

Quality-Speed Conundrum: Tradeoffs in Labor-Intensive Services

Krishnan S. Anand • M. Fazıl Paç • Senthil K. Veeraraghavan

Wharton School, University of Pennsylvania, Philadelphia, PA 19104.

anandk@wharton.upenn.edu • mpac@wharton.upenn.edu • senthilv@wharton.upenn.edu

September 2008

Abstract

In labor-intensive services such as primary health care, hospitality and education, the quality or value provided by the service increases with the time spent with the customer (with diminishing returns). However, longer service times (i.e., slower speed of service) also result in longer waits for customers. Thus, labor-intensive services need to make the tradeoff between service quality and service speed. By treating quality and speed as *independent* performance metrics, the extant academic research has not addressed the consequences of their interactions; whereas, their interaction is critical for labor-intensive services. In a queueing framework, we parameterize the degree of labor-intensity of the service. The service speed chosen by the service-provider affects the quality of the service through its labor-intensity. Customers queue for the service based on the quality of the service, delay costs and price. We study how a service provider can make the optimal “quality-speed tradeoff” in the face of such self-interested, rational customers. Our results demonstrate that the labor-intensity of the service is a critical driver of equilibrium price, service speed, demand, congestion in the queue and service provider revenues. We also model service rate competition among multiple servers, whose effects, we find, are very different from price competition. For instance, as the number of servers increases, the price increases and the servers become slower.

Keywords: Service Quality, Customer Behavior, Labor-Intensive Services, Queues, Cost Disease.

1. Introduction

‘Festina Lente’ [Make haste slowly]

– motto of *Aldus Manutius* (1449 - 1515).

In a wide variety of service industries, providing good customer service requires a high level of diligence and attention. Such “labor-intensive” services rely heavily on real-time human interactions between the service provider and the customer. Examples of such services are health care, legal and financial consulting, and personal care (such as spas, hair-dressing, beauty care and cosmetics).

Economists have noted that, in many industries (car manufacturing, computer assembly, retailing, etc.) productivity improvements have been rapid in the last few decades. Triplett and Bosworth (2004) found that from 1995 to 2001, labor productivity in services grew at an annual rate of 2.41%. Moreover, 17 of the 29 industries in their study realized productivity growth, including some that experienced very rapid growth (retail services, wholesale trade). However, as Varian (2004) points out, labor-intensive services have not realized the rapid productivity growth observed in the service sector. Industries in Triplett and Bosworth (2004)’s study, where productivity did *not* grow included miscellaneous repair services, hotels and lodging, amusement and recreation, education, and health services, all of which are labor-intensive services. For example, the health care industry displayed a negative growth, at -0.4% (Triplett and Bosworth, 2004. pp. 262-263).

A major difficulty in improving productivity in such labor-intensive services is the sensitivity of the *service quality* provided to the *speed of service*: As the service speed increases, the quality of the service inevitably declines. Often, the only way to increase productivity without sacrificing quality is to increase investments, which leads to increased prices. This phenomenon has been often termed as *Baumol’s cost disease* (Baumol 1993). James Surowiecki (2003) illustrates this point:

“Cost disease isn’t anyone’s fault. (That’s why it’s called a disease.) It’s just endemic to businesses that are labor-intensive. [...] you can control drug costs and limit expensive new procedures, but, when it comes to, say, hospital care and doctor visits, the only way to improve productivity is to shrink the size of the staff and have doctors spend less time with patients (or treat several patients at once). Thus the Hobson’s choice: to lower prices you have to lower quality.”

Primary health care practice in the United States epitomizes this problem. The scope of primary care includes health promotion, disease prevention, counseling, patient education, as well as diagnostic treatment of acute and chronic illnesses. If performed effectively, primary care practice holds the key to reducing the overall health care costs of patients by preventing

future illnesses. One of the often raised criticisms of the current primary health care system in the United States is that it does not offer patients effective primary care. Due to high levels of demand, doctors need to rush between patients, spending most of their time on treating acute illnesses (Yarnall *et al* 2003). In short, a persistent focus on the high pace of service leaves doctors with very little time for practicing valuable preventive care. According to an AARP report, a physician spends about 10.6 minutes on average with a patient, and sees about 112 patients a week on average (AARP Bulletin 2004). Consequently, the quality of the primary health care provided suffers.

These problems have led to the recent emergence of a new primary care model, termed “concierge medicine” (or, “boutique medicine”). For instance, MDVIP¹, founded in 2000, is a national network of 250 plus physicians who provide preventive and personalized health care. Concierge doctors affiliated to MDVIP care for a maximum of 600 patients. They offer their patients a highly customized primary care, spending as much as time as needed with each patient. In addition, the network provides these services with minimal delays. In return, concierge physicians charge higher fees, that also have the effect of limiting the demand for the service, thus reducing congestion.

In his testimony to the Joint Economic Committee of the U.S. Congress, Dr. Bernard Kaminetsky (formerly a Medical School faculty at New York University, and a concierge MD affiliated with MDVIP) testified that the traditional primary care practice is ineffective and dissatisfying for patients due to the fast pace of operations.² The primary health care example clearly demonstrates the quality degradation associated with a service system stretched to work at a fast pace while trying to serve a large number of patients. The longer duration of patient care in concierge practice leads to the service provider providing more valuable service to a limited number of customers.

The loss of quality due to a hurried pace of service provision, is not peculiar to the health care industry alone, but prevalent across a wide variety of labor-intensive industries (Baumol 1996). The time required to grade test papers, review files, and provide classroom education in colleges cannot be trimmed without loss of quality (*cf* Shaw and Black, 2001). In such services, a bias towards improving productivity and a relentless drive to cut operational costs may lead to lower revenues for the service provider due to quality erosion.

¹<http://www.mdvip.com/>. Another pioneering firm is *MD² International* at <http://www.md2.com>.

²“I was seeing 30 people a day and always rushing. Patients were dissatisfied.... I was dissatisfied.” Bernard Kaminetsky, M.D., F.A.C.P., in his testimony to the Joint Economic Committee of the United States Congress, April 28, 2004.

The aforementioned examples confirm that, focusing predominantly on improving productivity and reducing waiting times by increasing the speed of service leads to a reduction in the value of the service provided, which can eventually lead to lower demand. On the other hand, increasing the service value by increasing the time spent serving each customer has its pitfalls. First, it increases the cost of the service, as the productivity (number of customers served) falls. Second, it increases customers’ waiting times due to congestion effects from the slower service times. The first effect leads directly to higher prices; the second, to lower (net) customer value. In this paper, we study how a service provider can make the optimal “quality-speed” tradeoff in the face of strategic customers. We also analyze the equilibrium price, demand, congestion effects, customer valuations, and service provider profits, as a function of the labor-intensity of the services.

To summarize, customers in our model are strategic in that they join the service only if the net value (the value of the service net of congestion costs) exceeds the price charged by the service provider. Congestion costs are an outcome of the *aggregate* procurement decisions of all consumers in the market, since every customer who joins the service imposes a negative externality (in the form of additional waiting time) on all other customers. In turn, the tradeoff between the *quality* (service value) and *service speed* faced by the provider of a labor-intensive service forms the crux of our model. The extant academic research has not addressed the interaction between service value and service speed, or its consequences. In general, the extant literature treats service value and service times as independent performance metrics; whereas, their interaction is critical for labor-intensive services. In our queueing model, “labor-intensity” is indexed by a parameter α . The greater the labor-intensity of the service, the higher the value of α . We summarize some of the key results from our analytical model.

1. We find that the service provider increases the time of service as the labor-intensity of the service increases (i.e. as α increases). This result confirms the anecdotal and empirical observations of lower productivity/throughput in services that require larger investments in human capital and greater service interactions with customers.

As a result, the value of the service is *always* increasing in labor-intensity.

2. When the labor-intensity of the service increases, we observe one of two cases:
 - Case (i), Low to medium labor-intensity:* The price charge by the service provider falls, and yet the number of customers falls, and

Case (ii), High labor-intensity: The price increases and yet customer demand also increases.

To understand these effects, note that with an increase in labor-intensity, the service provider *slows down*. This increases the value of the service. At the same time, the slower service increases congestion in the system. Under low to medium intensities, the second effect predominates, reducing the net value accruing to customers. Hence the number of customers joining the service falls. This tends to reduce congestion, but not by enough to compensate for the slower service. In equilibrium, the price (which equals the net value accruing to customers) also falls.

Under high labor-intensity, slowing down the service significantly enhances the value to the customer from the service (by definition). This dominates any congestion effects. Hence more customers join the service, even at higher prices.

To summarize: Comparing two services of low to medium labor-intensity, the service that is more labor-intensive will be both less expensive and less congested than the other. But when we compare two services that are both highly labor-intensive, the service that is more labor-intensive is both more expensive and more congested than the other.

3. Heavily labor-intensive service providers benefit (in terms of both revenues and profits) by further increases in labor-intensity. When services are heavily labor-intensive, an increase in labor-intensity raises both price and demand (see point 2 above), and hence, the service provider's revenues.
4. We also develop a model of a price-setting service provider using multiple servers that compete with each other on service rates. We find that price charged by service provider *increases* as the number of servers increase. In other words, competition in service rates does *not* dampen prices. Further, the equilibrium waiting costs are invariant with respect to the number of servers. Our results illustrate the nature of pricing in the provision of labor-intensive services.

2. Related Literature

The extant academic research in Service Operations treats quality and speed as *independent* performance metrics. To our knowledge, there is no precedent in this literature that models

the labor-intensity of a service or studies the interactions between service quality and service speed, arising from labor-intensity. Nonetheless, two streams of extant research relate closely to key elements of our model. First, a number of papers model customers who choose whether or not to join a queue based on rational self-interest, as in our model. Second, a stream of research studies how service quality affects customer choices. These are reviewed below.

Rational Customers Joining Queues: Admission fees have long been considered an important tool to control congestion in service queues, dating back to the seminal paper by Naor (1969). Other papers that explore equilibrium queue joining, pricing and/or service rate decisions when customers have to wait for services include Armony and Haviv (2000), Cachon and Harker (2002), Gilbert and Weng (1998), Kalai *et al* (1992), Lederer and Li (1997), Li (1992), Li and Lee (1994), and Loch (1994). The reader is referred to Hassin and Haviv (2003)'s excellent review of this literature.

Productivity, Congestion and Quality: Papers in this genre include Chase and Tansik (1983), Gans (2002), Hopp *et al* (2007), Lu *et al* (2008), Oliva and Sterman (2001), Png and Reitman (1994), Ren and Wang (2008), Veeraraghavan and Debo (2008), and Wang *et al* (2008). Gans (2002) models a market where customers choose among servers based on the quality of the service. He finds that increasing the number of competing servers improves the industry's quality standards—a result similar to ours. Png and Reitman (1994), in their econometric analysis of gas stations, analyze customers' willingness to pay more for shorter lead times. They show that neither higher prices nor longer queues necessarily indicate higher quality. Veeraraghavan and Debo (2008) show that under *imperfect* information about service quality, customers might join longer queues even if the service is slower.

In Hopp *et al* (2007), the service-provider's revenues are increasing in the time devoted to 'discretionary' tasks. In their numerical studies, they show that, surprisingly, increasing skill levels may actually intensify congestion. Lu *et al* (2008) acknowledge that spending more time on a job may increase quality (and thus reduce rework). They study the effect of incentives on agents' output quality and the firm's profits, under different routing structures, in a production context. In both Hopp *et al* (2007) and Lu *et al* (2008), arrival rates are exogenously specified. Very often, service quality and realized customer demand interact in complex ways when customers are strategic. In a recent paper, Ren and Wang (2008), study the endogeneity of this critical relationship. In their large-scale empirical study covering over 4000 major hospitals in the US, they find that a larger patient volume does not necessarily lead to better service quality. Wang *et al* (2008) model the role of training for quality in the

diagnostic process in health care: With more training, nurses are more likely to make the correct diagnosis within the same contact time.

Two papers in organizational theory support our notion of the sensitivity of quality to service speed. In a simulation study, Oliva and Sterman (2001) show that when managers focus on improving productivity, service quality is eroded. In a normative model of service organizations, Chase and Tansik (1983) recommend that operational objectives for high-contact and low-contact services be different: High-contact service providers should focus on service effectiveness (i.e, quality), while low-contact service providers should focus on service efficiency (productivity).

3. A Model of Labor Intensive Service Provision

We consider a monopolist providing a labor-intensive service to a market of homogenous, self-interested consumers. We model the monopolist service setting using an unobservable $M/M/1$ queueing regime. Customers arrive to the market according to Poisson process at an exogenous mean rate Λ . We shall refer to Λ as the *potential demand* for the service. Upon arrival, every customer decides whether to procure the service (join the queue) or quit (balk from the services) based on the value of the service, expected cost of waiting and the price. As in many real contexts, the service procurement decision is made without observing the queue length. Patients do not always know exactly how many other patients are waiting in the doctor’s appointment queue, and clients often do not know how many other clients a legal-services team is planning to handle. We assume that all customers incur a waiting cost of c per unit of time spent in the system.

The service rate μ of the service provider is assumed to be common knowledge. The effective demand (effective arrival rate), λ , is the aggregate outcome of all customers’ procurement decisions (joining vs. balking). Then expected waiting cost for an arriving customer is given as follows:

$$WC(\mu, \lambda) = \begin{cases} \frac{c}{\mu-\lambda} & \text{if } 0 \leq \lambda \leq \mu \\ \infty & \text{if } \mu < \lambda \end{cases} \quad (1)$$

$\frac{c}{\mu-\lambda}$ in the above equation is the expected waiting cost for a customer arriving to an $M/M/1$ queue. If we were to use an $M/G/1$ queue, the mean waiting times can be immediately calculated by the Pollaczek-Khinchin formula (Ross 2006). However, to keep our model simple, we focus on an $M/M/1$ service system. An arriving customer rationally expects the

effective demand, considering all customers' equilibrium queue joining decisions. We will revisit the customers' queue joining decision in detail later in this section.

3.1 Service Value and Customer Equilibrium Decisions

Service Value: In many labor-intensive services, the value of the service provided decreases when the time spent with the customer is curtailed. We model the quality of the service through its associated service value V which increases with service time. Similarly in many situations, the marginal returns to increased service time is diminishing. We model the labor-intensive service by constructing the service value function V as a non-decreasing and concave function of service time τ . Thus, an additional minute of service increases the service value, but the incremental value of an additional minute decreases as the service duration is extended. The value of the labor-intensive service is given by:

$$V(\tau) = (V_b + \alpha/\tau_b - \alpha/\tau)^+ \quad (2)$$

where $x^+ = \max(x, 0)$. The parameter $\alpha \geq 0$ determines the sensitivity of the service value to service speed, and is an associated descriptor of the “nature” of the service. We denote α as *speed-sensitivity* or *labor-intensity* of the service provided. Clearly, higher values of α suggest a stronger dependence of the service value on the service speed (highly labor-intensive tasks). When α is high, reducing the service time results in the service value decreasing rapidly.

When α is zero, the value of the service provided equals V_b . Therefore, V_b could be thought of as a benchmark service value. Secondly, for any α when service time is τ_b , the value of the service provided is V_b . Therefore, τ_b could be considered as some benchmark service time, when the value of the service provided is V_b . For any labor-intensive service of type α , for the service to have positive value, service provider has to spend sufficient time with customer, i.e. $\tau \geq \underline{\tau}(\alpha) = \frac{\alpha\tau_b}{V_b\tau_b + \alpha}$.

The service value function defined above can be re-written in terms of the service rate, μ . Plugging in the service rate in to equation 2, we have:

$$V(\mu) = (V_b + \alpha\mu_b - \alpha\mu)^+ \quad (3)$$

μ_b denotes the service rate at which the service value is equal to the typical value, V_b . The service value, $V(\mu)$, is a decreasing function of the service rate μ . The fastest rate at which a non-negative valued service can be provided is given by $\bar{\mu} = V_b/\alpha + \mu_b$. The service

value does not change with changing service speed when the parameter α is equal to zero. Thus the ($\alpha = 0$) case is equivalent to the classical queueing model, where the value of the service is independent of the service speed.

Characterization of Service Rate Decision Space:

Clearly, the interaction between the service speed and the service value imposes a constraint on the service provider's operating region (i.e. the decision space of service rates he can choose from). Even if there were no queues, for a customer to expect non-negative net value (service value minus waiting cost during the service), the service value, $V(\mu)$, must exceed the expected cost of waiting. A service should be at the least valuable enough that a customer should not mind waiting *during* the process of service provision. Therefore, the value $V(\mu)$ must exceed the cost of waiting during the service, $\frac{c}{\mu}$, $V(\mu) - c/\mu \geq 0$. This assures that a customer can expect non-negative net value from the service when no other customer is joining the service. Note that a customer service procurement imposes negative externalities on others, as the expected waiting cost, $\frac{c}{\mu-\lambda}$, increases with the effective demand, λ .

Rewriting $V(\mu) - c/\mu \geq 0$, we have $V_b + \alpha\mu_b - \alpha\mu \geq c/\mu$, which in turn implies that $A_1(\alpha) \leq \mu \leq A_2(\alpha)$, where $A_1(\alpha), A_2(\alpha)$ are solutions to $V_b + \alpha\mu_b - \alpha\mu = c/\mu$. Thus,

- $\mu \geq A_1(\alpha) = \frac{V_b + \alpha\mu_b - \sqrt{(V_b + \alpha\mu_b)^2 - 4\alpha c}}{2\alpha}$. The service has to be fast enough. No one will wait forever *even* if the service value is high.
- $\mu \leq A_2(\alpha) = \frac{V_b + \alpha\mu_b + \sqrt{(V_b + \alpha\mu_b)^2 - 4\alpha c}}{2\alpha}$. The service cannot be *too fast*. It is not possible to provide valuable service at really high speeds of service.

For a labor-intensive service of type α , we denote this operating service-rate region by $\mathcal{F}(\alpha) = [A_1(\alpha), A_2(\alpha)]$. Note that the service provider's operating region depends on the characteristics of the service, α, V_b and μ_b , and the customers' cost of waiting, c . For the sake of notational convenience, we drop the dependencies on V_b, μ_b and c , and simply denote \mathcal{F} as dependent on α . The operating region and associated net value for service for any service rate in the operating region are denoted in Figure 1. Figure 1 shows that the service provider can choose from a larger range of service rates when the service is not very labor-intensive.

Remark 1 *As long as the typical service value V_b is greater than the expected waiting cost during a typical service, $\frac{c}{\mu_b}$, i.e. $V_b > \frac{c}{\mu_b}$, we have a non-empty operating region, $\mathcal{F}(\alpha)$ for a labor-intensive service of type α .*

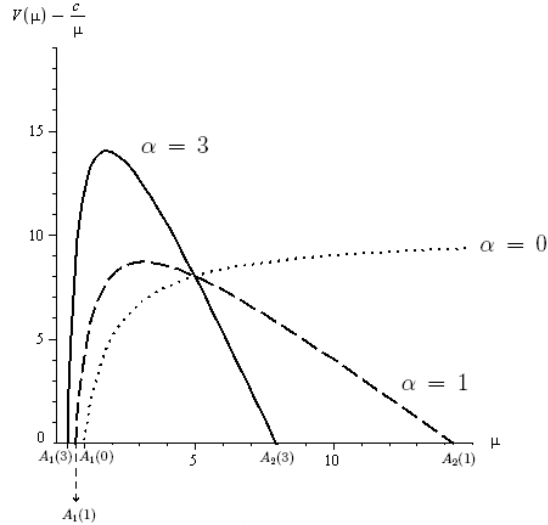


Figure 1: The net service value $(V(\mu) - c/\mu)$ and the operating region $\mathcal{F}(\alpha)$ shown for $\alpha = 0$ (dotted curve), $\alpha = 1$ (dashed curve) and $\alpha = 3$ (thick curve). Note that as $\alpha \rightarrow 0$, $A_1(\alpha) \rightarrow \frac{c}{V_b}$ and $A_2(\alpha) \rightarrow \infty$. However, for $\alpha > 0$ the service rates that provide non-negative net value are bounded in the interval $[A_1(\alpha), A_2(\alpha)]$.

Remark 1 establishes the non-empty operating region $\mathcal{F}(\alpha)$ in which the service provider can earn some positive revenue. Note that as the typical service value V_b exceeds the corresponding expected cost of waiting during the service c/μ_b (which is a natural assumption), we have an interval to choose the service rates from.

Customers' Queue Joining Decision: Self-interested customers arrive to the system according to a Poisson process at an exogenously determined rate Λ , and decide whether to join the unobservable service queue. As assumed, all customers are homogenous in their valuations of the service. Potential demand, Λ , price, p , service rate μ , and resulting service value, $V(\mu)$, are common knowledge to all arriving customers. We model the queue joining decision of individual customers as in Hassin and Haviv (2003). Let $\gamma_i(\mu, p)$ denote the probability that a customer i would join a queue where the server has service rate μ and admission price p . We consider symmetric equilibrium queue-joining strategies since all customers are homogenous. Simply, $\gamma_i(\mu, p) = \gamma(\mu, p) \forall i$. The queue joining decision of customers $\gamma(\mu, p)$ is based on the value of the service, the price and the expected cost of waiting.

There are two possible equilibrium outcomes under symmetric pure customer strategies: For a given service rate, μ and price, p , either all potential customers join the queue ($\gamma(\mu, p) = 1$), or none of them join the queue ($\gamma(\mu, p) = 0$). In the former case, even if all potential

customers join the net benefit is non-negative ($V(\mu) - (p + WC(\mu, \Lambda)) \geq 0$), therefore all customers choose to join the queue. Hence the equilibrium demand, $\lambda_e(\mu, p)$, is equal to the potential demand, Λ . In the latter case, the net benefit is negative for a customer joining the queue even if no other customers join the queue ($V(\mu) - (p + c/\mu) < 0$), and therefore the equilibrium demand is equal to zero.

Note that for $p + c/\mu < V(\mu) < p + WC(\mu, \Lambda)$, there exists no symmetric pure strategy queue-joining equilibrium. If no other customer joins the queue, an arriving customer has an incentive to join the queue. But if all other customers join the queue, an arriving customer is better off not joining the queue. In this case, the service provider can still earn positive revenues serving a fraction of the potential demand. We consider symmetric mixed customer strategies in which arriving customers randomize their queue joining decisions. $\gamma(\mu, p) \in (0, 1)$ denotes an arriving customer's probability of joining the service queue. We explore symmetric mixed strategy equilibria in which all arriving customers join the queue with probability, $\gamma_e(\mu, p)$, such that they are indifferent between joining the queue and not joining the queue when all customers choose $\gamma_e(\mu, p)$. The resulting equilibrium arrival rate is given by $\lambda_e(\mu, p) = \gamma_e(\mu, p)\Lambda$. At the equilibrium arrival rate (synonymous to equilibrium demand throughout the paper), the expected cost of the service to a customer is equal to the value of the service. The following equation then determines the equilibrium arrival rate, $\lambda_e(\mu, p)$:

$$V(\mu) - p = WC(\mu, \lambda_e(\mu, p)).$$

Having characterized the labor-intensive service value, and the equilibrium decisions of customers, we are now ready to focus on the service providers' revenue maximization objective.

3.2 Service Provider's Revenue Maximization

Service provider's objective is to maximize the revenues with respect to the service rate, μ and the price, p . Therefore the objective function of the service provider is given by:

$$\max_{\{p \geq 0, \mu \in \mathcal{F}\}} \{R(\mu, p) = p\lambda_e(\mu, p)\} \quad (4)$$

We solve the service provider's revenue maximization problem stepwise; First, we find the optimal price for a given service rate, μ . Then using the optimal price for every service rate μ , we find the revenue maximizing service rate in the operating region $\mathcal{F}(\alpha)$.

3.2.1 Service Provider's Price Decision

For a fixed service rate, μ in the operating region, $\mathcal{F}(\alpha)$, the service provider maximizes revenues with respect to price, p . We interpret μ as service rate rather than as service capacity. Setting a service rate in the operating region, $\mathcal{F}(\alpha)$, assures that $V(\mu) - c/\mu > 0$. Therefore there exists a positive price at which the service provider can earn positive revenues, for all $\mu \in \mathcal{F}\alpha$. The equilibrium demand, $\lambda_e(\mu, p)$, as a function of the price is given as follows:

$$\lambda_e(\mu, p) = \begin{cases} \Lambda & \text{if } 0 \leq p \leq V(\mu) - WC(\mu, \Lambda) \\ \mu - \frac{c}{V(\mu)-p} & \text{if } V(\mu) - WC(\mu, \Lambda) < p \leq V(\mu) - WC(\mu, 0) \\ 0 & \text{if } V(\mu) - WC(\mu, 0) < p. \end{cases} \quad (5)$$

Clearly, the equilibrium demand, $\lambda_e(\mu, p)$, is a non-increasing function of the price. If the potential demand for the service is low, such that the expected waiting cost when all customers join the queue is lower than the value of the service, $WC(\mu, \Lambda) < V(\mu)$, the service provider can serve all potential customers, hence clear the market, at low prices ($p \leq V(\mu) - WC(\mu, \Lambda)$). The resulting revenue for a given service rate, μ , as a function of the price is given by:

$$R(\mu, p) = \begin{cases} p\Lambda & \text{if } 0 \leq p \leq V(\mu) - WC(\mu, \Lambda) \\ p \left(\mu - \frac{c}{V(\mu)-p} \right) & \text{if } V(\mu) - WC(\mu, \Lambda) < p \leq V(\mu) - WC(\mu, 0) \\ 0 & \text{if } V(\mu) - WC(\mu, 0) < p. \end{cases} \quad (6)$$

When the potential demand is low, the service provider can clear the market. Therefore, the demand is not affected by increasing price, hence the revenues are increasing if $p \leq V(\mu) - WC(\mu, \Lambda)$. The monopolist service provider will not leave the consumers any surplus, and charge at least $p = V(\mu) - WC(\mu, \Lambda)$. This occurs because consumers join the service as long as the net benefit is non-negative. The service provider may increase the price further at the expense of losing some of the demand. This may lead to higher revenues if the demand is inelastic (i.e., price elasticity of demand is lower than one). On the other hand, when the potential demand Λ is large the service provider cannot clear the market even at a zero price. The service provider charges a price that maximizes its revenues by serving a fraction of the customers. The following proposition presents the service provider's optimal pricing policy for a given service rate μ .

Proposition 1 *Consider a labor-intensive service of type α . The equilibrium outcome depends on the threshold $\hat{\lambda}(\mu) = \mu - \sqrt{\frac{c\mu}{V(\mu)}}$, which defines the maximum number of customers*

the service provider will serve at a given service rate $\mu \in \mathcal{F}(\alpha)$, irrespective of the market demand. For any service rate $\mu \in \mathcal{F}(\alpha)$, the optimal price equals:

$$p^*(\mu) = \begin{cases} V(\mu) - WC(\mu, \Lambda) & \text{if } 0 \leq \Lambda \leq \hat{\lambda}(\mu) \\ V(\mu) - \sqrt{cV(\mu)/\mu} & \text{if } \hat{\lambda}(\mu) < \Lambda. \end{cases} \quad (7)$$

The resulting equilibrium arrival rate is equal to:

$$\lambda_e(\mu, p^*(\mu)) = \begin{cases} \Lambda & \text{if } 0 \leq \Lambda \leq \hat{\lambda}(\mu) \\ \hat{\lambda}(\mu) & \text{if } \hat{\lambda}(\mu) < \Lambda. \end{cases} \quad (8)$$

and the corresponding equilibrium revenue is equal to:

$$R(\mu, p^*(\mu)) = \begin{cases} (V(\mu) - WC(\mu, \Lambda))\Lambda & \text{if } 0 \leq \Lambda \leq \hat{\lambda}(\mu) \\ \mu V(\mu) - 2\sqrt{c\mu V(\mu)} + c & \text{if } \hat{\lambda}(\mu) < \Lambda. \end{cases} \quad (9)$$

Proposition 1 shows the optimal pricing decision and equilibrium demand (arrival rate) for any arbitrarily chosen service rate μ . We note that the threshold $\hat{\lambda}(\mu)$ defines the maximum number of customers the service provider would serve at a given service speed μ . When $\Lambda < \hat{\lambda}(\mu)$ the service provider clears the market (i.e., all arriving customers join the service provided by the server), and it also extracts all consumer surplus by pricing suitably. However, when the potential demand is higher (when $\Lambda \geq \hat{\lambda}(\mu)$) the service provider chooses not to serve more than $\hat{\lambda}(\mu)$ customers. Serving more customers will increase the congestion in the system, which means higher waiting costs for customers. Thus, to accommodate more customers, the service provider has to compensate the customers for the additional waiting costs they incurred. At the service rate μ , this compensatory measure can only be achieved by decreasing the price. As the arrivals to the system increase further, serving every additional customer requires a larger reduction in price, which eventually leads to a scenario when the increase in demand does not make up for the revenue lost due to the corresponding price reduction. When the service provider provides a service of value $V(\mu)$, he limits the number of customers admitted to the system by charging a suitable admission price which is independent of the potential demand in the market. Therefore, fluctuations in potential demand would not affect the optimal price, and hence revenues, as long as the potential demand remains higher than the threshold $\hat{\lambda}(\mu)$.

Proposition 1 shows that for every service speed within the operating region $\mu \in \mathcal{F}(\alpha)$, there exists a finite price $p^*(\mu)$ that maximizes the service provider's revenue. However, the optimal pricing strategies differ when the arrival demand rate is high (when the service provider does not clear the market), and when the arrivals are low (when it is optimal to

clear the market). Therefore, in order to consider the optimal service rate decision we have to consider each of the aforementioned cases separately. For expositional convenience, we begin by considering the large market scenario, when the service provider cannot cover all the demand in the market in Section §3.3 (when the potential demand $\Lambda > \lambda_\alpha^*$ where λ_α^* is a threshold dependent on α). Thereafter, we consider the small market scenario (when $\Lambda \leq \lambda_\alpha^*$) in Section §3.4. We then conclude this section by consolidating the results in Section §3.5.

3.3 Large Market Scenario

In this section, we examine the large market scenario. Suppose that the service provider chooses to provide service at rate μ for labor-intensive service of type $\alpha > 0$. Lemma 1 characterizes the behavior of optimal prices and the effective equilibrium demand with respect to changes in the chosen service rate μ .

Lemma 1 *For any $\alpha > 0$, the optimal price for a given service rate, $p^*(\mu)$ and the resulting equilibrium demand $\lambda_e(\mu, p^*(\mu))$ are unimodal in the service rate, μ when $\Lambda > \bar{\lambda}_\alpha > \lambda_\alpha^*$.*

Lemma 1 characterizes the behavior of the optimal price and the resulting equilibrium demand for every service rate within the possible operating region. Since the optimal price function is unimodal in μ , we note that the service provider offers the highest price not when the service value is highest, but at some “intermediate” service value (at some ‘interior’ service rate in $\mathcal{F}(\alpha)$). This occurs because the value of the service *net* of waiting costs has a finite upper bound. Thus, high prices in such labor-intensive services may *not* be indicative of high quality. When the service provider spends a long time on every customer, other customers in queue have to wait longer even when we account for customers who balk the service because of long waiting times. Thus, in equilibrium, it may be that the customers are willing to pay more for a (relatively) faster service of lower value than a more valuable service that takes more time.

Lemma 1 also characterizes the equilibrium demand (rate) for any service rate decision. Even in markets where the potential demand rate is infinite (i.e., $\Lambda \rightarrow \infty$), increasing the service speed does not lead to increase in effective demand. At high service speeds, some customers drop out from queueing for the service. Note that both service value and waiting times are low. However, the value of the service diminishes faster than the cost of waiting when the service rate is increased, and therefore, some customers balk from the service.

3.3.1 Service Provider's Optimal Service Rate Decision:

The service provider's objective is to maximize the revenues with respect to the price and the service rate. Having derived the optimal price for each service rate μ , we now focus on the service provider's service rate decision that maximizes its revenues. Therefore, the revenue maximization objective can be written as:

$$\max_{\{\mu \in \mathcal{F}(\alpha)\}} \left\{ \max_{\{0 \leq p \leq V(\mu)\}} \{R(\mu, p)\} \right\} \quad (10)$$

Employing the optimal pricing results from Proposition 1, we can reduce the optimal revenue maximization involving prices and service rates, to a problem requiring just the service rate decision. Lemma 2 provides the result of maximization in equation (10).

Lemma 2 *Consider any labor-intensive service of type $\alpha > 0$. Let the potential demand be high, i.e., $\Lambda > \lambda_\alpha^*$. Then,*

1. *The optimal service rate is equal to $\mu^* = \frac{V_b + \alpha \mu_b}{2\alpha}$.*
2. *The corresponding optimal price is equal to $p^*(\mu^*) = \frac{V_b + \alpha \mu_b - 2\sqrt{c\alpha}}{2}$.*
3. *The equilibrium demand (arrival rate) at the optimal price and service rate equals $\lambda_e(\mu^*, p^*(\mu^*)) = \frac{V_b + \alpha \mu_b - 2\sqrt{c\alpha}}{2\alpha}$.*

Therefore, the optimal revenue for the service is equal to $R(\mu^, p^*(\mu^*)) = \frac{(V_b + \alpha \mu_b - 2\sqrt{c\alpha})^2}{4\alpha}$.*

Lemma 2 shows that there exists a unique, interior service rate μ^* in $\mathcal{F}(\alpha)$ that maximizes the revenues. Furthermore, note that the optimal service setting is independent of the potential demand, Λ . Throughout the paper, for notational convenience, unless it is necessary, we suppress the dependency of the decision variables (μ and p) on α . For instance, μ^* is the optimal service rate for a labor-intensive service of type α .

At this juncture, we point out how the labor-intensive nature of the offered service affects the optimal decision parameters. Recall that α denotes how fast the quality of the service degrades with increasing service speed. Clearly, when α equals zero, we have a traditional non-labor-intensive service. Note that the maximum number of customers that a server serves decreases, as the service becomes more labor-intensive (as α increases).

Lemma 2(1) shows that the optimal service rate, μ^* , is decreasing in α . As the service becomes more labor-intensive, in equilibrium, the service provider has a higher incentive to

spend more time on each customer. Also note that the optimal service value is lower (higher) than the benchmark service value V_b for $\alpha < (>)V_b/\mu_b$. Faster service is chosen over valuable service if the labor-intensity is relatively small ($\alpha < V_b/\mu_b$), and valuable service is chosen over faster service otherwise.

From Lemma 2(2) we note that the optimal price, $p^*(\mu^*)$, is unimodal in α , decreasing for $\alpha < c/\mu_b^2$, and increasing for $\alpha > c/\mu_b^2$. Although the optimal service time is increasing as the service becomes more labor-intensive, it does not imply an increase in the *net* value of service provided. If α is low (i.e. $\alpha < c/\mu_b^2$), the congestion effects dominate the increase in service value. Therefore the optimal price for the service is *decreasing* in α . As the task become more labor-intensive (i.e. α increases), prices fall! However, when α is high ($> c/\mu_b^2$), the gains in the service value due to increased service time dominate the increase in the equilibrium waiting cost, therefore the optimal price is increasing in high α .

Let us consider the behavior of the equilibrium demand, $\lambda_e(\mu^*, p^*(\mu^*))$ from Lemma 2(3). The equilibrium demand is decreasing in α for $\alpha < V_b^2/c$, and is increasing in α for $\alpha > V_b^2/c$. As the labor-intensity of the service increases, the minimum service time required to produce positive valued service increases. This implies that the maximum number of customers that can be served by a server decreases. Thus, the equilibrium demand falls with increasing α as the service provider serves slower. But once α is greater than a certain threshold (V_b^2/c), an increase in α leads to a higher equilibrium demand and consequently to a higher price.

Finally, results of Lemma 2 also describe the effect of waiting costs. The optimal service rate, μ^* , is independent of the waiting cost, c . Interestingly, if customers are more impatient in a labor-intensive service, the additional waiting cost does not result in a faster service. As one might expect, higher waiting costs leads to lowering of prices, $p^*(\mu^*)$, but they also lead to lower equilibrium demand $\lambda_e(\mu^*, p^*(\mu^*))$. Consequently the optimal revenues, $R(\mu^*, p^*(\mu^*))$, decrease with increasing waiting cost.

Having characterized the optimal decision parameters, we now explore the relative impact of choosing non-optimal service rates when providing a labor-intensive service. Again, consider a labor-intensive service of type α . Let the optimal service rate be $\mu^* \in \mathcal{F}(\alpha)$. Let the corresponding optimal price for a given service rate $\mu \in \mathcal{F}(\alpha)$ be $p^*(\mu)$ and the resulting equilibrium demand at that price and service rate be $\lambda_e(\mu, p^*(\mu))$. Then we have the following property.

Lemma 3 [Property of α -symmetry:] For a labor-intensive service of type α , $p^*(\mu)$

and $\lambda_e(\mu, p^*(\mu))$ have the following symmetric relationship around the optimal service rate μ^* for any given $\mu \in \mathcal{F}(\alpha)$.

$$p^*(\mu^* + \epsilon) = \alpha \lambda_e(\mu^* - \epsilon, p^*(\mu^* - \epsilon)),$$

where $\epsilon = (\mu - \mu^*)$.

Lemma 3 clearly demonstrates that prices and effective demand are two levers related to each other by the labor-intensity parameter α . To better illustrate the property we derived in Lemma 3, we divide the operating region $\mathcal{F}(\alpha)$ into 3 sub-regions as shown in Figure 2. Region 1 corresponds to low service rates, Region 2 corresponds to intermediate service rates, and Region 3 corresponds to high service rates.

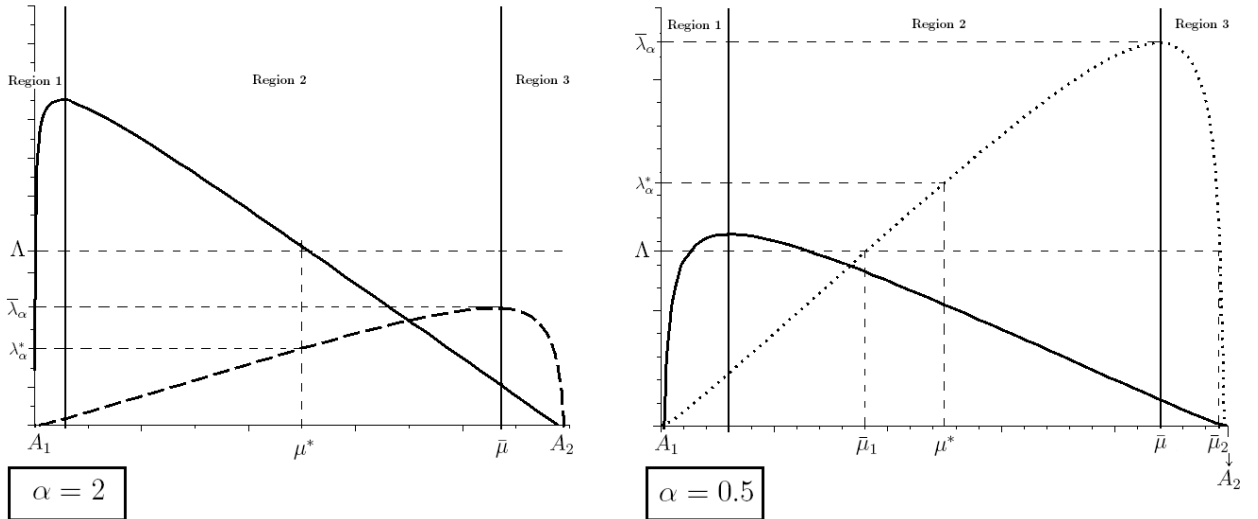


Figure 2: The symmetry of $p^*(\mu)$ (denoted by thick line) and $\lambda_e(\mu, p^*(\mu))$ (denoted by dotted curve) around μ^* for a labor-intensive services of type $\alpha = 2$ (left panel) and $\alpha = 0.5$ (right panel). The optimal service rate, in a market with infinite demand is μ^* and the corresponding equilibrium arrival rate is $\lambda_\alpha^* = \lambda_e(\mu^*, p^*(\mu^*))$. The throughput maximizing service rate is $\bar{\mu}$ and the corresponding throughput is denoted by $\bar{\lambda}_\alpha$. A representative potential demand Λ is indicated in the figures. On the left panel, the potential demand Λ is higher than $\bar{\lambda}_\alpha$, and the right panel, $\Lambda < \bar{\lambda}_\alpha$. For illustrative purposes, $V_b = 10$, $\mu_b = 5$ in the figure.

When the service rates are low (Region 1), there is an over-investment in time of service for both the customers and the service provider. Although the service provided is of high value, the cost of waiting is also high. Can the service provider then increase the service speed? Increasing the service rate will lead to some loss of service value. However, the gains

from the reduction in waiting cost will overcome the service value loss, (at low μ , waiting costs drop precipitously as μ increases), hence the net value of the service to a customer will be higher. The increase in the net value of the service allows the service provider to charge customers a higher price. Furthermore, the service rate increase also leads to higher throughput. Therefore, the service provider has an opportunity to both increase the price and the number of customers served simultaneously.

For intermediate service rates (Region 2), increasing the service rate no longer increases the net value of the service because the reduction in the service value is higher than the reduction in waiting cost. Therefore at a given price, increasing the service rate leads to lower equilibrium demand (and consequently, lower revenues). However, by increasing the service rate, the service provider gains the capability of serving more customers. By lowering the price, service provider can increase the demand. Figure 2 indicates that the firm chooses to serve more customers at a lower price as the optimal price at μ^* is lower than the maximum price that can be charged at the end point of Region 1.

When the service rates are high (Region 3), increasing the service rate is not desirable, as it leads to a lower price *and* lower equilibrium demand. In this region, the reduction in service value due to increasing service speed, is far greater than the gains made from customer waiting cost reduction. Thus the equilibrium demand is decreasing despite the reduction in price. This implies that, when the potential demand is high the optimal service rate lies in the intermediate service rate region (Region 2).

Figure 2 also illustrates the co-existence of two different service systems in a market that can earn identical revenues. For example an exogenous constraint (regulatory control on prices, industry standards on contact time, etc.) can force a monopolist to choose one of the following two options. A service provider may choose to provide high quality service and limit the demand charging a high price, or it may choose to provide service to a large number of customers at a lower price but with lower value. Comparable revenues can be attained through either of the two service decisions. Modeling labor-intensity through α allows us to capture the presence of such options in service provision.

Finally, we show that even when the demand is high, for labor-intensive services, concentrating solely on a productivity (or throughput) based objective leads to lower revenues. The following Lemma illustrates the point.

Lemma 4 *For a labor-intensive service of type $\alpha > 0$, the revenue maximizing equilib-*

rium demand $\lambda_e(\mu^*, p^*(\mu))$ is strictly lower than the maximum equilibrium demand, $\bar{\lambda}_\alpha = \max_{\{\mu \in \mathcal{F}\}} \{\lambda_e(\mu, p^*(\mu))\}$. The service rate maximizing the equilibrium throughput is greater than the optimal service rate, i.e., $\bar{\mu} = \operatorname{argmax}_{\{\mu \in \mathcal{F}\}} \{\lambda_e(\mu, p^*(\mu))\} > \mu^*$.

The crux of Lemma 4 is that an objective of maximizing throughput would lower the quality of service in the market (since the optimal service rate would be higher). Furthermore, even when there is ample demand, focussing on maximizing throughput is also suboptimal for the service provider. The symmetric relation of Lemma 3 implies that charging the highest price, $\max_{\{\mu \in \mathcal{F}\}} p^*(\mu)$, is also suboptimal. Thus, in labor-intensive services, the service provider maximizes revenues by serving less than $\bar{\lambda}_\alpha$ customers but charging them a higher price at μ^* than he could charge by serving more customers faster. This result is visualized in Figure 2 for a particular α . Note that μ^* (Revenue maximizing service rate) is lower than $\bar{\mu}$ (throughput maximizing service rate).

3.4 Small Market Scenario

Having derived the characteristics of the system under the large market scenario, we now analyze the case of a small market, where the potential demand Λ is low. We begin with the following Lemma that characterizes the optimal price that clears the market demand.

Lemma 5 *For any labor-intensive service of type α , when the potential demand is such that $\Lambda < \bar{\lambda}_\alpha$, there exist $\bar{\mu}_1(\Lambda)$ and $\bar{\mu}_2(\Lambda)$ in $\mathcal{F}(\alpha)$ such that the optimal price $p^*(\mu)$ clears the market, i.e. $\lambda_e(\mu, p^*(\mu)) \geq \Lambda$ for $\mu \in [\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)]$.*

Lemma 5 shows that for low values of potential demand there exists a closed interval of service rates $[\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)] \subset \mathcal{F}(\alpha)$, where it is optimal to clear the market. A corresponding price $p^*(\mu)$ can be chosen so that the market demand is cleared for any service rate μ in this interval. Note that $\Lambda = \lambda_e(\bar{\mu}_1(\Lambda), p^*(\bar{\mu}_1(\Lambda))) = \lambda_e(\bar{\mu}_2(\Lambda), p^*(\bar{\mu}_2(\Lambda)))$.

If the potential demand, Λ , is higher than $\bar{\lambda}_\alpha$ then the service provider cannot clear the market at any service rate μ in the operating region $\mathcal{F}(\alpha)$. Using the result of Lemma 5 we write the resulting equilibrium arrival rate as:

$$\lambda_e(\mu, p^*(\mu)) = \begin{cases} \hat{\lambda}(\mu) = \mu - \sqrt{\frac{c\mu}{V(\mu)}} & \text{if } A_1(\alpha) \leq \mu < \bar{\mu}_1(\Lambda) \\ \Lambda & \text{if } \bar{\mu}_1(\Lambda) \leq \mu \leq \bar{\mu}_2(\Lambda) \\ \hat{\lambda}(\mu) = \mu - \sqrt{\frac{c\mu}{V(\mu)}} & \text{if } \bar{\mu}_2(\Lambda) \leq \mu \leq A_2(\alpha), \end{cases} \quad (11)$$

A representative example of the equilibrium demand in a small market scenario can be seen in the right panel of Figure 2. Note that service provider covers the entire demand Λ in the interval $[\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)]$. In the same vein, we can rewrite the optimal price, $p^*(\mu)$, for a given service rate as follows:

$$p^*(\mu) = \begin{cases} V(\mu) - \sqrt{cV(\mu)/\mu} & \text{if } A_1(\alpha) \leq \mu < \bar{\mu}_1(\Lambda) \\ V(\mu) - WC(\mu, \Lambda) & \text{if } \bar{\mu}_1(\Lambda) \leq \mu \leq \bar{\mu}_2(\Lambda) \\ V(\mu) - \sqrt{cV(\mu)/\mu} & \text{if } \bar{\mu}_2(\Lambda) \leq \mu \leq A_2(\alpha). \end{cases} \quad (12)$$

Therefore, the corresponding equilibrium revenue is equals:

$$R(\mu, p^*(\mu)) = \begin{cases} \mu V(\mu) - 2\sqrt{c\mu V(\mu)} + c & \text{if } A_1(\alpha) \leq \mu < \bar{\mu}_1(\Lambda) \\ (V(\mu) - \frac{c}{\mu-\Lambda})\Lambda & \text{if } \bar{\mu}_1(\Lambda) \leq \mu \leq \bar{\mu}_2(\Lambda) \\ \mu V(\mu) - 2\sqrt{c\mu V(\mu)} + c & \text{if } \bar{\mu}_2(\Lambda) \leq \mu \leq A_2(\alpha). \end{cases} \quad (13)$$

We derive the service provider's optimal settings (optimal service rate and optimal price) using the equilibrium revenue function given by equation (13).

Lemma 6 *Under low potential demand $\Lambda < \lambda_\alpha^*$ for any labor-intensive service of type α :*

1. *The optimal service rate is equal to $\mu^* = \Lambda + \sqrt{c/\alpha}$*
2. *The corresponding optimal price is equal to $p^*(\mu^*) = V_b + \alpha\mu_b - \alpha\Lambda - 2\sqrt{\alpha c}$.*
3. *The equilibrium demand (arrival rate) at the optimal price and service rate $\lambda_e(\mu^*, p^*(\mu^*)) = \Lambda$ (full market coverage).*

Lemma 6(1), shows that the optimal service rate, μ^* is decreasing in α (just as in the large market scenario case). The service provider spends more time on each customer as the service becomes labor-intensive. Before examining part (2), let us examine the result (3) on equilibrium arrival rates. The equilibrium arrival rates are constant since the service provider serves all customers in equilibrium. Therefore, note that as labor-intensity α increases, the optimal service rate falls while the equilibrium arrival rate remains unchanged. This leads to increased waiting costs, as α increases. In fact, the expected waiting cost is $\sqrt{c\alpha}$. In equilibrium, if the market demand is low, customers wait longer as the service becomes more labor-intensive.

Part (2) of the Lemma 6 notes that the optimal price is convex in α . When $\Lambda > \mu_b$ the optimal price is always decreasing. When $\Lambda < \mu_b$, for $\alpha < \frac{c}{(\mu_b - \Lambda)^2}$, the optimal price is decreasing, and for $\alpha > \frac{c}{(\mu_b - \Lambda)^2}$ it is increasing. From Lemma 6(1), we noted that when α

increases the service provider increases the time of service per customer, thus providing a higher value service. However when $\alpha < \frac{c}{(\mu_b - \Lambda)^2}$, the higher waiting cost (due to increased service time) dominates any increase in service value, leading to a degradation in the net value of the service for the customers. To accommodate this loss in value, the service provider has to cut prices as α increases. Hence, the optimal price is decreasing in α for this region. Thus, comparing two services of low to medium labor-intensity (i.e. $\alpha < \frac{c}{(\mu_b - \Lambda)^2}$), the more labor-intensive will be both less expensive and less congested than the other.

However, when α is high ($> \frac{c}{(\mu_b - \Lambda)^2}$), the gains in service value are significant enough to dominate equilibrium waiting costs as α increases. Therefore, the optimal price is increasing in α when labor-intensity is high (i.e. α is greater than described threshold). In contrast to the above comparison when α was low, when we compare two services that are both highly labor-intensive ($\alpha > \frac{c}{(\mu_b - \Lambda)^2}$), the service with higher α is both more expensive and more congested than the other.

Finally, the optimal service rate increases with waiting costs. For the same α , the service provider increases the pace of operations to mitigate higher waiting costs. As the value degrades, the optimal price charged also decreases. Even though the equilibrium demand has not changed and the waiting times are shorter, the increased waiting cost (c) forces the service provider to reduce prices. Thus, for services in which customers are less patient, the quality of the service is *lower* than a labor-intensive service in which customers are more patient. Customer waiting costs only precipitate the fall in quality.

3.5 Inference from Single Server Results

Note that Lemma 1 and Lemma 5 describe the the properties of the optimal price and the corresponding equilibrium demand for $\Lambda > \bar{\lambda}_\alpha$ and $\Lambda \leq \bar{\lambda}_\alpha$ respectively. Thus they cover the entire space of potential demand rates. Similarly, Lemma 2 and Lemma 6 provide the optimal service rate decisions when the potential demand Λ is such that $\Lambda > \lambda_\alpha^*$ (large market scenario) and $\Lambda \leq \lambda_\alpha^*$ (small market scenario). Thus, the two Lemmas together cover the optimal service rate decisions for all potential demands. Lemma 4 compares the maximum throughput $\bar{\lambda}_\alpha$ and the optimal equilibrium throughput λ_α^* . Figure 2 provides an illustrative comparison of λ_α^* and $\bar{\lambda}_\alpha$. We can now combine all the results proven in the aforementioned Lemmas, and state the following proposition that summarizes the service providers' revenue maximizing policy.

Proposition 2 *The service rate maximizing the service provider's revenues is given as follows:*

$$\mu^* = \begin{cases} \Lambda + \sqrt{c/\alpha} & \text{if } \Lambda \leq \frac{V_b + \alpha\mu_b}{2\alpha} - \sqrt{c/\alpha} \\ \frac{V_b + \alpha\mu_b}{2\alpha} & \text{if } \Lambda > \frac{V_b + \alpha\mu_b}{2\alpha} - \sqrt{c/\alpha}. \end{cases} \quad (14)$$

the optimal price is given by:

$$p^*(\mu^*) = \begin{cases} V_b + \alpha\mu_b - \alpha\Lambda - 2\sqrt{\alpha c} & \text{if } \Lambda \leq \frac{V_b + \alpha\mu_b}{2\alpha} - \sqrt{c/\alpha} \\ \frac{V_b + \alpha\mu_b - 2\sqrt{\alpha c}}{2} & \text{if } \Lambda > \frac{V_b + \alpha\mu_b}{2\alpha} - \sqrt{c/\alpha}, \end{cases} \quad (15)$$

the optimal equilibrium demand is given by:

$$\lambda_e(\mu^*, p^*(\mu^*)) = \begin{cases} \Lambda & \text{if } \Lambda \leq \frac{V_b + \alpha\mu_b}{2\alpha} - \sqrt{c/\alpha} \\ \frac{V_b + \alpha\mu_b - 2\sqrt{\alpha c}}{2\alpha} & \text{if } \Lambda > \frac{V_b + \alpha\mu_b}{2\alpha} - \sqrt{c/\alpha}, \end{cases} \quad (16)$$

and the service provider's optimal revenue is given by:

$$R(\mu^*, p^*(\mu^*)) = \begin{cases} \Lambda(V_b + \alpha\mu_b - \alpha\Lambda - 2\sqrt{\alpha c}) & \text{if } \Lambda \leq \frac{V_b + \alpha\mu_b}{2\alpha} - \sqrt{c/\alpha} \\ \frac{(V_b + \alpha\mu_b - 2\sqrt{\alpha c})^2}{4\alpha} & \text{if } \Lambda > \frac{V_b + \alpha\mu_b}{2\alpha} - \sqrt{c/\alpha}, \end{cases} \quad (17)$$

Proposition 2 characterizes the optimal service rate and price policy of the service provider for all levels of potential demand, Λ . In all cases, the monopolist extracts the consumer surplus fully. When the potential demand for the service is low ($\Lambda \leq \frac{V_b + \alpha\mu_b}{2\alpha} - \sqrt{c/\alpha}$), the service provider serves all potential customers, maximizing the net value of an arriving customer ($V(\mu) - WC(\mu, \Lambda)$). The resulting surplus is fully extracted by price as self-interested customers continue procuring the service as long as the net benefit is non-negative. Finally, note that the revenues are lowest at some intermediate α . Thus, when service are very labor intensive (high α), reducing the labor intensity might end-up reducing the revenues of the firm.

4. Service Provider with Multiple Servers

In this section, we consider the effect of multiple servers owned by a single service provider (firm) who provides a labor-intensive service of type α . Although the service provider sets the price to maximize total revenues, the individual servers have the flexibility to set their own service speed (and hence, quality). Examples of such scenarios include tax consultants at a single tax firm who handle customers under the firm's pricing policy, or surgeons who belong to the same health network or the same hospital which in turn determines the admission

price for surgeries and medical operations. The firm consists of multiple servers. We initially restrict our attention to two servers for modeling ease, and then show our results extend to multiple servers. Each server individually decides its service rate μ , (and hence, the corresponding service value) to maximize its own revenues under the endogenously-decided price p set by the firm. Arriving customers decide on which server to go to, based on the expected waiting times and the service value offered by the servers under the prices set by the firm. Thus, we model each server using an $M/M/1$ queueing regime. The separate queues setting demonstrates two different servers competing in a common market of homogenous customers.

Customers in the market arrive to the service provider, and decide whether to join the queues at the servers, and if they join, they also decide which queue to join. Customers' queue joining decision occurs without observing the queue lengths, and it is based on the price, expected waiting cost and service quality offered at each server. The queue joining decision of a customer is given by $\gamma_j(\mu_1, \mu_2, p, \Lambda)$, for $j = 0, 1, 2$, where γ_0 denotes the probability of balking, and γ_1 and γ_2 denote the probability of joining queue 1 and 2 respectively. Under pure consumer strategies, i.e. $\gamma_i = 1$ for some i , either one of the servers serves all the customers (Λ), or none of the servers serves any customers. We find that none of those cases occur in equilibrium. We thus focus on mixed consumer strategies to reflect the effects of competition on service provider's strategy.

Again, as in Section 3, we divide our analysis into cases of large market scenario and small market scenario. When the market is sufficiently large, i.e. $\Lambda \geq 2 \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha} = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}$, we show that there are no competition effects. In equilibrium, both the servers choose the service rates as if they were monopolies, and the firm chooses single-server monopoly price. Under low potential demand, i.e. when $\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}$, we show that the firm chooses a price such that all consumer surplus is extracted, and the market demand is fully split between the two servers.

Given $\Lambda \geq \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}$, in equilibrium, the customers are indifferent between the queues, and therefore join one of the queues (or balk). However, we find that competition has no effect on the servers' optimal strategies; the equilibrium decisions are identical to the single-server case. The market is large enough that both the servers can choose their optimal service rates, without having to compete for the customer demand at the other server. The servers' optimal service rate decisions in the case of high potential demand is captured in the following proposition.

Proposition 3 *When the potential demand for the service is high, i.e. $\Lambda \geq \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}$, servers act as monopolists. The optimal service rate set by the servers is given by: $\mu_i^* = \frac{V_b + \alpha\mu_b}{2\alpha}$. The optimal price is $p^* = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2}$.*

Proposition 3 simply states that in a market large enough (i.e. large Λ), the price charged by the service provider remains unaffected by the presence of another server in the service provider's network. Since the single server monopoly results derived in the previous section are retained for high demand, all our insights on labor-intensive services continue to hold.

When the market is smaller, i.e. when $\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}$, competition affects the service providers' strategies. The servers compete by adjusting their service rates, while the firm charges some admission price for the service. One ponderable question rests on the servers' service rate decision: Do they opt for fast service rate (to improve throughput), or do they choose higher service value (to allow the firm to choose a higher price)? When both servers have positive market share, we find that the net value ($V(\mu) - WC(\mu_i, \lambda_i) - p$) provided by the servers are equal and positive in the equilibrium. Server i 's equilibrium demand is λ_i , and the whole market is covered by the two servers (i.e., $\Lambda = \lambda_1 + \lambda_2$). The following proposition shows that in equilibrium the servers share the market demand equally, and the service provider extracts the consumer welfare by a charging a corresponding price p^* .

Proposition 4 *When the potential demand for the service is low, $\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}$, the two servers share the market demand, by setting their service at $\mu_i^e = \frac{\Lambda}{2} + \sqrt{c/\alpha} \forall i$.*

Proposition 4 shows, in equilibrium the service provider provides higher service value at a slower rate through its servers, than it would do so if there was only one server. However, note that in equilibrium, the customers still continue to wait the same time to acquire service as in the corresponding single server case. The expected waiting cost of an arriving customer, $WC(\mu_i^e, \Lambda/2) = \sqrt{c\alpha}$ is equal to the expected waiting cost in the case of single server, $WC(\mu^*, \Lambda) = \sqrt{c\alpha}$. Therefore, the net value of the service increases when there are more servers competing for market share under a single service provider. Our structural results continue to hold when there are n servers competing on service rate. In markets that are sufficiently large, the servers behave as local monopolies. When $\Lambda < n \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$, each server provides a slower service than it would in the single-server case.

Corollary 1 *The service provider charges a higher price when there are two servers compared to the case when there is only one server. When the firm has only one server, the*

optimal price charged is $p_M = V(\Lambda + \sqrt{c/\alpha}) - \sqrt{c\alpha}$. When there are two servers competing for the market share the price charged for the service equals $p_2^* = V(\mu_i^e) - \sqrt{c\alpha} > p_M$.

The behavior of the optimal equilibrium price p^* with respect to labor-intensity α remains identical (i.e. optimal price is still convex in α), except that α where the price is minimized has changed from the single-server case.

More interestingly, Corollary 1 shows that when there are competing servers, the total revenues exceed the revenue gained by a single monopolistic server by providing higher net value to consumers. The additional consumer surplus that is generated is extracted by the firm by charging a higher price. Since the size of the market (equilibrium arrival rate) remains unchanged, the total revenue is higher than the monopoly revenue. This result is intriguing. Competition in service rates *benefits* the service provider's revenues (compared to the single server monopoly), unlike competition in prices.

In fact, we also note that the prices (and the total market revenue) is increasing in the number of servers competing. This is captured in the following corollary.

Corollary 2 When $\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}$, if there are n servers competing on market share, the admission price is non-decreasing in n . Furthermore, the optimal prices exhibit decreasing differences property as n increases.

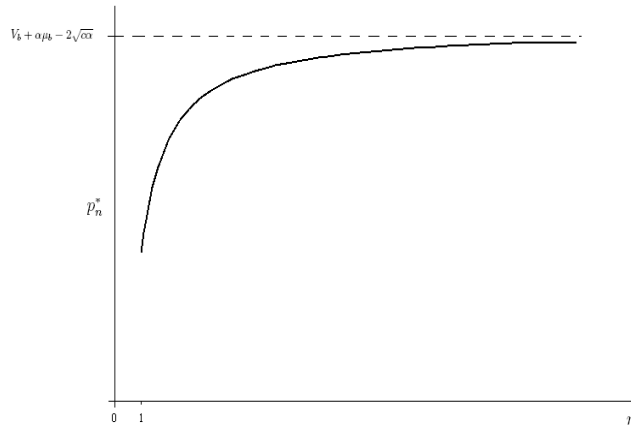


Figure 3: Admission price p_n^* is increasing and concave in the number of servers, n .

Corollary 2 shows that the price charged by the network service provider is *increasing* as the number of servers providing the service is increasing. Also see Figure 3. Note that

demand is likely to be in small-market region, as the number of competing servers increase. The result is intriguing; we note that the higher price comes due to the higher quality of service provided in the market. Therefore, a firm focussed on revenues (such as in the concierge care example), can lead to provision of higher service value (although, for a limited set of customers). Although the servers compete among themselves for the customers, they choose to produce higher service value instead of servicing faster, which in turn allows the firm to charge higher admission prices. Our result provides a hypothesis that can be tested empirically. In fact, for all n , collusion between the servers would also lead to identical revenues.

Finally, there could be (dis)economies of scale in the design of labor-intensive service provision. For example, there might be convex costs involved in adding more servers when providing labor-intensive services. In such cases, we can calculate the optimal number of servers a service providers might choose to have, when providing labor-intensive services.

5. Conclusion and Future Directions

In labor-intensive services, the service quality depends on the time spent with the customer. We have argued that traditional queueing models do *not* apply to such settings, wherein the tradeoff between *quality* and *speed* is at the crux of the service-provider’s problem. In our queueing model, service quality degraded in both the service speed and the labor-intensity. The service-provider’s optimal choice of an *intermediate* service rate in the face of self-interested, rational customers, reflects the above tradeoff.

Assuming convex and increasing investment costs to improve the service rate, within a conventional queueing model, might also result in intermediate optimal service rates; however investment costs do not *directly* affect the equilibrium queue choice of customers. Thus our model provides fundamentally new insights into the nature of labor-intensive services.

An implication of service degradation with speed is that service-level (quality) targets are met only at slow service times, necessitating a larger investment in capacity/service rates. This raises the costs of providing the service. Thus, *Baumol’s cost disease* (discussed in Section 1) is a direct consequence of the labor-intensity of the service. What could exacerbate this disease is our analytical result that, as the service gets more labor-intensive, the service provider *slows down*, and increases the time spent with each customer. This supports the empirical observation (*cf* Baumol (1993), Surowiecki (2003), Triplett and Bosworth (2004),

Varian (2004), Yarnall *et al* (2003)) that labor-intensive services suffer from low productivity/throughput compared to the service sector overall.

We find that in equilibrium, the service-provider's price and customer demand for the service move in tandem when the labor-intensity of a service changes. Both are convex functions of the labor intensity, and they rise and fall together as the labor-intensity of the service changes, in stark contrast to the price-demand relationship in conventional markets (After all, downward-sloping demand curves are *de rigueur* in Economic models). This result arises from (i) the quality sensitivity of customers, coupled with (ii) the quality-speed inter-relationship in labor-intensive services. Since revenues are a product of price and demand, this has clear implications for the service-provider. For example, if he were already operating a highly labor-intensive service, increasing the labor intensity further would improve his revenues (as we proved). Methods to alter the labor-intensity of a service include reconfiguring the service process to make it less sensitive to service time, training the workforce to reduce degradation of service quality with time, and bundling services appropriately so that the resultant composite service has greater labor-intensity.

The structural results from our monopoly analysis continue to hold even if the individual servers are forced into service-rate competition under the price set by the service provider. If the potential market is sufficiently large, the servers act as local monopolies, and the single-server results apply. If the market is small enough, the servers must compete for customers, and they do so by operating at *slower* service rates. As the number of competing servers increases, the servers slow down further. As a consequence, server-competition simultaneously (i) increases the price charged by the service-provider of the labor-intensive service, and (ii) enhances the service value in equilibrium, while (iii) holding the equilibrium congestion (waiting) costs constant. These results, which are in sharp contrast to previous research in service operations and economics, are driven by labor-intensity.

Several directions seem promising for future research. One obvious extension is to study different kinds of heterogeneity. Competing servers could vary in their labor-intensities (based on their different positioning, choice of service-process and investments in work-force training). Alternatively, one could model heterogeneity in customers' waiting costs— After all, we expect some customers to be more patient than others. These settings would probably lead to asymmetric equilibria in service rates, prices, market-shares and profits. Whether *labor-intensity differentiation* is a viable competitive strategy or not is an interesting research question.

A second extension would be to model multiple service providers that independently set their prices and service rates. Deriving solutions analytically in closed form for this problem, as we did in the present paper, is known to be intractable. Hence, a simplification of the model in other ways (such as eliminating customer choice by assuming an exogenously specified demand rate) and/or employing other methodologies such as computational approaches will almost certainly be required.

A third interesting extension of this research would be to model information asymmetry—in the value of the service in general, and in labor intensity in particular (These were assumed to be common knowledge in the present research). Presumably, there are occasions when customers do not know the exact content and complexity of the service. Service-providers have an incentive to mis-represent the value of their services. Debo *et al* (2008) model incentive effects in the context of credence services; similar issues are pertinent to labor-intensive services.

References

- AARP Bulletin Today. 2004. Want Your Doctor to Pamper You? Pay Extra. Oct. 12, 2004.
- Armony M., M. Haviv. 2000. Price and Delay Competition between Two Service Providers. *The European Journal of Operational Research*, **147**(1), 32–50.
- Baumol, W. J., W. G. Bowen. 1966. Performing Arts: The Economic Dilemma. New York: The Twentieth Century Fund.
- Baumol, W. J. 1993. Health Care, Education, and the Cost Disease: A Looming Crisis for Public Choice. *Public Choice*. Volume 77, Number 1, September, 1993. 17–28.
- Baumol, W. J. 1996. Children of Performing Arts, the Economic Dilemma: The Climbing Costs of Health care and Education. *Journal of Cultural Economics*, **20**, 183–236.
- Cachon, G., P. Harker. 2002. Competition and Outsourcing with Scale Economies. *Management Science*, **48**(10), 1314–1333.
- Chase, R.B., D.A. Tansik. 1983. The Customer Contact Model for Organization Design. *Management Science*, **29**(9), 1037–1050.
- Debo, L.G., L.B. Toktay, L. N. Van Wassenhove. 2008. Queueing for Expert Services. *Management Science*, **54**(8), 1497–1512.

- Gans, N. 2002. Customer Loyalty and Supplier Quality Competition. *Management Science*, **48**(2), 207–221.
- Gilbert, S. M., Z. K. Weng. 1998. Incentive Effects Favor Non-Consolidating Queues in a Service System: The Principal Agent Perspective. *Management Science*, **44**(12), 1662–1669.
- Hassin, R., M. Haviv. 2003. To Queue or not to Queue: Equilibrium behavior in queuing systems. Kluwer Academic Publishers, Norwell, MA.
- Hopp, W.J., S. M. R. Iravani, G. Y. Yuen. 2007. Operations Systems with Discretionary Task Completion. *Management Science*, **53**(1), 61–77.
- Kalai, E., M. Kamien, M. Rubinovitch. 1992. Optimal Service Speeds in a Competitive Environment. *Management Science*, **38**(8), 1154–1163.
- Kaminetsky, B. 2004. Testimony to the Joint Economic Committee of the United States Congress. Consumer-Directed Doctoring: The Doctor is in, Even if Insurance is Out. Wednesday, April 28, 2004.
- Lederer, P. J., L. Li. 1997. Pricing, production, scheduling, and delivery-time competition. *Operations Research* **45**(3), 407–420.
- Li, L. 1992. The role of inventory in delivery time-competition. *Management Science*, **38**(2) 182-197.
- Li, L., Y. S. Lee. 1994. Pricing and Delivery-Time Performance in a Competitive Environment. *Management Science*, **40**(5), 633–646.
- Loch, C. 1994. Incentive compatible equilibria in markets with time competition. Working paper, INSEAD, Fontainebleau, France.
- Lu, L., J. Van Mieghem, C. Savaskan. 2008. Incentives for Quality Through Endogenous Routing. *Manufacturing and Service Operations Management*. Forthcoming.
- Naor, P. 1969. The Regulation of Queue Sizes by Levying Tolls. *Econometrica*, **37**(1), 15–24.
- Oliva, R., R. J. Sterman. 2001. Cutting Corners and Working Overtime: Quality Erosion in the Service Industry. *Management Science*, **47**(7), 894–914.
- Png, I., D. Reitman. 1994. Service Time Competition. *RAND Journal of Economics*, **25**(4), 619–634.
- Ren, Z.J., X. Wang. 2008. Should Patients be Steered to High Volume Hospitals? An

- Empirical Investigation of Hospital Volume and Operations Service Quality. Working Paper. Boston University.
- Ross, Sheldon M. 2006. Introduction to Probability Models. Academic Press. Ninth Edition.
- Shaw, K. A., D. A. Black. 2001. View: Why College Costs So Much. *NY Times*, April 8.
- Surowiecki, J. 2003. What Ails us? *The New Yorker*, July 7, 2003.
- Triplett, J. E., B. P. Bosworth. 2004. Productivity in the U.S. Services Sector, New Sources of Economic Growth. Brookings Institution Press, Washington, D.C.
- Varian, H. 2004. Economic Scene; Information Technology May Have Been What Cured Low Service-Sector Productivity. *NY Times*, Published: February 12, 2004.
- Veeraraghavan, S., L. G. Debo. 2008. Joining Longer Queues: Information Externalities in Queue Choice. *Manufacturing and Service Operations Management*. Forthcoming.
- Yarnall, K. S. H., K. I. Pollak, T. Ostbye, K. M. Krause and J. L. Michener, 2003. Primary Care: Is There Enough Time for Prevention? April 2003, Vol 93, No. 4, *American Journal of Public Health* 635-641.
- Wang, X., L. G. Debo, A. Scheller-Wolf. 2008. Design and Analysis of Diagnostic Service Centers. Working paper, Carnegie Mellon University, Pittsburgh, USA.

Online Technical Appendix

Quality-Speed Conundrum: Tradeoffs in Labor-Intensive Services

Krishnan Anand, M. Fazıl Paç, Senthil K. Veeraraghavan

Proof of Proposition 1: We begin by showing the optimal price, $p^*(\mu)$ for $\Lambda > A_2(\alpha)$. In this case, the service provider cannot serve all potential customers even when the price is equal to zero. The equilibrium arrival rate, $\lambda_e(\mu, p)$, is determined by the following equation in this case:

$$V(\mu) - p = WC(\mu, \lambda_e(\mu, p)). \quad (1)$$

The revenue of the service provider, $R(\mu, p)$, is given by:

$$R(\mu, p) = p \left(\mu - \frac{c}{V(\mu) - p} \right). \quad (2)$$

Recall that the service value is an upper-bound for the price, i.e. $V(\mu) \geq p$. Therefore the revenue function is concave in the price, p , for the set of admissible prices (for $0 \leq p \leq V(\mu)$), as the second order condition is negative:

$$0 > \frac{\delta^2 R(\mu, p)}{\delta p^2} = -\frac{2c}{(V(\mu) - p)^2} - \frac{2pc}{(V(\mu) - p)^3}$$

The optimal price, maximizing the service provider's revenues for a given service rate μ is $p^*(\mu) = V(\mu) - \sqrt{cV(\mu)/\mu}$. We find the optimal price using the first order condition:

$$0 = \frac{\delta R(\mu, p)}{\delta p} = \mu - \frac{c}{V - p} - \frac{pc}{(V - p)^2}.$$

The first order condition is satisfied at $p^* = V(\mu) - \sqrt{cV(\mu)/\mu}$, which is in the set of admissible prices, $p \in [0, V(\mu)]$. Plugging $p^*(\mu)$ into equation (1) we find the resulting equilibrium arrival rate as:

$$\lambda_e(\mu, p^*(\mu)) = \mu - \sqrt{\frac{c\mu}{V(\mu)}}.$$

The equilibrium arrival rate, $\lambda_e(\mu, p^*(\mu))$, is independent of the potential demand, Λ . This shows that the optimal price at service rate μ is equal to $V(\mu) - \sqrt{cV(\mu)/\mu}$ for all $\Lambda \geq \mu - \sqrt{\frac{c\mu}{V(\mu)}}$. So far, we have derived the optimal price p^* for all $\Lambda \geq \mu - \sqrt{\frac{c\mu}{V(\mu)}}$.

To complete the proof we need to derive the optimal price for $\Lambda < \mu - \sqrt{\frac{c\mu}{V(\mu)}}$. Note that the service provider can serve all potential customers at a non-negative price for $\Lambda \leq$

$\mu - \sqrt{\frac{c\mu}{V(\mu)}}$: For a given service rate μ , the equilibrium demand, $\lambda_e(\mu, p)$, is decreasing in price. Therefore the maximum number of customers that can be served (maximum throughput) at rate μ , $\bar{\Lambda}(\mu)$, is found by setting the price equal to zero. Using the following equation we find $\bar{\Lambda}(\mu)$.

$$V(\mu) = \frac{c}{\mu - \bar{\Lambda}(\mu)} \Rightarrow \bar{\Lambda}(\mu) = \mu - \frac{c}{V(\mu)}$$

If $\bar{\Lambda}(\mu)$ is greater than the potential demand Λ , then the service provider can serve all potential customers, charging a price greater than zero. For $\Lambda < \mu - \sqrt{\frac{c\mu}{V(\mu)}}$, $\bar{\Lambda}(\mu) > \Lambda$:

$$\bar{\Lambda}(\mu) = \mu - \frac{c}{V(\mu)} \geq \mu - \sqrt{\frac{c\mu}{V(\mu)}} > \Lambda, \text{ since } V(\mu) \geq \frac{c}{\mu} \text{ for all } \mu \in \mathcal{F}(\alpha).$$

For $\Lambda < \mu - \sqrt{\frac{c\mu}{V(\mu)}}$ all arriving customers join the queue, if the net benefit of joining the queue when all others join the queue at price p is non-negative, i.e. $V(\mu) - p - WC(\mu, \Lambda) \geq 0$. The net benefit is decreasing in price, and it is non-negative for $p \leq V(\mu) - \frac{c}{\mu - \Lambda}$. Increasing the price further reduces the equilibrium arrival rate. Then the service provider's revenue as a function of price can be written as:

$$R(\mu, p) = \begin{cases} p\Lambda & \text{if } 0 \leq p \leq V(\mu) - WC(\mu, \Lambda) \\ p \left(\mu - \frac{c}{V(\mu) - p} \right) & \text{if } V(\mu) - WC(\mu, \Lambda) < p \leq V(\mu) - \frac{c}{\mu} \\ 0 & \text{if } p > V(\mu) - \frac{c}{\mu}, \end{cases} \quad (3)$$

Differentiating the revenue function with respect to price we get:

$$\frac{\delta R(\mu, p)}{\delta p} = \begin{cases} \Lambda & \text{if } 0 \leq p \leq V(\mu) - \frac{c}{\mu - \Lambda} \\ \mu - \frac{c}{V(\mu) - p} - \frac{pc}{(V(\mu) - p)^2} & \text{if } V(\mu) - \frac{c}{\mu - \Lambda} < p \leq V(\mu) - \frac{c}{\mu} \\ 0 & \text{if } p > V(\mu) - \frac{c}{\mu}, \end{cases} \quad (4)$$

The revenue, $R(\mu, p)$ is clearly increasing in the price for $p \leq V(\mu) - \frac{c}{\mu - \Lambda}$. Increasing the price further at $p = V(\mu) - \frac{c}{\mu - \Lambda}$ will decrease the demand (throughput) but it may still increase the revenues. Note that the revenue function for $p \geq V(\mu) - \frac{c}{\mu - \Lambda}$ is equivalent to the revenue function given by equation (2), which is maximized at $p = V(\mu) - \sqrt{\frac{cV(\mu)}{\mu}}$. The revenues are decreasing in price at $p = V(\mu) - \frac{c}{\mu - \Lambda}$ because $V(\mu) - \frac{c}{\mu - \Lambda} > V(\mu) - \sqrt{\frac{cV(\mu)}{\mu}}$ for $\Lambda < \mu - \sqrt{\frac{c\mu}{V(\mu)}}$:

$$\sqrt{\frac{cV(\mu)}{\mu}} = \frac{c}{\mu - (\mu - \sqrt{\frac{c\mu}{V(\mu)}})} > \frac{c}{\mu - \Lambda}.$$

As a result, the optimal price at service rate μ , for $\Lambda < \mu - \sqrt{\frac{c\mu}{V(\mu)}}$ is $p^*(\mu) = V(\mu) - \frac{c}{\mu - \Lambda}$. The resulting equilibrium arrival rate is equal to $\lambda_e(\mu, p^*(\mu)) = \Lambda$.

$$p^*(\mu) = \begin{cases} V(\mu) - WC(\mu, \Lambda) & \text{if } 0 \leq \Lambda \leq \hat{\lambda}(\mu) \\ V(\mu) - \sqrt{cV(\mu)/\mu} & \text{if } \hat{\lambda}(\mu) < \Lambda. \end{cases} \quad (5)$$

Thus we have derived p^* for all Λ . ■

Preparatory Results for Lemmas 1-6: Before we prove the lemmas, we prove two main preparatory results.

Result 1: Service provider's revenue function $R(\mu, p)$ is non-decreasing in the potential demand, Λ .

Proof: For a given service rate μ and price p , service provider's revenue as a function of the potential demand, Λ , is given as follows:

$$R(\mu, p) = \begin{cases} p\Lambda & \text{if } WC(\mu, \Lambda) \leq V(\mu) - p \\ p \left(\mu - \frac{c}{V(\mu) - p} \right) & \text{if } WC(\mu, 0) < V(\mu) - p < WC(\mu, \Lambda) \\ 0 & \text{if } V(\mu) - p < WC(\mu, 0). \end{cases} \quad (6)$$

$R(\mu, p)$ is continuous in Λ . To show this, we only need to show that the function is continuous at the transition point $WC(\mu, \Lambda) = V(\mu) - p$. Note that the last region, $V(\mu) - p < WC(\mu, 0)$, is independent of the potential demand, Λ .

$WC(\mu, \Lambda) = \frac{c}{\mu - \Lambda}$ is increasing in Λ for $\Lambda \leq \mu$. Rewriting the transition point, $WC(\mu, \Lambda) = V(\mu) - p$, and solving for Λ we get:

$$\frac{c}{\mu - \Lambda} = V(\mu) - p \Rightarrow \Lambda = \mu - \frac{c}{V(\mu) - p}.$$

Therefore $R(\mu, p)$ is continuous in Λ for $\Lambda \geq 0$.

Clearly, $R(\mu, p)$ is increasing in Λ for $WC(\mu, \Lambda) \leq V(\mu) - p$ and constant in Λ for $WC(\mu, \Lambda) > V(\mu) - p$. Which proves that $R(\mu, p)$ is non-decreasing in Λ for $\Lambda \geq 0$. ■

Result 2: For $\Lambda \geq \bar{\lambda}_\alpha = \max_{\{\mu \in \mathcal{F}(\alpha)\}} \{\lambda_e(\mu, p^*(\mu))\}$, the optimal price, $p^*(\mu)$, and the resulting equilibrium arrival rate, $\lambda_e(\mu, p^*(\mu))$, have the following symmetric relationship around $\beta = \left(\frac{V_b + \alpha \mu_b}{2\alpha} \right)$ for any $\mu \in \mathcal{F}(\alpha)$.

$$p^*(\beta + \epsilon) = \alpha \lambda_e(\beta - \epsilon, p^*(\beta - \epsilon)),$$

where $\epsilon = \mu - \frac{V_b + \alpha \mu_b}{2\alpha}$.

Proof: For $\Lambda > \bar{\lambda}_\alpha = \max_{\{\mu \in \mathcal{F}(\alpha)\}} \{\lambda_e(\mu, p^*(\mu))\}$, the optimal price for a given service rate is

given by $p^*(\mu) = V_b + \alpha\mu_b + \alpha\mu - \sqrt{c(V_b + \alpha\mu_b - \alpha\mu)/\mu}$, and the corresponding equilibrium demand is given by $\lambda_e(\mu, p^*(\mu)) = \mu - \sqrt{\frac{c\mu}{V_b + \alpha\mu_b - \alpha\mu}}$ (From Proposition 1).

Plugging in β for $\frac{V_b + \alpha\mu_b}{2\alpha}$ we get:

$$p^*(\mu) = 2\alpha\beta - \alpha\mu - \sqrt{c(2\alpha\beta - \alpha\mu)/\mu},$$

$$\lambda_e(\mu, p^*(\mu)) = \mu - \sqrt{\frac{c\mu}{2\alpha\beta - \alpha\mu}}.$$

Let $\epsilon = \mu - \beta$, therefore $\mu = \beta + \epsilon$. We can rewrite $p^*(\mu)$ and $\lambda_e(\mu, p^*(\mu))$ as:

$$p^*(\mu) = 2\alpha\beta - \alpha(\beta + \epsilon) - \sqrt{c(2\alpha\beta - \alpha(\beta + \epsilon))/(\beta + \epsilon)} = \alpha(\beta - \epsilon) + \sqrt{c\alpha(\beta - \epsilon)/(\beta + \epsilon)}.$$

$$\lambda_e(\mu, p^*(\mu)) = \beta + \epsilon - \sqrt{\frac{c(\beta + \epsilon)}{\alpha(\beta - \epsilon)}}$$

$$\Rightarrow p^*(\beta + \epsilon) = \alpha\lambda_e(\mu - \epsilon, p^*(\mu - \epsilon)).$$

Hence $p^*(\mu)$ and $\alpha\lambda_e(\mu, p^*(\mu))$ are symmetric around β . ■

Having proven Results 1 and 2, we are now ready to prove the lemmas.

Proof of Lemma 1: For, $\Lambda > \bar{\lambda}_\alpha$, $p^*(\mu)$ and $\lambda_e(\mu, p^*(\mu))$ are uni-modal (increasing and then decreasing) in the service rate, μ : We begin by proving uni-modality of the optimal price $p^*(\mu)$ in μ . For $\Lambda > \bar{\Lambda}_\alpha$, the optimal price for service rate $\mu \in \mathcal{F}(\alpha)$ is given by $p^*(\mu) = V(\mu) - \sqrt{\frac{cV(\mu)}{\mu}}$.

$p^*(\mu)$ is equal to zero for $\mu = A_i(\alpha)$ for $i = 1, 2$. We pick an interior point β in the operating region, $\mathcal{F}(\alpha)$ such that $\beta = \frac{V_b + \alpha\mu_b}{2\alpha}$. Clearly $\beta \in (A_1, A_2)$ for $\alpha, c > 0$.

The optimal price for β , $p^*(\beta)$, is non-negative as long as the condition in Remark 1 holds. For $\alpha, c \geq 0$,

$$p^*(\beta) = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha} > 0 \Leftrightarrow (V_b + \alpha\mu_b)^2 - 4c\alpha > 0.$$

If $V_b > c/\mu_b$, then $(V_b + \alpha\mu_b)^2 - 4c\alpha > (V_b - \alpha\mu_b)^2 > 0$, since $c < V_b\mu_b$ from Remark 1.

We will prove the uni-modality of $p^*(\mu)$ by showing that the first derivative $\frac{\delta p^*(\mu)}{\delta \mu}$ crosses 0 only once in $(A_1(\alpha), A_2(\alpha))$ and this point is the maximizer of the price, $p^*(\mu)$, with respect to μ for $\mu \in (A_1(\alpha), A_2(\alpha))$. Hence the first order condition is satisfied at a unique, interior point.

$$FOC : \frac{\delta p^*(\mu)}{\delta \mu} = -\alpha + \frac{cK}{2\mu^2 \sqrt{\frac{cV(\mu)}{\mu}}} = 0$$

The first derivative (right hand side of the first order condition) is continuous for $\mu \in (A_1(\alpha), A_2(\alpha))$. Reorganizing the terms we can write the first order condition as:

$$2\alpha\mu^2 \sqrt{\frac{cV(\mu)}{\mu}} = c(V_b + \alpha\mu_b).$$

For notational convenience let $K = (V_b + \alpha\mu_b)$. Therefore $V(\mu) = K - \alpha\mu$. Plugging in K and squaring both sides of the equation we get:

$$\begin{aligned} 4\alpha\mu^4 \frac{c(K - \alpha\mu)}{\mu} &= c^2 K^2 \\ \Rightarrow \frac{cK^2}{4\alpha^2} &= \mu^3(K - \alpha\mu). \end{aligned} \quad (7)$$

Note that the left hand side of equation (7) is constant with respect to μ . We show that the right hand side crosses this constant only once for $\mu \in [A_1(\alpha), A_2(\alpha)]$, therefore the first order condition is satisfied only once in the operating region.

The right hand side, $\mu^3(K - \alpha\mu)$, is uni-modal in the service rate μ :

$$\frac{\delta}{\delta \mu} \mu^3(K - \alpha\mu) = \mu^2(3K - 4\alpha\mu).$$

This shows that the right hand side is increasing in μ for $\mu < \frac{3K}{4\alpha}$ and decreasing for $\mu > \frac{3K}{4\alpha}$.

We now show that the RHS term in equation (7) is less than $\frac{cK^2}{4\alpha^2}$ when $\mu = A_1(\alpha)$ and greater than $\frac{cK^2}{4\alpha^2}$ when $\mu = A_2(\alpha)$, which proves that the right hand side crosses the left hand side only once in the operating region, $\mathcal{F}(\alpha)$, since the right hand-side is uni-modal in μ .

If we plug in K for the value of $A_i(\alpha)$, we get:

$$A_i(\alpha) = \frac{K \mp \sqrt{K^2 - 4\alpha c}}{2\alpha} \quad \text{for } i = 1, 2.$$

Let $RHS(\mu) = \mu^3(K - \alpha\mu)$. Let $LHS = \frac{cK^2}{4\alpha^2}$

$$RHS(A_1(\alpha)) = \frac{c(K - \sqrt{K^2 - 4\alpha c})}{4\alpha^2} \leq LHS \quad \text{and}$$

$$RHS(A_2(\alpha)) = \frac{c(K + \sqrt{K^2 - 4\alpha c})}{4\alpha^2} \geq LHS,$$

which shows the desired result. Note that the weak inequalities in the above equations are strict if $\alpha > 0$ and they are replaced with equalities if $\alpha = 0$.

The point at which the first order condition is satisfied is a maximizer because the price is equal to zero at the end points of the operating region, $\mathcal{F}(\alpha)$, i.e. at $A_1(\alpha)$ and $A_2(\alpha)$ ($p^*(A_i(\alpha)) = 0$). Furthermore the price is positive in the operating region, $\mathcal{F}(\alpha)$, since the optimal price is positive at an interior point β , i.e. $p^*(\beta) > 0$ and the first derivative of $p^*(\mu)$ crosses zero only once.

Thus we have shown that $p^*(\mu)$ is uni-modal (increasing and then decreasing) in the service rate μ . Uni-modality of the corresponding equilibrium demand, $\lambda_e(\mu, p^*(\mu))$ follows from Result 2, which implies a symmetry relation between $p^*(\mu)$ and $\lambda_e(\mu, p^*(\mu))$. ■

Proof of Lemma 2:

1. For $\Lambda > A_2(\alpha)$, the objective function in equation 10 is uni-modal in the service rate, μ : The revenue function $R(\mu, p^*(\mu)) = \mu(V_b + \alpha\mu_b - \alpha\mu) - 2\sqrt{c\mu(V_b + \alpha\mu_b - \alpha\mu)} + c$ is continuous in μ for $\mu \in [A_1(\alpha), A_2(\alpha)]$.

The revenue function is differentiable in μ for $\mu \in [A_1(\alpha), A_2(\alpha)]$:

The first derivative $\frac{\delta R(\mu, p^*(\mu))}{\delta \mu} = V_b + \alpha\mu_b - 2\alpha\mu - \frac{c(V_b + \alpha\mu_b) - 2c\alpha\mu}{\sqrt{c\mu(V_b + \alpha\mu_b - \alpha\mu)}}$, exists and is continuous for $\mu \in [A_1(\alpha), A_2(\alpha)]$.

The first derivative, $\frac{\delta R(\mu, p^*(\mu))}{\delta \mu} = 0$, crosses 0 at three points; $A_1(\alpha)$, $\mu^* = \frac{V_b + \alpha\mu_b}{2\alpha}$ and $A_2(\alpha)$. $\mu^* \in (A_1(\alpha), A_2(\alpha))$ for $\alpha, c > 0$. As a result, μ^* is either the unique maximizer or the unique minimizer of $R(\mu, p^*(\mu))$ for $\mu \in [A_1(\alpha), A_2(\alpha)]$.

We show that μ^* maximizes the revenue function:

$R(A_1(\alpha), p^*(A_1(\alpha)))$ and $R(A_2(\alpha), p^*(A_2(\alpha)))$ are equal to zero because both the optimal price, $p^*(A_i(\alpha))$, and the resulting equilibrium arrival rate, $\lambda_e(A_i(\alpha), p^*(A_i(\alpha)))$, are clearly equal to zero, since the service value, $V(\mu)$, is equal to the waiting cost during the service, c/μ , at these points.

Now we show that $R(\mu^*, p^*(\mu^*))$ is greater than zero.

$R(\mu^*, p^*(\mu^*)) = \frac{(V_b + \alpha\mu_b - 2\sqrt{c\alpha})^2}{4\alpha} > 0$ for all $\alpha > 0$, since $V_b + \alpha\mu_b - 2\sqrt{c\alpha} > 0$ from Remark 1. This shows that $R(\mu, p^*(\mu))$ increasing in μ for $\mu \in (A_1(\alpha), \mu^*)$ and decreasing in μ for $\mu \in (\mu^*, A_2(\alpha))$, which proves that the optimal service rate is $\mu^* = \frac{V_b + \alpha\mu_b}{2\alpha}$.

2. From Proposition 1 $p^*(\mu^*) = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2}$.
3. From Proposition 1 $\lambda_e(\mu^*, p^*(\mu^*)) = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$.

This proves that the optimal service setting for $\Lambda > A_2(\alpha)$.

We now show that the above operating setting is optimal, even for all $\Lambda \geq \lambda_e^*(\mu^*, p^*(\mu^*)) = \frac{V_b + \alpha\mu_b - \sqrt{c\alpha}}{2\alpha}$.

The revenue function $R(\mu, p)$ is non-decreasing in the potential demand, Λ , for all $\mu, p, \Lambda \geq 0$ from Result 1. Therefore the optimal revenue $R(\mu^*, p^*(\mu^*))$ is non-decreasing in Λ for all $\Lambda \geq 0$.

The service provider can achieve $R(\frac{V_b + \alpha\mu_b}{2\alpha}, \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2})$ by serving $\frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$ customers at rate $\mu = \frac{V_b + \alpha\mu_b}{2\alpha}$ charging price $p = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2}$ for all $\Lambda \geq \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$. In other words, the optimal revenue for $\Lambda > A_2(\alpha)$ can be achieved by the identical operating setting (price and service rate) when the potential demand is lower than $A_2(\alpha)$ ($\frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha} \leq \Lambda \leq A_2(\alpha)$).

The optimal revenues are non-decreasing in the potential demand, Λ , therefore the above setting is optimal for all $\Lambda \geq \lambda_\alpha^* = \frac{V_b + \alpha\mu_b - \sqrt{c\alpha}}{2\alpha}$, since it is optimal for $\Lambda > A_2(\alpha)$ and $A_2(\alpha) > \frac{V_b + \alpha\mu_b - \sqrt{c\alpha}}{2\alpha}$. ■

Proof of Lemma 3:[α -symmetry] Lemma 2 indicates that β defined in Result 2 is equal to the optimal service rate, μ^* , for $\Lambda > \bar{\lambda}_\alpha$. Result of the Lemma immediately follows from Result 2, plugging in μ^* for β . ■

Proof of Lemma 4: The revenue maximizing equilibrium demand, $\lambda_e(\mu^*, p^*(\mu^*))$ is smaller than the maximum equilibrium demand, $\bar{\lambda}_\alpha = \max_{\{\mu \in \mathcal{F}(\alpha)\}} \{\lambda_e(\mu, p^*(\mu))\}$. By definition we have $\bar{\lambda}_\alpha > \lambda_e(\mu^*, p^*(\mu^*))$.

We know that $\lambda_e(\mu, p^*(\mu))$ is uni-modal in μ from Lemma 1. We first show that $\lambda_e(\mu, p^*(\mu))$ is increasing in the service rate at $\mu = \mu^*$.

Differentiating the equilibrium demand $\lambda_e(\mu, p^*(\mu))$ with respect to μ we get:

$$\frac{\delta \lambda_e(\mu, p^*(\mu))}{\delta \mu} = 1 - \frac{cK}{(K - \alpha\mu)\sqrt{c\mu(K - \alpha\mu)}}.$$

Evaluating the first derivative at μ^* we get:

$$\left. \frac{\delta \lambda_e(\mu, p^*(\mu))}{\delta \mu} \right|_{\mu=\mu^*} = 1 - 2\sqrt{ac}/K.$$

To prove the result we need to show that $1 - 2\sqrt{ac}/K \geq 0$. Re-organizing the equation, we get:

$$K^2 \geq 4\alpha c.$$

The inequality holds for all $K, \alpha, c \geq 0$, which follows from Remark 1:

Remark 1 suggests that $V_b > c/\mu_b$. Recall that $K = V_b + \alpha\mu_b$. Therefore the condition in Remark 1 is equivalent to $(K - \alpha\mu_b)(K - V_b) > \alpha c$. Both μ_b and V_b are non-negative. Therefore we can rewrite the condition of Remark 1 as follows:

$$(K - V_b)V_b > c\alpha \text{ for } V_b \in [0, K]. \quad (8)$$

Inequality 8 holds for all $V_b \in [0, K]$, therefore it holds for the value of V_b maximizing $(K - V_b)V_b$ in $[0, K]$.

$(K - V_b)V_b$ is maximized when $V_b = K/2$.

$$\max_{\{0 \leq V_b \leq K\}} \{(K - V_b)V_b\} = K^2/4 \Rightarrow \frac{K^2}{4} > \alpha c.$$

Therefore, Remark 1 implies the result $K^2 \geq 4\alpha c$, which in turn implies that

$$\left. \frac{\delta \lambda_e(\mu, p^*(\mu))}{\delta \mu} \right|_{\mu=\mu^*} \geq 0.$$

We have therefore shown that $\lambda_e(\mu, p^*(\mu))$ is increasing in the service rate at $\mu = \mu^*$. Therefore, the throughput maximizing service rate $\bar{\mu} \geq m\mu^*$.

The equilibrium price $p^*(\mu)$ is decreasing in μ at $\mu = \mu^*$. Otherwise both the price, $p^*(\mu)$, and the equilibrium demand, $\lambda_e(\mu, p^*(\mu))$, would be increasing in μ at μ^* , which contradicts with the optimality of μ^* . ■

Proof of Lemma 5: When the potential demand $\Lambda < \bar{\lambda}_\alpha$, there exists $\mu_1(\Lambda)$ and $\mu_2(\Lambda)$ in $\mathcal{F}(\alpha)$ such that $\lambda_e(\mu, p^*(\mu)) \geq \Lambda$ for $\mu \in [\mu_1(\Lambda), \mu_2(\Lambda)]$. The result of the Lemma follows from the fact that $\lambda_e(A_i(\alpha), p^*(A_i(\alpha))) = 0$ for $i = 1, 2$ and from the uni-modality of $\lambda_e(\mu, p^*(\mu))$ (Lemma 1). ■

Proof of Lemma 6: We will show that in the low potential demand scenario ($\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$) the service provider's optimal service rate is $\mu^* = \Lambda + \sqrt{c/\alpha}$, the price is $p^*(\mu^*) = V_b + \alpha\mu_b - \alpha\Lambda - 2\sqrt{c\alpha}$ and the optimal equilibrium demand is equal to Λ .

1. For the small market scenario, service provider's objective function is given by equation 13:

$$R(\mu, p^*(\mu)) = \begin{cases} \mu(K - \alpha\mu) - 2\sqrt{c\mu(K - \alpha\mu)} + c & \text{if } A_1(\alpha)\mu < \bar{\mu}_1(\Lambda) \\ (V(\mu) - \frac{c}{\mu - \Lambda})\Lambda & \text{if } \bar{\mu}_1(\Lambda) \leq \mu \leq \bar{\mu}_2(\Lambda) \\ \mu(K - \alpha\mu) - 2\sqrt{c\mu(K - \alpha\mu)} + c & \text{if } \bar{\mu}_2(\Lambda) \leq \mu < A_2(\alpha). \end{cases} \quad (9)$$

The objective function is continuous in μ as $\Lambda = \lambda_e(\mu, p^*(\mu))$ and $V(\mu) - WC(\mu, \Lambda) = p^*(\mu)$ at $\bar{\mu}_1(\Lambda)$ and $\bar{\mu}_2(\Lambda)$, which implies that the revenues are equal at the transition points between regions.

Recall that Lemma 5 shows that there exists $\bar{\mu}_1(\Lambda)$ and $\bar{\mu}_2(\Lambda)$ such that all potential customers join the queue at the optimal price, $p^*(\mu)$, for all $\mu \in [\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)]$ when $\Lambda < \bar{\lambda}_\alpha$. $\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$ and $\bar{\Lambda} > \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$ from Lemma 4. Therefore $A_1(\alpha) < \bar{\mu}_1(\Lambda) < \bar{\mu}_2(\Lambda) < A_2(\alpha)$ in the small market scenario.

Let Region A be $A_1(\alpha) < \mu \leq \bar{\mu}_1(\Lambda)$, Region B be $\bar{\mu}_1(\Lambda) < \Lambda \leq \bar{\mu}_2(\Lambda)$ and Region C be $\bar{\mu}_2(\Lambda) < \mu < A_2(\alpha)$. We will show that the optimal service rate is in Region B, for $\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$.

Note that in Region A and Region C the objective function is equivalent to that of the large market scenario ($R(\mu, p^*(\mu)) = \mu(K - \alpha\mu) - 2\sqrt{c\mu(K - \alpha\mu)} + c$), which is maximized at $\mu = \frac{K}{2\alpha}$ (Lemma 2).

$\bar{\mu}_2(\Lambda) = \max\{\mu | \lambda_e(\mu, p^*(\mu)) = \Lambda\}$, is greater than $\frac{K}{2\alpha}$ since $\lambda_e(\mu, p^*(\mu))$ is uni-modal by Lemma 1 and $\bar{\mu} = \operatorname{argmax}_{\{\mu \in \mathcal{F}(\alpha)\}} \{\lambda_e(\mu, p^*(\mu))\} > \frac{K}{2\alpha}$ by Lemma 4. Therefore the objective in equation 9 is decreasing in μ in Region C.

We show that the objective function is increasing in μ in Region 1, by showing that $\bar{\mu}_1(\Lambda) < \frac{K}{2\alpha}$. By definition $\lambda_e(\bar{\mu}_1(\Lambda), p^*(\bar{\mu}_1(\Lambda))) = \Lambda$. $\lambda_e(\mu, p^*(\mu))$ is uni-modal in μ from Lemma 1 and $\operatorname{argmax}_{\{\mu \in \mathcal{F}(\alpha)\}} \{\lambda_e(\mu, p^*(\mu))\} > \frac{K}{2\alpha}$ from Lemma 4. These facts imply that for $\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha} = \lambda_e(\frac{K}{2\alpha}, p^*(\frac{K}{2\alpha}))$, $\bar{\mu}_1(\Lambda) < \frac{K}{2\alpha}$.

As a result the service rate maximizing the objective function is in Region 2. Differentiating the objective function, $R(\mu, p^*(\mu))$, with respect to the service rate we get:

$$\frac{\delta R(\mu, p^*(\mu))}{\delta \mu} = \begin{cases} (K - 2\alpha\mu) \left[1 - \frac{c}{\sqrt{c\mu(K - \alpha\mu)}} \right] & \text{if } A_1(\alpha) < \mu < \bar{\mu}_1(\Lambda) \\ \Lambda \left(\frac{c}{(\mu - \Lambda)^2} - \alpha \right) & \text{if } \bar{\mu}_1(\Lambda) \leq \mu \leq \bar{\mu}_2(\Lambda) \\ (K - 2\alpha\mu) \left[1 - \frac{c}{\sqrt{c\mu(K - \alpha\mu)}} \right] & \text{if } \bar{\mu}_2(\Lambda) \leq \mu < A_2(\alpha). \end{cases} \quad (10)$$

The first order condition is given by:

$$FOC : 0 = \frac{\delta R(\mu, p^*(\mu))}{\delta \mu} = \Lambda \left(\frac{c}{(\mu - \Lambda)^2} - \alpha \right).$$

We show that the optimal service rate is an interior point of Region 2:

Recall that $\Lambda = \lambda_e(\mu, p^*(\mu)) = \mu - \sqrt{\frac{c\mu}{K-\alpha\mu}}$ at $\mu = \bar{\mu}_i(\Lambda)$ for $i = 1, 2$. Therefore we can write $\bar{\mu}_1(\Lambda) = \Lambda + \sqrt{\frac{c\bar{\mu}_1(\Lambda)}{K-\alpha\bar{\mu}_1(\Lambda)}}$. Plugging in this value into $\frac{\delta R(\mu, p^*(\mu))}{\delta\mu}$ we get:

$$\left. \frac{\delta R(\mu, p^*(\mu))}{\delta\mu} \right|_{\mu=\bar{\mu}_1(\Lambda)} = \Lambda \left[\frac{K - 2\alpha\bar{\mu}_1(\Lambda)}{\bar{\mu}_1(\Lambda)} \right].$$

The value of $\frac{\delta R(\mu, p^*(\mu))}{\delta\mu}$ is positive for $\bar{\mu}_1(\Lambda) \leq \frac{K}{2\alpha}$. $\bar{\mu}_1(\Lambda)$ is monotonically increasing in Λ for $0 \leq \Lambda \leq \bar{\Lambda}$ since $\lambda_e(\mu, p^*(\mu))$ is uni-modal (increasing and then decreasing) in μ (Lemma 1). Therefore the objective function is increasing in μ at $\mu = \bar{\mu}_1$, since $\bar{\mu}_1(\Lambda) \leq \frac{K}{2\alpha}$. This proves that the optimal service rate, μ^* , is greater than $\bar{\mu}_1(\Lambda)$ for low potential demand ($\Lambda < \frac{K-\sqrt{c\alpha}}{2\alpha}$).

Similarly, we show that $R(\mu, p^*(\mu))$ is decreasing in μ at $\mu = \bar{\mu}_2(\Lambda)$ by plugging in $\Lambda + \sqrt{\frac{c\bar{\mu}_2(\Lambda)}{K-\alpha\bar{\mu}_2(\Lambda)}}$ for $\bar{\mu}_2(\Lambda)$:

$$\left. \frac{\delta R(\mu, p^*(\mu))}{\delta\mu} \right|_{\mu=\bar{\mu}_2(\Lambda)} = \Lambda \left[\frac{K - 2\alpha\bar{\mu}_2(\Lambda)}{\bar{\mu}_2(\Lambda)} \right] < 0.$$

The above value is negative since $\bar{\mu}_2(\Lambda) > \frac{K}{2\alpha}$. Therefore an interior point of $[\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)]$ satisfies the first order condition for $\Lambda < \frac{K}{2\alpha} - \sqrt{c/\alpha}$:

The unique solution of the first order condition for $\mu \in [\bar{\mu}_1(\Lambda), \bar{\mu}_2(\Lambda)]$ is $\mu^* = \Lambda + \sqrt{c/\alpha}$, which proves the result of Lemma 6.

2. The optimal price, $p^*(\mu^*)$, is equal to $V_b + \alpha\mu_b - \alpha\Lambda - 2\sqrt{\alpha c}$, from Proposition 1.

3. The resulting equilibrium demand, $\lambda_e(\mu^*, p^*(\mu^*))$, is equal to the potential demand Λ by Proposition 1. ■

Proof of Proposition 2: Proposition 2 presents the optimal service rate and optimal price for all values of potential demand, Λ . Results follow from Lemmas 2 and 6. ■

Proof of Proposition 3: Proposition 3 shows that for $\Lambda \geq \frac{V_b + \alpha - 2\sqrt{c\alpha}}{\alpha}$, competing servers can achieve monopoly revenues, at using the optimal monopoly operating setting, μ^* and $p^*(\mu^*)$. Note that the potential demand Λ must be at least two times the optimal monopoly equilibrium demand, $\lambda_e(\mu^*, p^*(\mu^*))$.

To prove the proposition we begin by showing that for any potential demand $\Lambda \geq 0$, a server serving at rate μ and charging price p , is better-off in the single server setting, than in the multi-server competition setting. Recall that the revenue of the server in the single server setting is given by $R(\mu, p)$. Let $R_1(\mu_1, \mu_2, p)$ be the revenue of server 1, providing service with rate μ_1 , when server 2 is providing service with rate μ_2 , at price p in the two server competition case. $R(\mu, p) \geq R_1(\mu, \mu_2, p)$ for all $\mu, \mu_2 \in \mathcal{F}(\alpha)$ and $p \geq 0$.

$$R_1(\mu, \mu_2, p) = \begin{cases} p\Lambda & \text{if } V(\mu) - p - WC(\mu, \Lambda) \geq \max\{0, V(\mu_2) - p - WC(\mu_2, 0)\} \\ p\lambda_{1e}(\mu, \mu_2, p, \Lambda) & \text{if } WC(\mu, \Lambda) > V(\mu) - p \geq c/\mu \\ 0 & \text{if } V(\mu) - p - WC(\mu, 0) \leq \max\{0, V(\mu_2) - p - WC(\mu, \Lambda)\}. \end{cases} \quad (11)$$

where $\lambda_{1e}(\mu_1, \mu_2, p, \Lambda)$ is the equilibrium arrival rate for server 1.

Let examine the first line in equation (11). In the multi-server competition setting, server 1 can serve all potential customers if the net benefit of an arriving customer from joining server 1 when all other customers join server 1, is non-negative and greater than the net benefit of joining server 2 when no other customer joins server 2, i.e. $V(\mu) - p - WC(\mu, \Lambda) \geq \max\{0, V(\mu_2) - p - WC(\mu_2, 0)\}$. Recall that in the single server setting non-negativity of the net benefit ($V(\mu) - p - WC(\mu, \Lambda) \geq 0$) is sufficient to serve all potential customers. Clearly server 1 is better-off in a single server setting, than in a multi server setting when $V(\mu) - p - WC(\mu, \Lambda) \geq 0$.

Let us examine the case when $WC(\mu, \Lambda) > V(\mu) - p \geq c/\mu$ (line 2 in Equation (11)). In the single server setting, customers join the queue until the net benefit of joining equals zero. However in a multi-server setting, the net benefit of joining server 2 may be positive when $\lambda_e(\mu, p)$ customers join server 1 and $\Lambda - \lambda_e(\mu, p)$ customers join server 2. Therefore customers will deviate to server 2, until an equilibrium is reached, i.e. until the net benefit of joining server 1 equals the net benefit of joining server 2.

Hence, $R(\mu, p) \geq R_1(\mu, \mu_2, p)$ for all $\mu, \mu_2 \in \mathcal{F}(\alpha)$ and $p \geq 0$, which implies:

$$\max_{\{\mu \in \mathcal{F}(\alpha), p \geq 0\}} \{R(\mu, p)\} \geq \max_{\{\mu \in \mathcal{F}(\alpha), p \geq 0\}} \{R_1(\mu, \mu_2, p)\} \quad (12)$$

for all $\Lambda \geq 0$.

Therefore, in the multi server setting when $WC(\mu, \Lambda) > V(\mu) - p \geq c/\mu$, the equilibrium arrival rate for server 1, $\lambda_{1e}(\mu, \mu_2, p, \Lambda)$, is less than or equal to the monopoly equilibrium arrival rate, $\lambda_e(\mu, p) = \left(\mu - \frac{c}{V(\mu) - p}\right)$.

For $\Lambda \geq \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}$, the optimal equilibrium demand in the single server setting is given by $\lambda_e(\mu^*, p^*(\mu^*)) = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$.

In the two server setting the service provider can serve $2\lambda_e(\mu^*, p^*(\mu^*))$ customers, charging $p^*(\mu^*)$ when both servers serve at rate $\mu_i^* = \mu^*$, hence doubling the optimal monopoly revenues. In this case, the equilibrium demand at each server is equal to $\lambda_e(\mu^*, p^*(\mu^*))$. Clearly, the net benefit of an arriving customer from joining either server is equal to zero, therefore an arriving customer is indifferent between joining server 1, joining server 2 and not procuring the service. Hence this setting is an equilibrium for customers.

When the service provider charges price $p^*(\mu^*)$, (μ^*, μ^*) is a Nash Equilibrium for the servers, since they achieve the maximum revenue (shown in the LHS of the equation (12)) by choosing μ^* . (i.e. μ^* is the best response of a server, to any service rate strategy adopted by the other server).

Proof of Proposition 4: Proposition 4 indicates that for $\Lambda < \frac{V_b + \alpha - 2\sqrt{c\alpha}}{\alpha}$, there exists a symmetric Nash equilibrium (μ^e, μ^e) such that all potential customers procure the service and agents equally share the potential demand. Further, we have $\mu^e = \frac{\Lambda}{2} + \sqrt{c/\alpha}$.

We prove the proposition by showing that none of the players (the servers and customers) have incentive to deviate from μ^e . An arriving customer's net benefit from joining server i when half of the customers join server i is given by:

$$NB_i(\mu_i, p, \Lambda/2) = V(\mu_i) - p - \frac{c}{\mu_i - \Lambda/2}.$$

μ^e maximizes $NB_i(\mu_i, p, \Lambda/2)$: Differentiating $NB_i(\mu_i, p, \Lambda/2)$ with respect to μ_i we get:

$$\frac{\delta NB_i(\mu_i, p, \Lambda/2)}{\delta \mu_i} = -\alpha + \frac{c}{(\mu_i - \Lambda/2)^2}$$

The second order condition indicates that $NB_i(\mu_i, p, \Lambda/2)$ is concave in μ_i for $\mu_i \geq \Lambda/2$:

$$\frac{\delta^2 NB_i(\mu_i, p, \Lambda/2)}{\delta \mu_i^2} = \frac{-2c}{(\mu_i - \Lambda/2)^3} < 0.$$

The first order condition is satisfied at $\mu^e = \frac{\Lambda}{2} + \sqrt{c/\alpha}$, hence μ^e maximizes $NB(\mu_i, p, \Lambda/2)$.

The resulting net benefit is given by:

$$NB(\mu^e, p, \Lambda/2) = V_b + \alpha\mu_b - \alpha\Lambda/2 - 2\sqrt{c\alpha} - p.$$

Clearly the net benefit of an arriving customer from joining server i is non-negative for $p \leq V_b + \alpha\mu_b - \alpha\Lambda/2 - 2\sqrt{c\alpha}$, when agents serve at rate μ^e and customers mix equally between servers 1 and 2.

Server i , serving $\Lambda/2$ customers, has no incentive to deviate from μ^e , because deviating from μ^e will decrease the net benefit of an arriving customer from joining server i , which will result in lower equilibrium demand for the server i . Therefore (μ^e, μ^e) is a Nash equilibrium. Thus we prove that (μ^e, μ^e) is a Nash equilibrium for service agents for all $p \leq V_b + \alpha\mu_b - \alpha\Lambda/2 - 2\sqrt{c\alpha}$.

The maximum price that can be charged by the service provider is $p_2^* = V_b + \alpha\mu_b - \alpha\Lambda/2 - 2\sqrt{c\alpha}$. The service provider optimizes the revenues by charging p_2^* and fully extracting the

consumers' surplus. It can be seen that this result extends to n servers, and holds for all $\Lambda < n \frac{V_b + \alpha - 2\sqrt{c\alpha}}{2\alpha}$. ■

Proof of Corollary 1: The corollary suggests that a multi-server service provider can charge a higher price than a single server service provider when $\Lambda \leq \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}$.

Recall that the optimal price in the single server setting is given by:

$$p^*(\mu^*) = \begin{cases} V_b + \alpha\mu_b - \alpha\Lambda - 2\sqrt{c\alpha} & \text{if } \Lambda \leq \frac{V_b + \alpha\mu_b}{2\alpha} - \sqrt{c/\alpha} \\ \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2} & \text{if } \Lambda > \frac{V_b + \alpha\mu_b}{2\alpha} - \sqrt{c/\alpha}, \end{cases} \quad (13)$$

Note that the optimal price is non-increasing in the potential demand, Λ .

We prove the corollary for $\Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$ first.

In this case, the optimal price for a single server is $p_M = p^*(\mu^*) = V_b + \alpha\mu_b - \alpha\Lambda - 2\sqrt{c\alpha}$.

The optimal price in the two server setting is $p_2^* = V_b + \alpha\mu_b - \alpha\Lambda/2 - 2\sqrt{c\alpha}$.

Clearly the maximum price that can be charged by the service provider in the 2 server setting is higher than the optimal price in the single server setting: $p_2^* - p_M = \alpha\Lambda/2 > 0$.

For $\frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha} \leq \Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}$, the optimal price for the single server setting is $p_M = p^*(\mu^*) = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2}$ and the optimal price for the two server setting is $p_2^* = V_b + \alpha\mu_b - \alpha\Lambda/2 - 2\sqrt{c\alpha}$.

$$p_2^* - p_M = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2} - \alpha\frac{\Lambda}{2} > 0, \text{ since } \Lambda < \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{\alpha}.$$

Proof of Corollary 2: The corollary indicates that the optimal price is non-decreasing in the number of servers, n , and the marginal price increase approaches to zero as $n \rightarrow \infty$.

Recall that the optimal price is equal to $p^*(\mu^*) = \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2}$ for $\Lambda \geq n \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$.

To prove the corollary we first derive the equilibrium service rate for n agents and the resulting optimal price, when $\Lambda < n \frac{V_b + \alpha\mu_b - 2\sqrt{c\alpha}}{2\alpha}$.

We use the result of Proposition 4 to derive the symmetric Nash equilibrium. Agents maximize the net benefit $NB(\mu_i, p, \Lambda/n)$. Therefore the equilibrium service rate is $\mu^e(n) = \frac{\Lambda}{n} + \sqrt{c/\alpha}$. The resulting optimal price is given by:

$$p_n^* = V_b + \alpha\mu_b - \alpha\Lambda/n - 2\sqrt{c\alpha}.$$

Clearly p_n^* is increasing in the number of agents, n .

The marginal increase in price is given by:

$$p_{(n+1)}^* - p_n^* = \frac{\alpha\Lambda}{n(n+1)},$$

which approaches to zero as $n \rightarrow \infty$. ■