., "Queuing with Impatient Customers Who leave at Random," *Journal of Industrial Engineering*, 13, pp. 84-90, 1962.

Ancker, C. J. and Gafarian, A. V., "Some Queuing Problems with Balking and Reneging – I," *Operations Research*, 11, pp. 88-100, 1963.

Ancker, C. J. and Gafarian, A. V., “Some Queuing Problems with Balking and Reneging – II,” *Operations Research*, 11, pp. 928-937, 1963.

Assaf, D. and Haviv, M., “Reneging from Processor Sharing Systems and Random Queues,” *Mathematics of Operations Research*, 15, 129-138, 1990.

Barrer, D. Y., “Queuing with Impatient Customers and Indifferent Clerks,” *Operations Research*, 5, 644-649, 1957.

Barrer, D. Y., “Queuing with Impatient Customers and Ordered Service,” *Operations Research*, 5, 650-656, 1957.

Coffman, E. G., Muntz, R. R., and Trotter, H., “Waiting Time Distributions for Processor-Sharing Systems,” *Journal of the Association for Computing Machinery*, 17(1), pp. 123-130, 1970.

Dreyfus, S. E. and Law, A. M., *The Art And Theory Of Dynamic Programming*, Academic Press, New York, 1977

Gnedenko, B. V. and Kovalenko, I. N., *Introduction to Queuing Theory*, Birkäuser, Boston, 1989.

Green, H., “The Great Yuletide Shakeout,” *Business Week*, pp. EB 18-28, November 1, 1999.

Kleinrock, L., “Analysis of a Time-Shared Processor,” *Naval Research Logistics Quarterly*, 11, 59-73, 1964.

Kleinrock, L., “Time-shared Systems: A Theoretical Treatment,” *Journal of the Association for Computing Machinery*, 14(2), pp. 242-261, 1967.

Kleinrock, L., *Queueing Systems: Volume 2, Computer Applications*, Wiley, New York, 1976.

Marchand, M. G., “Priority Pricing,” *Management Science*, 20(7), pp. 1131-1140, July 1974.

Mardesich, J., “The Web Is No Shopper’s Paradise,” Fortune, pp. 188-198, November 8, 1999.

Moe, W. W. and P. S. Fader, “Modeling Online Store Visit Patterns as a Measure of Customer Satisfaction,” Wharton School Working Paper, 2001.

Morrison, J. A., “Response-Time Distribution for a Processor-Sharing System,” SIAM Journal of

Applied Mathematics, 45, 152-167, 1985.

O’Donovan, T. M., “Direct Solutions of M/G/1 Processor-Sharing Models,” *Operations Research*, 22, pp. 1232-1235, 1974.

Rao, S. S., “Queuing with Balking and Reneging in M / G / 1 Systems,” *Metrika*, 12, pp. 173-188, 1968.

Sagan, H., *Introduction to the Calculus of Variations*, McGraw-Hill, New York, 1969.

Sakata, M., Noguchi, S., and Oizumi, J., “An analysis of the M/G/1 Queue under Round-Robin Scheduling,” *Operations Research*, 19, 371-385, 1971.

Stallings, W., *Operating Systems: Internals and Design Principles*, Prentice Hall, New Jersey, 1997.

Stanford, R. “Reneging Phenomena in Single Channel Queues,” *Mathematics of Operations Research*, 4, 162-178, 1979.

Tan, Y., “Value-based Design of Electronic Commerce Servers,” *Doctoral Dissertation*, University of Washington, 2000.

Turner, J. S., “New Directions in Communications,” *IEEE Communications Magazine*, 10, 1986.

Weinstock, R., *Calculus of Variations*, McGraw-Hill, New York, 1952.

Whitt, W., “Improving Service by Informing Customers about Anticipated Delays,” *Management Science*, 45(2), pp. 192-207, 1999.

Appendix

Proof of Proposition 1

Discrete Classes. Let us first solve the discrete version of the problem. Assume that there are N priority classes; class-k orders have arrival rate l_k, processing rate m_k and reneging rate n_k. Let E_k be the expected number of class-k orders in the system in the steady state.

We follow the method used by Kleinrock (1964). Introduce a tagged order (assume it is one of the priority class-k orders). Each order in class-k is given a time slice (quantum) of g_kQ, where

Q is infinitesimal. The cycle time is the duration between two consecutive ejects of the tagged order from the server, namely

$$\sum_{i=1}^{N} g_i Q E_i + g_k Q \,.$$

With a probability $m_k g_k Q$, an order of class-*k* will finish needed processing and exit from the system during the processing time $g_k Q$ allocated to this order. After a cycle, expected number of class-*k* orders will be

$$\left(1 - m_k g_k Q\right) E_k + l_k \left(\sum_{i=1}^{N} g_i Q E_i + g_k Q \right) - n_k \left(\sum_{i=1}^{N} g_i Q E_i + g_k Q \right) E_k \,.$$

The first term is the expected number of class-*k* orders returning to the system. The second term is the expected number of class-*k* orders arriving during a cycle. The third term is the expected number of class-*k* orders reneging during a cycle.

In the steady state, the above expression should be the same as E_k. Therefore we get

$$m_k g_k E_k = l_k \left(\sum_{i=1}^{N} g_i E_i + g_k \right) - n_k \left(\sum_{i=1}^{N} g_i E_i + g_k \right) E_k \,,$$

or

$$E_k = \frac{l_k (g_k + G)}{m_k g_k + n_k (g_k + G)} \,,$$

where the constant

$$G = \sum_{i=1}^{N} g_i E_i = \sum_{i=1}^{N} \frac{l_i (g_i + G) g_i}{m_i g_i + n_i (g_i + G)} \,. \tag{A1}$$

The above constant can be solved self-consistently.

The expected loss function for class-*k* orders is then

$$L_k = n_k E_k = \frac{l_k n_k (g_k + G)}{m_k g_k + n_k (g_k + G)} \,.$$

This is the expected number of class-*k* orders lost per unit time.

Existence of Unique Positive Value of *G*. Let's rewrite Equation A1 as

$$j\,(G) = \sum_{i=1}^{N} \frac{l_i\,(g_i + G)g_i}{m_i g_i + n_i\,(g_i + G)} - G = 0\,.$$

The function $j\,$(G) has *N* negative poles at $G = -(m_i / n_i + 1)g_i$, and (*N* – 1) negative zeros (roots) in between these poles. For the region where $G \geq 0$,

$$j\,(0) = \sum_{i=1}^{N} \frac{l_i g_i}{m_i + n_i} > 0\,,$$

and $j\,(\infty) \to -\infty$. Its second derivative,

$$\frac{d^2 j\,(G)}{dG^2} = -\sum_{i=1}^{N} \frac{2l_i m_i n_i g_i^{\,2}}{\left(m_i g_i + n_i\,(g_i + G)\right)^3} < 0\,.$$

Therefore $j\,(G)$ is a concave function over $G \geq 0$, with $j\,(0) > 0$ and $j\,(\infty) \to -\infty$. It must cross zero once and only once at a positive value of *G*.

Rescaling and Continuous Limits. The existence of a positive G allows us to rescale the weight vector, explicitly,

$$g_k / G \to g_k\,.$$

The rescaled g_k must satisfy Equation A1 (where G is set to 1). With the continuous value distribution *f*(*h*), required processing time $t(h)$ (inverse the service rate), and waiting time *w*(*h*), we can take continuous limit with respect to value, while retaining the discreteness of priority classes. For orders that have values in the interval [*h*, *h* + *dh*], if belonging to priority class-*k*, the loss function becomes,

$$l_k\,(h)dh = \frac{t\,(h)(g_k + 1)}{t\,(h)(g_k + 1) + w(h)g_k}\, l f\,(h)dh\,,$$

where $l_k(h)$ is loss function density. Proposition 1 follows if we further have $g_k \to g(h)$.

Proof of Proposition 2

We adopt the method of calculus of variations (Sagan, 1969; Weinstock, 1952) that is widely used to find the functional form that will optimize a given objective function.

Optimal Processing Allocation. We start by writing the Lagrange of Problem 1,

$$\mathsf{L} = \int_{h\in H} L(g(h),C)\,dh\,,$$

where, the integrand, or Lagrange density, is

$$\begin{aligned} L(g(h),h) = {} & \frac{w(h)g(h)h}{t(h)\big(g(h)+1\big)+w(h)g(h)}\, lf(h) \\ & + y\left(\frac{1}{l} - \frac{w(h)t(h)\big(g(h)+1\big)g(h)}{t(h)\big(g(h)+1\big)+w(h)g(h)}\right) lf(h) + x(h)g(h). \end{aligned}$$

y is the Lagrange multiplier for constraint given by Equation 4. $x(h)$ is the Lagrange multiplier density for constraints $g(h)\geq 0, \forall h\in H$. It is positive if $g(h) = 0$, and zero if $g(h) > 0$. The first variation yields,

$$\begin{aligned} \frac{\partial L(g(h),h)}{\partial g(h)} = {} & \frac{w(h)t(h)\, lf(h)}{\big(t(h)\big(g(h)+1\big)+w(h)g(h)\big)^2} \\ & \times \Big(h - y\,\Big(t(h)\big(g(h)+1\big)^2 + w(h)g(h)^2\Big)\Big) + x(h). \end{aligned} \qquad \text{(A2)}$$

Setting the above equation to zero, we get, if $g(h) > 0$,

$$h = y\,\Big(t(h)\big(g(h)+1\big)^2 + w(h)g(h)^2\Big); \qquad \text{(A3)}$$

and if $g(h) = 0$,

$$x(h) = lf(h)w(h)\left(y - \frac{h}{t(h)}\right).$$

Obviously, if

$$\frac{h}{t(h)} \geq y\,,$$

x(*h*) becomes zero, therefore the corresponding constraints are non-binding. The above results are summarized in Proposition 2. The value of Lagrange multiplier y can be solved by substituting the expression of $g^*(h)$ into constraint given by Equation 4.

Second Variation and Concavity. Let's perform the second variation around $g^*(h)$. This is done by differentiating Equation A2 with respect to $g(h)$ and substituting in Equation A3. We have, for $g^*(h) > 0$,

$$\left.\frac{\partial^2 L(g(h),h)}{\partial g(h)^2}\right|_{g(h)=g*(h)} = -\frac{2yw(h)t(h)lf(h)}{t(h)\left(g^*(h)+1\right)+w(h)g^*(h)} .$$

This is strictly negative as $y > 0$. Therefore, the solution $g^*(h)$ yields maximum value for the profit function p.

For the solution that lies on the boundary, i.e. $g^*(h) = 0$, the profit will decrease, at a rate $-x(h)$, if $g(h)$ is increased from $g^*(h) = 0$.

Optimal Capacity. The optimal capacity can be obtained by differentiating the Lagrange, explicitly,

$$\frac{\partial L}{\partial C} = \int_{h\in H} \frac{w(h)t^0(h)\left(g(h)+1\right)g(h)\left(h+yw(h)g(h)\right)}{\left(t^0(h)\left(g(h)+1\right)+w(h)Cg(h)\right)^2} lf(h)dh - g = 0 .$$

Substituting in Equation A3, the above expression can be simplified to,

$$\frac{y}{C^*}\int_{h\in H} \frac{w(h)t(h)\left(g^*(h)+1\right)^2 g^*(h)}{t(h)\left(g^*(h)+1\right)+w(h)g^*(h)} lf(h)dh = g \ . \tag{A4}$$

Solving Equations A3, A4, and constraint given by Equation 4, we can find the optimal policy described by $g^*(h)$, C^*, and y.

In the following we show the capacity solved above is optimal. First, taking derivative of Equation 4, and combined with Equation A3, we have,

$$\frac{\partial y}{\partial C^*} = \frac{y}{C^*}\frac{a_{12}}{a_{11}}.$$

The second derivative of Lagrange with respect to C is,

$$\frac{\partial^2 L}{\partial C^2} = \frac{y}{C^{*2}}\left(a_{22} - a_{12}\frac{C^*}{y}\frac{\partial y}{\partial C^*}\right) = \frac{y}{C^{*2}}\left(a_{22} - \frac{a_{12}{}^2}{a_{11}}\right). \qquad \text{(A5)}$$

The coefficients are,

$$a_{11} = \int_{h\in H} \frac{wt\left((g^*+1)^2 t + (g^*)^2 w\right)^2}{2\left((g^*+1)t + g^* w\right)^3}\, lfdh;$$

$$a_{12} = \int_{h\in H} \frac{wt\left((g^*+1)^4 t^2 - (g^*)^2 (g^*+1)^2 tw - 2(g^*)^3 (g^*+1) w^2\right)}{2\left((g^*+1)t + g^* w\right)^3}\, lfdh;$$

$$a_{22} = \int_{h\in H} \frac{wt\left((g^*+1)^4 t^2 - 4(g^*)^2 (g^*+1)^3 tw - 4(g^*)^3 (g^*+1)^2 w^2\right)}{2\left((g^*+1)t + g^* w\right)^3}\, lfdh.$$

We find that coefficients a_{11}, a_{12}, and a_{22} have the following properties:

1. $a_{11} > 0$;
2. $a_{11} > a_{12} > a_{22}$;
3. $a_{11} + a_{22} < 2a_{12}$.

Then, it is straightforward to show that Equation A5 is negative. If $a_{22} < 0$, it is obvious that $a_{11}a_{22} - a_{12}{}^2 < 0$; while if $a_{22} \ge 0$, we have $2\sqrt{a_{11}a_{22}} \le a_{11} + a_{22} < 2a_{12}$.

Proof of Corollary 1

Since $g^*(h)$ is nonnegative, from Equation A4, we have,

$$C^* \ge \frac{y}{g}\int_{h\in H} \frac{w(h)t(h)\left(g^*(h)+1\right)g^*(h)}{t(h)\left(g^*(h)+1\right) + w(h)g^*(h)}\, lf(h)dh = \frac{y}{g}.$$

Equation A3 yields an inequality, $h \ge y\,t(h)(g(h)+1)^2$. Substitute into Equation A4, we get

$$C^* \le \frac{1}{g} \int_{h \in H} \frac{hw(h)g^*(h)}{t(h)\left(g^*(h)+1\right)+w(h)g^*(h)}\, lf(h)dh = \frac{S^*}{g}.$$

where S^* is the expected revenue.

Proof of Corollary 2

Part i is straightforward, and,

$$\frac{\partial g^*(h)}{\partial\left(w(h)/t(h)\right)} = -\frac{g^*(h)^2}{2}\left(1+\left(\frac{1}{y}\frac{h}{t(h)}-1\right)\left(1+\frac{w(h)}{t(h)}\right)\right)^{-1/2} < 0.$$

Proof of Corollary 3

Part i. The derivative of the expected profit with respect to the demand is, according to envelope theorem,

$$\frac{\partial p^*}{\partial l} = \frac{\partial}{\partial l} p\left(g^*(h), C^*\right) = \int_{h \in H} \frac{w(h)g^*(h)\left(h - y\,t(h)\left(g^*(h)+1\right)\right)}{t(h)\left(g^*(h)+1\right)+w(h)g^*(h)}\, f(h)dh.$$

Substituting Equation A3, we have,

$$\frac{\partial p^*}{\partial l} = y \int_{h \in H} w(h)\left(g^*(h)\right)^2 f(h)dh > 0.$$

Parts ii and iii. Taking derivatives (with respect to l) of Equations 4 and A4, and making using of Equation A3, we get,

$$\frac{a_{11}}{y}\frac{\partial y}{\partial l} - \frac{a_{12}}{C^*}\frac{\partial C^*}{\partial l} = \frac{1}{l};$$

$$\frac{a_{12}}{y}\frac{\partial y}{\partial l} - \frac{a_{22}}{C^*}\frac{\partial C^*}{\partial l} = \frac{1}{l}\frac{gC^*}{y};$$

Solving these equations, we obtain,

$$\frac{\partial}{\partial l}\left(\frac{y}{C^*}\right) = \frac{1}{{a_{12}}^2 - a_{11}a_{22}}\left(\frac{a_{12}-a_{22}}{l}\frac{y}{C^*} - \frac{a_{11}-a_{12}}{l}g\right).$$

This becomes positive if

$$\frac{y}{C^*} > a_0 = g\,\frac{a_{11} - a_{12}}{a_{12} - a_{22}}\,.$$

Corollary 2 and properties of coefficients a_{11}, a_{12}, and a_{22} immediately render,

$$\frac{\partial C^*}{\partial l} = \frac{C^*}{{a_{12}}^2 - a_{11}a_{22}}\left(\frac{a_{11}}{l}\frac{gC^*}{y} - \frac{a_{12}}{l}\right) > 0\,.$$

Proof of Corollary 4

Part i. To show that the expected profit decreases with the increased capacity cost, we find that,

$$\frac{\partial p^*}{\partial g} = \frac{\partial}{\partial g}p\left(g^*(h), C^*\right) = -C^* < 0\,.$$

Parts ii and iii. Similarly we take derivatives (with respect to *g*) of Equations 4 and A4, this yields,

$$\frac{a_{11}}{y}\frac{\partial y}{\partial g} - \frac{a_{12}}{C^*}\frac{\partial C^*}{\partial g} = 0\,;$$

$$\frac{a_{12}}{y}\frac{\partial y}{\partial g} - \frac{a_{22}}{C^*}\frac{\partial C^*}{\partial g} = -\frac{C^*}{y}\,.$$

Solving these two equations simultaneously, we have,

$$\frac{\partial}{\partial g}\left(\frac{y}{C^*}\right) = \frac{a_{11} - a_{12}}{{a_{12}}^2 - a_{11}a_{22}} > 0\,;$$

$$\frac{\partial C^*}{\partial g} = -\frac{a_{11}C^*}{{a_{12}}^2 - a_{11}a_{22}}\frac{C^*}{y} < 0\,.$$

Proof of Lemma 1

This is obvious, as C_{j-1} appears, explicitly, in Π_j in a form g_jC_{j-1}.

Proof of Proposition 3

We follow Equation 7, and write the *j*-th period problem,

$$\max_{g_j(h),C_j} \; p_j\left(l_j, C_{j-1}; g_j(h), C_j\right) + d \cdot \Pi_{j+1}(l_{j+1}, C_j)$$

subject to:

$$1 = \int_{h \in H_j} \frac{w_j(h) t_j^0(h)\left(g_j(h)+1\right) g_j(h)}{t_j^0(h)\left(g_j(h)+1\right) + w_j(h) C_j g_j(h)} \, l_j f_j(h) dh;$$

$$l_{j+1} = \int_{h \in H_j} \frac{w_j(h) g_j(h) r_j + t_j(h)\left(g_j(h)+1\right) p_j}{t_j(h)\left(g_j(h)+1\right) + w_j(h) g_j(h)} \, l_j f_j(h) dh;$$

$$g_j(h) \geq 0, \quad \forall h.$$

Similar to single period problem, we can define a Lagrange,

$$\begin{aligned}
L_j = {} & p_j\left(l_j, C_{j-1}; g_j(h), C_j\right) + d \cdot \Pi_{j+1}(l_{j+1}, C_j) \\
& + y_j \left(1 - \int_{h \in H_j} \frac{w_j(h) t^0(h)\left(g_j(h)+1\right) g_j(h)}{t^0(h)\left(g_j(h)+1\right) + w_j(h) C_j g_j(h)} \, l_j f_j(h) dh\right) \\
& + c_j \left(\int_{h \in H_j} \frac{w_j(h) C_j g_j(h) r_j + t^0(h)\left(g_j(h)+1\right) p_j}{t^0(h)\left(g_j(h)+1\right) + w_j(h) C_j g_j(h)} \, l_j f_j(h) dh - l_{j+1}\right) \\
& + \int_{h \in H_j} x_j(h) g_j(h) dh,
\end{aligned}$$

where, y_j, c_j, and $x_j(h)$ are Lagrange multipliers (or density). The first variation gives,

$$h + c_j(r_j - p_j) = y_j\left(t_j(h)\left(g_j(h)+1\right)^2 + w_j(h) g_j(h)^2\right), \tag{A6}$$

where,

$$c_j = d \frac{\partial \Pi_{j+1}}{\partial l_{j+1}}.$$

The results described in Proposition 3 follow.

The optimal capacity for *j*-th period, C_j^*, is given by,

$$\frac{y_j}{C_j^*}\int_{h\in H_j}\frac{w_j(h)t^0(h)\left(g_j^*(h)+1\right)^2 g_j^*(h)}{t^0(h)\left(g_j^*(h)+1\right)+w_j(h)C_j^*g_j^*(h)}\,l_j f_j(h)dh = g_j - d\,\frac{\partial\Pi_{j+1}}{\partial C_j} = g_j - dg_{j+1}.$$

Similar to the single-period problem, the derivation of this result utilizes the solution $g_j^*(h)$, given by Equation A6, and Lemma 1. Note that, if one requires that the capacity is non-decreasing, (namely, $C_j \geq C_{j-1}$), a positive number, *e*$_j$, (the Lagrange multiplier) should be subtracted from the right hand side of the above expression.

Proof of Corollary 5

Using the definition of $\Pi_j(l_j, C_{j-1})$, given by Equation 7, and the envelope theorem, we have,

$$\frac{\partial\Pi_j}{\partial l_j} = \int_{h\in H_j}\frac{w_j(h)g_j^*(h)(h+c_j r_j)+t_j(h)\left(g_j^*(h)+1\right)\left(c_j p_j - y_j w_j(h)g_j^*(h)\right)}{t_j(h)\left(g_j^*(h)+1\right)+w_j(h)g_j^*(h)}\,f_j(h)dh.$$

Substituting in Equation A6 reduces to the recursive equation in Corollary 5.