Submitted to
manuscript (Please, provide the manuscript number!)

Perils and Benefits of Free Trials in Large Scale Service Systems: An Operational Perspective

(Authors' names blinded for peer review)

Despite their pervasive use in practice, the operational implications of offering free trials is not well understood. Motivated by this, we consider a service firm that has the option of offering free trials to a new market while catering to an existing market of price and delay sensitive customers with price and capacity optimized prior to offering free trials. Then, we pose the following question: Under what conditions is offering free trials beneficial and what is the extent of this benefit? This question is relevant because with free trials the firm essentially gives up part of its capacity to potentially generate more revenue in the future by converting its free trial customers into paying customers. Hence, it is not apriori clear if offering free trials is beneficial. To answer this question, we pursue an asymptotic analysis that is relevant to large scale services. First, we consider the price and capacity optimization of the firm without free trials. Then, we incorporate free trials, and finally, we assume that the firm can adjust capacity as it offers free trials. Somewhat surprisingly, we find that offering free trials is beneficial only for small free trial market sizes with correspondingly small relative benefits that become negligible for large scale systems. On the other hand, offering free trials is always beneficial when capacity can be adjusted at relatively low cost. Thus, service firms should opt to simultaneously increase capacity to unlock the potential benefits of offering free trials, and otherwise not offer free trials at all.

Key words: service systems, free trial promotions, operations-marketing interface, capacity sizing, pricing, many-server queues, Halfin-Whitt (QED) regime.

---

1. Introduction

1.1. Motivation and Overview of Results

Free trials are pervasively used to promote high tech communication, information, and entertainment services to consumers nowadays. Many free trial promotions allow consumers to access and experience a specific service for a specific duration without charge. For example, the popular movie streaming service Netflix offers a 30-day free trial period to its customers while other streaming services such as Hulu or Vudu have similar promotions. After the free trial period, a portion of customers convert to paying regular customers. This portion is referred to as the trial-

to-adoption conversion rate or simply conversion rate. Thus, offering a free trial provides a service firm the opportunity to increase long term revenues by tapping into a new market of customers and eventually converting part of them into paying customers. Although appealing from a marketing perspective, the potential future benefits of free trials do not come for free and have a potential drawback for a service firm: By offering free trials, a service firm essentially gives up part of its capacity with the hope of generating future demand for long term benefits. Thus, the following questions emerges: *Is offering free trials really beneficial to a service provider when we also factor in its operational implications*?

In many ways, free trial promotions are the service equivalent of free sample promotions. Particularly, both free sample promotions and free trial promotions intend to generate new demand by letting customers experience a good or a service, respectively. Thus, free trial promotions and free sample promotions are similar from a marketing perspective. Despite this similarity, free trials have a distinguishing feature as they pertain to congestion-prone services: In service systems, the introduction of a free trial promotion requires that some of the capacity is used to serve demand that does not generate revenue. Hence, when a service firm offers free trials, it essentially gives away part of its capacity for free and operates at a lower effective capacity, which might decrease the rate at which revenue is generated. Furthermore, the lower effective capacity along with the increased demand might also increase the congestion in the system due to queueing effects. This, in turn, might discourage regular customers from using the service. Moreover, free trial customers who experience the same increased congestion might also be less likely to convert to paying customers. As a result, offering free trials can become a double-edged sword and it is not a priori clear if the service provider really benefits from offering free trials.

Motivated by this question, we consider a large scale congestion-prone service firm that considers offering a free trial service to a new market of customers in addition to its current customers, and we assume that the service firm has ex ante optimized its price and capacity. We aim to answer to the following related research questions:

- *What are the implications of offering free trials for a congestion-prone service firm? Specifically, under which conditions is offering free trials beneficial to a service provider, and what is the extent of such a benefit? What is the effect of the conversion rate and free trial market size?*
- *Under what conditions does a service firm that maximizes capacity and price operate prior to offering free trials?*
- *Under what conditions does a service firm operate with free trials assuming neither price nor capacity can be changed? That is to say, what is an impact of free trials on the system?*
- *Can the firm be better off by leveraging some form of control such as capacity adjustment? If yes, what are the prescriptions of the optimal controls?*

- *Under what conditions does a firm operate when capacity is controlled? Specifically, is it be possible to mitigate the impact of free trial services on congestion so that the system experiences a congestion as if there is no free trial service?*

It is worthwhile to note that there is limited research on free trials in the OR/OM literature and the operational effect of offering free trials is not well understood. On the other hand, recent studies in marketing literature focus on other aspects of free trials such as timing of free trials (see, for example, Foubert and Gijsbrechts (2016)) or customer retention (see, for example, Datta et al. (2015)). Our research methodology, which is driven from stochastic modeling and queueing theory, allows us to glean new and important insights, and contribute to the synergistic interface of marketing and operations management.

In order to answer the questions posed above, we first need to understand the operating conditions of the service system where the capacity and price are jointly optimized prior to offering free trials. Hence, we begin with an $M/M/C$ multi-server queue, also known as the Erlang-C queue, to model the system without free trials. We assume that the service system has a large capacity, and that this capacity is initially used to serve demand from regular customers who have heterogeneous valuations for the service (i.e., valuations that differ within regular customers), and are price and delay sensitive. We assume that that the system operates in equilibrium that results from an intrinsic feedback between demand and delay and that the service provider sets its price and capacity optimally to maximize her profit rate prior to offering free trials.

Next, we incorporate free trial customers on top of the regular customers as follows. Free trials are offered on an ongoing basis for an arbitrarily long planning horizin, but only to a new market of customers and for a given duration. Free trial customers are not familiar with the service (i.e., they either do not know their valuations and/or system characteristics such as the average delay). As a result, free trial customers opt to pay for service only after experiencing the service. After the free trial period, a certain portion of the free trial customers convert and become stochastically identical to regular paying customers. We still assume that the system operates in equilibrium that results from the intrinsic feedback between demand and delay.

We assume that the service firm cannot change its price (that was chosen optimally prior to offering free trials) as free trials are offered and we consider two cases depending on the form of control available: (a) **No Control**. The firm cannot change capacity and continues to operate under the same price and capacity that is optimized prior to offering free trials. Since the system manager has no control variable at her disposal, the no control case is essentially performance analysis of an ex ante service system with the addition of free trials (b) **Capacity Optimization**. The firm can only adjust its capacity as free trials are offered. Hence, the capacity optimization case is essentially an optimal design problem where we initially set the capacity and price optimally

and only capacity can be changed as free trials are offered. In both cases, our main goal is to understand the effect of free trials on the revenue rate as well as the quality of service as measured by the expected delay.

Before continuing to our results, we will elaborate on our solution methodology. We first note that conducting an exact analysis of the system with free trials or without free trials is difficult. This is because there exists no closed form expression for the equilibrium arrival rate for a given price and capacity. Thus, we can only analyze performance or optimize over control variables numerically. But, from such a direct numerical approach, we cannot glean much structural insights (with regards to optimal profit/revenue rate). Hence, we pursue an asymptotic analysis that is relevant for larger scale systems. It is this asymptotic approach that will allow us to derive structural insights (that are otherwise not available), which, in turn, will help us understand the effect of offering free trials.

Our main result under ex ante *joint* price and capacity optimization is that the quality-efficiency-driven (QED) regime arises as the optimal operating regime (see Theorem **??**). To our knowledge, this is the first result that establishes QED as the optimal regime in a *joint* pricing and capacity optimization problem, under mild assumptions that are fairly common in the revenue management and pricing literature. This approach allows us to shed light into the operating characteristics of large systems. Furthermore, it allows us to derive closed form scaling relations for the optimal price, capacity, and revenue rate as well as other performance measures of interest. Specifically, we find that, both the price and revenue rate admit a two part decomposition, which includes a first order term (which ignores the stochasticity) and a second order correction term (which accounts for stochasticity). This structural insight follows from an analysis that parallels Maglaras and Zeevi (2003).

Next, we consider the case of uncontrolled system with free trials. Our main result is that the optimal system operates under the ED (efficiency-driven) heavy traffic regime when the regular and free trial market size scale proportionally to the capacity. We derive fluid based scaling relationships for the revenue rate that highlight the impact of offering free trials on the capacity and delay, which also allows us to assess the benefit of free trials. Specifically, our results suggest that the revenue rate as well as the delay increases towards a limit as free trials are offered and more and more conversions occur. Somewhat controversially though, we find that offering free trials is as not beneficial when compared to the ex ante optimized system without free trials: For smaller free trial sizes, offering free trials offers relatively small benefit while larger free trial market sizes can yield significant losses.

Our final result concerns the case when the service provider is willing to invest in increasing her capacity. We establish that if the regular and free trial market sizes scale proportionally, then capacity optimization with free trials leads to an optimal system operating under the QED heavy

traffic regime. Therefore, the larger system experiences the smaller delays and higher utilization. Moreover, we examine the optimal first order profit rate as a function of the free trial market size and determine explicit conditions for offering free trials to be beneficial. Specifically, we observe that offering free trials is beneficial when capacity costs are low enough, whose threshold depends also on the valuation function.

Our key contributions are summarized in three-fold:

- **Actionable Insights into the Benefit of Offering Free Trials.** If an ex ante optimized service system offers free trials without any further control, our analysis reveals offering free trials is either not beneficial at all or it yields a small relative benefit. This relates to the aforementioned double-edged sword effect. On the other hand, when the capacity can be changed at a relatively low cost, it turns out that offering free trials is *always* beneficial. Hence, our analysis yields the important managerial insight that a service firm should either not offer free trials at all (if capacity cannot be changed at a low cost), otherwise the service firm should opt to increase its capacity along with free trials. Another important insight that emerges from our analysis is that any effort to increase conversion rates should be accompanied with an appropriate capacity investment. Otherwise, a higher conversion rate will not have a significant effect on the long run benefit of offering free trials. Conversely, it is not advisable to increase the free trial market size to the compensate for a small conversion rate as it can be detrimental to the service firm. Effect of Offering Free trials: Practical Implications

  — The question is can we convert free trials to paying throughput.

  — For a system with capacity and price optimized ex ante, the answer is only if the free trial market size is of order of the square root of the capacity

  — Even so, the benefit is of order of the square root of the capacity and thus limited

  — Trying to increase the conversion rate wouldn't improve the system much while increasing the free trial market size to compensate for small conversion rates can be detrimental.

  — Our solution is increasing capacity as you go.

- **Optimal Design of Service Systems with and without Free Trials.** There is a vast literature on the design of optimal queueing systems (see, for example, Stidham (2009)). We contribute to this literature by (1) modeling service systems without free trials and studying the joint price and capacity optimization, and (2) modeling service systems with free trials and studying the capacity optimization problem. Our asymptotic approach is relevant to large scale service systems, and allows us to provide explicit analytical expressions for the optimal price (or price and capacity) and the optimal revenue rate (or profit rate) and derive insight that is not possible via exact analysis.

• **Asymptotic Queueing Regimes for Large Scale Service Systems.** Under ex ante joint price and capacity optimization, we find that the optimal system operates under the QED heavy traffic regime. To our knowledge, this is the first result that establishes that QED is the optimal operating regime under joint price and capacity optimization. A practical implication of the QED regime is, prior to offering free trials, the service system enjoys high utilization levels while offering a good service quality in the form of negligible delays. When free trials are incorporated into this system without further control, we find that the system moves into the ED heavy traffic regime and that the now non-negligible delay is increasing in time. This degradation in quality-of-service can in practice result in lower than assumed conversion rates and reduced the revenue rates. On the other hand, when the service firm can adjust its capacity as free trials are offered, the system moves back into the QED regime. Thus, capacity optimization plays a dual role in that it also facilitates that the conversion process occurs as effectively as possible.

The remainder of our paper is organized as follows: We first review the relevant literature in Section 1.2. Then, in Section 2, we describe our model in detail. In Section EC.1, we assume that the system manager can only change the static per usage price, and study the revenue maximizing policy. We first assume that the system operates in the QED heavy traffic regime and establish scaling relations. Then, under our key elasticity assumption, we prove that the optimal system indeed operates in the QED heavy traffic regime in Section **??**. Next, we study the effect of offering free trials in Section **??** assuming the aforementioned elasticity condition. In Section EC.1.2, we consider what happens if the elasticity condition does not hold, and show that the optimal system operates under the QD heavy traffic regime. In Section EC.2, we consider the joint price and capacity optimization problem. We first construct an asymptotically optimal policy and show that QED is the optimal operating regime in Section **??**. Then, in Section **??**, we study the effect of offering free trials under joint optimization. In Section **??**, we extend our model to include heterogeneity across regular and free trial customer valuations. We make concluding remarks in Section **??**. All proofs are deferred to the Electronic Companion.

1.2. Literature Review

Our paper is related to four streams of literature which we elaborate on next.

Economics of queues: There is a large body of research that is pioneered by Naor (1969) and studies pricing to manage congestion externalities in queueing systems. In terms of the demand model assumed, our work belongs to the stream of literature that includes Mendelson (1985), Stidham (1985), Mendelson and Whang (1990), and Van Mieghem (2000). The common assumption in this stream of papers is that customers are price and delay sensitive but cannot observe the queue length. Hence the arrival rate is determined through an equilibrium. We refer the reader to Chapter 3 of Hassin and Haviv (2003) and Chapter 2 of Stidham (2009) for further details.

Many-server queues: Following a seminal paper by Halfin and Whitt (1981), there is also a vast literature of papers that study queueing systems with many servers. The standing assumption in this stream of papers is the celebrated square-root staffing rule which stipulates that the number of servers scales as $n + \beta\sqrt{n}$, where n is a proxy for system size. Such a scaling gives rise to Halfin-Whitt or QED heavy traffic regime where utilization approaches one while expected delay approaches zero. Thus, papers in this stream stipulate that the QED heavy traffic regime is the rational choice and assume the aforementioned scaling. For further details on many server queues and the QED heavy traffic regime, we refer the reader to the surveys by Gans et al. (2003), Dai and He (2012), Ward (2012), and van Leeuwaarden et al. (2017).

Economically optimal operating regime of queues: There is also a more recent literature where the QED Halfin-Whitt regime (for many servers), or conventional heavy traffic regime (for a fixed number of servers) arises endogenously as the optimal operating regime (rather than assuming a heavy traffic condition such as the square-root staffing rule). Related papers include Garnett et al. (2002), Maglaras and Zeevi (2003), Whitt (2003), Armony and Maglaras (2004a,b), Maglaras and Zeevi (2005), Plambeck and Ward (2006), Randhawa and Kumar (2008), Kumar and Randhawa (2010), Nair et al. (2016), and Maglaras et al. (2017). Among these papers, Whitt (2003) considers a model similar to ours where customers are sensitive to expected delay and shows that only the ED heavy traffic regime arises as the limiting operating regime. Our model is built upon Maglaras and Zeevi (2003) where it is shown that the QED regime is optimal under price optimization provided that an elasticity condition holds. We also find that a service system with free trials operates under QED when an analogous elasticity condition holds. Furthermore, we show that the QD regime arises as the optimal regime when the said condition does not hold, and that QED is the optimal regime under price and capacity optimization. In a recent paper, Maglaras et al. (2017) study the incentive compatible pricing decision of a service firm facing price and delay sensitive customers, and show that the QD and ED heavy traffic regimes arise endogenously as a result of price discrimination and service differentiation. Specifically, the high priority class operates under QD heavy traffic regime while the low priority class operates under the ED heavy traffic regime.

Free trials in service systems: The literature on free trials in service systems is rather new, and hence scarce. In the marketing literature, Foubert and Gijsbrechts (2016) study a firm offering free trials but focuses on the effect of evolving service quality. In our model, we consider a firm that already offers a service with steady quality to its regular paying customers. To our knowledge only two papers in the OR/OM literature study a similar problem, but via a non-asymptotic (single-server) approach: Zhou et al. (2014) and Lian et al. (2016) consider a service firm, modeled as an $M/M/1$ queue, which offers a service to a market of price and delay sensitive customers. Our model is significantly different than the mentioned two papers in several important ways: First, we

model the service system under consideration as a multi-server queue. This allows for a broader range of modelling capabilities where multiple customers can be served simultaneously or where capacity is shared amongst customer as in Maglaras and Zeevi (2003). In addition, our multi-server framework allows the service system to achieve statistical economies of scale and gives rise to an entirely different heavy traffic regime that is otherwise not possible (see also Section 3.3 in Maglaras and Zeevi (2003) for further discussion). Zhou et al. (2014) and Lian et al. (2016) also assume that all customers have the same valuation for service. Our model explicitly captures the heterogeneity in valuations which has several important implications on our results. Specifically, our elasticity assumption underpins the optimality of the QED regime asymptotically and allows us to provide tractable closed form expressions that yield insight that is otherwise not possible. Furthermore, we find that the elasticity and type of the demand distribution are also significant factors in determining whether free trials are beneficial.

2. Model Description

In this section, we provide our queueing model dynamics along with a customer choice model in order to establish equilibrium demand and customer's expected delay. We also describe economic objectives of the service provider.

2.1. Queueing Dynamics

Consider a service provider that operates under C units of capacity which corresponds to the number of processing resources available. The service provider uses these resources to serve its current customers, which we refer to as regular customers hereafter, and a new market of customers, which we refer to as free trial customers hereafter. We assume the regular customers arrive according to a Poisson process with rate Λ_1 per day, and that their service requirements follow an i.i.d. exponential distribution with mean $1/\mu$ for $\mu \in (0, \infty)$.

Next, we describe the free trial process in detail. We assume each free trial cycle lasts for a fixed duration of τ days and that the free trial service is offered on an ongoing basis for an arbitrarily long planning horizon of T days. Each day, the system faces arrivals from two types of customers: paying customers which is compromised of regular customers and free trial customers from previous cycles who converted to be paying customers, and the current free trial customers who are experiencing the service for free. Free trial customers have service requirements that are identical to regular customers and arrive into the system according an independent Poisson processes with rate Λ_2/τ per day. We let δ_t denote the conversion rate of the customers who start their free trial at day t. Hence, when the free trial of those customers expires at day $t+\tau$, a proportion δ_t of them converts into paying customers. Then, we let $\sigma_t := \frac{1}{\tau}\sum_{j=1}^{t-\tau}\delta_j$ denote the cumulative portion of customers

that converted up to time t, and we see that Λ_1 and $\Lambda_{2,t} := \Lambda_2 \sigma_t$ represent market size (i.e., the maximum potential arrival rate) of regular and converted free trial customers at day t, respectively. We assume that paying customers make entry decisions based on the customer choice model which is described next, and as a result, the effective arrival rate in equilibrium remains Poisson. Hence, the queueing dynamics both prior and after offering free trials is that of $M/M/C$ (Erlang-C) model with an endogenous arrival rate to be made specific. Finally, to ensure stability of the system, we assume that $\Lambda_2 < C$ and that within each day the system operates under its stationary distribution.

2.2. Customer Choice Model

We assume that customers are endowed with i.i.d. service valuations v (reservation price) for service, that are also independent of the arrival processes and service requirements. We assume that converted free trial customers have the same distribution as regular customers and that they become identical to regular customers once they experience the system and are familiar with the service offer. We let P denote the probability distribution of the valuation v and we assume that the cumulative distribution function F is an increasing generalized failure rate[1] (IGFR) distribution that is strictly increasing on its support, and that it has a continuous density f and a finite mean. Potential customers in free trial cycle t join the system only if their valuation is greater or equal to the expected value of the *full price* of service, which we describe next.

The full price of service in free trial cycle t includes the per usage fee for the service and the additive linear delay cost. The firm charges a fixed usage fee $\$p_t > 0$ for service. Current free trial customers do not pay this fee during the free trial promotion. The paying customers are also sensitive to delay and incur a cost $\$q > 0$ per unit time of delay. However, customers are not endowed with their actual delay at the time of arrival, and thus base their decision on average delay. In other words, we assume that the system operates as an unobservable queue.

2.3. Equilibrium Demand and Equilibrium Expected Delay

We assume that the system operates in the equilibrium steady state every day. Then, the equilibrium arrival rate into the system for any fixed service price $\$p > 0$ at day $t \geq \tau$ is

$$\lambda_t(p) := \left(\Lambda_1 + \frac{\Lambda_2}{\tau} \sum_{j=1}^{t-\tau} \delta_j \right) P(v \geq p + qE[D_t]) + \Lambda_2 = (\Lambda_1 + \Lambda_{2,t}) \bar{F}(p + qd_t) + \Lambda_2, \tag{1}$$

where $\bar{F}(\cdot) := 1 - F(\cdot)$, and $d_t := E[D_t] := E[D_t(p)]$ is the equilibrium expected delay in free trial cycle t. We interpret (1) as follows: Under the equilibrium expected delay $E[D_t]$, the full price

[1] We say that the user valuation v has increasing generalized failure rate (IGFR), or equivalently that the cumulative valuation distribution $F(x)$ is an IGFR distribution if the generalized failure rate $g(x) := xh(x)$ is an increasing function, where $h(x) := \frac{f(x)}{\bar{F}(x)}$ is the failure (hazard) rate. IGFR is a fairly common assumption in the revenue management/pricing literature and is satisfied by many distributions (see, for example, Lariviere (2006) and Ziya et al. (2004)).

of service becomes $p + qE[D_t]$, and thus a customer joins the system with probability $P(v \geq p + qE[D_t])$. Those customers who join and pay for service are either regular customers or free trial customers whose converted. Thus, the total arrival rate becomes that of paying customers, $\left(\Lambda_1 + \frac{\Lambda_2}{\tau}\sum_{j=1}^{t-\tau}\delta_j\right) P(v \geq p + qE[D_t])$, plus the current free trial customers, Λ_2.

Implicit in the equilibrium notion of (1) is an intrinsic feedback between delay and demand. Indeed, (1) implies that the equilibrium arrival rate and delay are *consistent* in the following way: The average steady state delay that arises in an Erlang-C queue where the arrival rate is equal to the equilibrium arrival rate, also yields the same equilibrium arrival rate under the assumed customer choice model. Our first result, which we present next, states that for all free trial cycles, the said equilibrium exists and is unique.

PROPOSITION 1. (SYSTEM EQUILIBRIUM) *Given a capacity $C > 0$ and price $p > 0$, there exists a unique equilibrium arrival rate and a corresponding eqilibrium expected delay for all $t = 0, \ldots, T$.*

2.4. Economic Objective of the Service Provider

We consider two related economic objectives in our paper: ex ante *joint* pricing and capacity optimization and capacity optimization, where the former corresponds to the initial system optimization prior to offering free trials. Our analysis has a dual purpose; in addition to determining the optimal operating policy , it also enables us to study the effect of offering free trials.

Ex ante joint pricing and capacity optimization: We assume the service system has already been optimized prior to offering free trials (i.e., at period $t = 0$). In other words, at $t = 0$, the system is operating under the optimal price $p_0^\star$ and the capacity $C_0^\star$ so as to maximize the profit rate $P_0(p, C)$, which is defined as

$$P_0(p, C) := p\lambda(p) - w\mu C = p\Lambda_1 \bar{F}(p + qE[D_0]) - w\mu C, \tag{2}$$

where $w > 0$ is the cost per capacity per day. Hence, the pair $(p_0^\star, C_0^\star)$ is the solution of the following optimization problem:

$$\max_{p>0, C>0} P_0(p, C). \tag{3}$$

Capacity optimization with free trials: We now assume the service provider offers free trials and the price is held fixed at $p_0^\star$, where $p_0^\star$ is the solution of the initial joint pricing and capacity optimization (5). Then, the sole decision variable at the disposal of the service provider is the system capacity. The objective of the service provider is to maximize the expected total revenue over the T days. Thus, we view the expected total profit rate as a function of the capacity vector $\mathbf{C}_T := [C_1, \ldots, C_T]$ and express it as

$$P_T(\mathbf{C}_T) = \sum_{t=1}^{T} [p_0^\star(\lambda_t(p_t) - \Lambda_2) - w\mu C_t] = \sum_{t=1}^{T} \left[(\Lambda_1 + \Lambda_{2,t})\, p_0^\star \bar{F}(p_0^\star + qE[D_t]) - w\mu C_t\right], \tag{4}$$

and thus, the objective of the service provider is

$$\max_{\mathbf{C}_T \in \mathbb{R}_+^T} P_T(\mathbf{C}_T). \tag{5}$$

Note that the capacity decision at day t does not impact the profit rate in the later days. Then, the objective function given in (4) decouples and the optimization problem in (5) becomes reduces to separately solving the following optimization problem for all $t = 1, \ldots, T$:

$$\max_{C_t > 0} (\Lambda_1 + \Lambda_{2,t})\, p_0^\star \bar{F}(p_0^\star + qE[D_t]) - w\mu C_t. \tag{6}$$

3. Model Analysis

We consider large scale systems in which their market sizes and capacities grow proportionally large. We first consider an initial system prior to offering free trials, where capacity and price are to be jointly optimized. Next, we add free trials on top but without any controls and examine how the optimal system operates. Finally, we study the system with capacity sizing optimization and delve into the same problem; how the system scales under optimal policy.

Motivation for Large Scale Analysis. Recall from Proposition 1 that there exists a unique equilibrium for a fixed price p and capacity C. However there exists no closed form expression for the equilibrium expected delay or arrival rate. Hence, the optimization problem in (3) can only be solved numerically. While we use this approach to compute exact optimal quantities in our numerical studies, it falls short of providing much insight into the operation of the system or the effect of offering free trials. This motivates us to take an approximate asymptotic approach and consider large scale systems. We describe our approach in the next section.

Scaling of Capacity and Market Sizes. Throughout our asymptotic analysis, we consider a sequence of systems indexed by a generic variable n (to appear as a superscript) with the understanding that as n grows large so do the market size(s) and the capacity. Specifically, under initial system optimization, we let the regular market size grow large with n. In this case, the optimal capacity is a decision variable and grows large naturally while the free trial market size is irrelevant. Then, we incorporate free trials (but without any further control) by letting the capacity, regular market size, and free trial market size grow large proportionally. Finally, under capacity optimization with free trials, we let the regular market size and free trial market size grow large proportionally (while the optimal capacity grows as the decision variable). We also note that the mean service requirement $1/\mu$, the valuation cdf $F(\cdot)$, the delay cost q, and the conversion rate δ_t do not scale with n and remain fixed throughout these sequences. As a result, individual customer characteristics remain the same while the aggregate market size(s) grow large. For all cases, we first present our asymptotic results, and then we translate these asymptotic results and interpret them in the context of what they imply for the original system.

3.1. Setting the Benchmark: Initial System Optimization (Optimality of the QED regime)

We assume that the firm has ex ante optimized the capacity and the static per usage price before offering a free trial service. That is, we first study the joint price and capacity optimization of a service system without free trials. It turns out that, for large scale systems, the optimal price and capacity are such that the optimal system operates under the so-called Quality-Efficiency-Driven (QED) heavy traffic regime where small average delays are accompanied by high utilization levels.

THEOREM 1. *Suppose the regular market size scales proportionally to the scaling factor n. Then, the optimal system operates in the QED regime and*

$$\lambda_0^{\star,n} := \lambda(C_0^{\star,n}, p_0^{\star,n}) = C_0^{\star,n}\mu - \beta^\star\sqrt{C_0^{\star,n}}\mu + o(\sqrt{n}) \quad \textit{and} \quad p_0^{\star,n} = \bar{p}_0(C_0^{\star,n}) + \frac{\pi^\star}{\sqrt{n}} + o(\frac{1}{\sqrt{n}})$$

$$P_0^{\star,n} = C_0^{\star,n}\bar{p}(C_0^{\star,n}) - \sqrt{C_0^{\star,n}}\mu\left(\bar{p}(C_0^{\star,n})\beta^\star - \pi^\star\right) - C_0^{\star,n}\mu w.$$

where $\bar{p}_0 := \bar{p}(C_0^\star)$ is such that $\Lambda_1\bar{F}(\bar{p}_0) = C_0^\star\mu$. Here, $\pi^\star$ and $\beta^\star := \beta(\pi^\star)$ are the optimal second order price/capacity corrections defined as

$$\pi^\star := \arg\min_{\pi\in\mathbb{R}}\{\bar{p}\beta\left(\pi\right) - \pi\}, \tag{7}$$

and for fixed $\pi \in \mathbb{R}$, $\beta(\pi)$ is the unique solution of

$$\frac{1}{q}\left(\frac{\bar{F}\left(\bar{p}\right)}{f\left(\bar{p}\right)}\beta - \pi\right) = \frac{\phi\left(\beta\right)}{\beta\left(\beta\Phi\left(\beta\right) + \phi\left(\beta\right)\right)}, \tag{8}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal probability density function and cumulative distribution function, respectively. The optimal profit scales as

Note that the QED heavy traffic regime *emerges* as the optimal regime (as opposed to being assumed). Consequently, we are able to derive insights with regards to the operations of optimally designed large scale service systems. Since we take such a system as our benchmark, this will also help us understand the effect of offering free trials. To do so, we have to first "translate" the asymptotic results in 1 and interpret them in the context of a a finite size system.

Practical Implications for Finite Size Systems. Theorem 1 allows us to shed light on the operation of optimally designed large scale service systems as follows: First, we see that large scale service systems operate under the QED heavy traffic regime and thus enjoy high utilization levels (high efficiency) accompanied by delays that are negligible, i.e., the equilibrium delay is a second order effect (high quality of service). Second, because delay is a second order effect, the optimal arrival rate, optimal price, and the optimal profit rate admit a two-part representation including a first order term (that ignores stochasticity) and a second order term (that accounts

for stochasticity). Then, we can interpret the asymptotic results in Theorem 1 in the context of a finite size system. Specifically, when the optimal capacity is $C_0^\star$, the optimal arrival rate scales as

$$\lambda_0^\star \approx C_0^\star \mu - \beta^\star \sqrt{C_0^\star}\mu \tag{9}$$

while the the optimal price scales as

$$p_0^\star(C) \approx \bar{p}_0(C_0^\star) + \frac{\pi_0^\star(C_0^\star)}{\sqrt{C_0^\star}}. \tag{10}$$

Furthermore, the optimal profit rate scales as

$$P_0^\star \approx C_0^\star \bar{p}(C_0^\star) - \sqrt{C_0^\star}\mu\left(\bar{p}(C_0^\star)\beta^\star - \pi^\star\right) - C_0^\star \mu w. \tag{11}$$

Remark 1. (An Asymptotically Optimal Capacity Control)

Based on Theorem 1, we can also construct an asymptotically optimal control by setting the capacity equal to

$$\hat{C}_0 := \arg\max_{C>0} \left(C\mu - \Lambda_2\right)\bar{p}\left(C\right) - w\mu C, \tag{12}$$

and the price equal to

$$\hat{p}_0 := \hat{\bar{p}}_0 + \hat{\pi}_0/\sqrt{\hat{C}_0}, \tag{13}$$

where $\hat{\bar{p}}_0 := \bar{p}_0(\hat{C}_0)$ and $\hat{\pi}_0$ is the minimizer of the second order profit correction term in square brackets in (7). Then, it follows from Theorem 1 and Theorem 3 in Maglaras and Zeevi (2003) that the pair $(\hat{C}_0, \hat{p}_0)$ is asymptotically optimal. Moreover, $\bar{p}_0(C_0^{\star,n}) \to \hat{\bar{p}}_0$ as $n \to \infty$. Hence, the optimal price $p_0^{\star,n}$ also converges to $\hat{\bar{p}}_0$ which only depends on the valuation distribution and the linear staffing cost.

3.2. Uncontrolled System (Emergence of the ED regime)

We assume that once the firm starts to offer free trial promotions, the firm cannot change neither capacity nor price (no control), or it can change the capacity at the beginning of each free trial cycle to maximize the total profit (capacity optimization). In this section, we examine the former situation.

We start our asymptotic analysis by considering a sequence of systems where the capacity, regular market size, and free trial market size grow large proportionally. Specifically, we assume that the capacity, the regular market size, and the free trial market size increase proportionally according to

$$\begin{aligned} C^n &:= n, \\ \Lambda_1^n := \kappa_1 \mu n, \ &\text{and} \ \ \Lambda_2^n := \kappa_2 \mu n, \end{aligned} \tag{14}$$

respectively, for $\kappa_1 := \frac{\Lambda_1}{C_0^\star \mu} > 0$ and $0 \leq \kappa_2 := \frac{\Lambda_2}{C_0^\star \mu} < 1$. Then, we see that the market size of converted free trial customers scales as

$$\Lambda_{2,t}^n := \Lambda_2^n \sum_{j=1}^{t-1} \delta_j = \sigma_t \kappa_2 \mu n, \tag{15}$$

for all $t \geq 1$. Note that the original system is indeed embedded in this sequence and can be recovered by setting $n = C_0^\star$. Again, we only scale the regular and free trial market sizes along with capacity, but not the individual customer characteristics. Then, $\bar{p}_t$, the price that matches the capacity to demand in the absence of delay, is constant throughout this sequence for fixed t and given by

$$(\Lambda_1^n + \Lambda_2^n \sigma_t)\, \bar{F}(\bar{p}_t) + \Lambda_2^n = C^n \mu, \tag{16}$$

or equivalently via

$$\bar{p}_t = \bar{F}^{-1}\left(\frac{C^n\mu - \Lambda_2^n}{\Lambda_1^n + \Lambda_2^n \sigma_t}\right) = \bar{F}^{-1}\left(\frac{C\mu - \Lambda_2}{\Lambda_1 + \Lambda_2 \sigma_t}\right) = \bar{F}^{-1}\left(\frac{1-\kappa_2}{\kappa_1 + \kappa_2 \sigma_t}\right). \tag{17}$$

Similarly, we let $\bar{p}_0$ be defined analogously for the system prior to offering free trials, i.e., for a system with $\Lambda_2 = 0$ as

$$\bar{p}_0 = \bar{F}^{-1}\left(\frac{C^n\mu}{\Lambda_1^n}\right) = \bar{F}^{-1}\left(\frac{C\mu}{\Lambda_1}\right) = \bar{F}^{-1}\left(\kappa_1^{-1}\right). \tag{18}$$

Using the definition in (17), we see that $\bar{p}_t$ is increasing in Λ_1 and Λ_2, and that $\bar{p}_t > \bar{p}_0$ for all $t \geq 1$. The constant price $\bar{p}_t$ will indeed play a pivotal role in our asymptotic analysis in what follows. Particularly, the fact that $\bar{p}_t > \bar{p}_0$ suggests that it is desirable for the system manager to increase the price in response to the increased demand due to the addition of free trials. However, the price cannot be changed and must remain fixed at $p_0^\star$, the optimal price that is set prior to offering free trials. Consequently, the desired increase in the full price of service can only be achieved with an increase in the expected delay. Our main result of this section formalizes this intuition.

Theorem 2. (Emergence of ED with Free Trials) *Suppose regular market size and free trial market scale proportionally to the ex ante optimal capacity. Then, with free trials the system moves into the Efficiency-Driven (ED) heavy traffic regime:*

$$d_t^n = \bar{d}_t + \frac{1}{(\Lambda_1^n + \Lambda_2^n \sigma_t)\bar{d}_t f(\bar{p}_t + q\bar{d}_t)} + o(1/n)$$

where $\bar{d}_t := \frac{\bar{p}_t - \bar{p}_0}{q}$. *Furthermore, optimal revenue scales as*

$$R_t^n(C_0^{\star,n}, p_0^{\star,n}) = p_0^{\star,n}\left[(C_0^{\star,n}\mu - \Lambda_2^n) - 1/\bar{d}_t\right] + o(1).$$

Remark 2. (Intuition and Relation to Earlier Literature) The intuition behind Theorem...

Practical Implications for Finite Size Systems. Once free trials are incorporated, we see from Theorem 2 that the system operates under the Efficiency-Driven heavy traffic regime where the delay is no longer negligible. For large systems, we can approximate the delay in day t by the first order delay as $d_t^n \approx \bar{d}_t$. Similarly, can approximate the revenue in day t by the first order revenue as $R_t(C_0^\star, p_0^\star) \approx p_0^\star \left[(C_0^\star \mu - \Lambda_2) - 1/\bar{d}_t\right]$. Moreover, we expect the delay to increase in time and converge to an upper bound since the first order delay $\bar{d}_t$ is increasing in t. On the other hand, we expect the revenue to increase in time and converge to an upper bound. To verify these insights we plot the exact revenue against the first order fluid approximation in Figure 1.

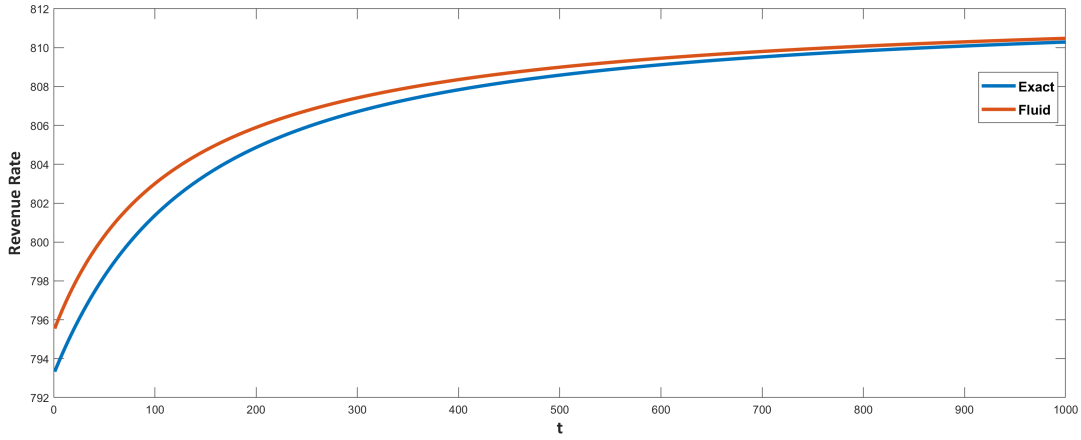

Figure 1 % **Effect of Offering Free Trials under Free Trial Market Rates**

In Figure 1, we assume that the valuations follow a uniform distribution with $v \sim U[0,4]$. We also set the initially optimized capacity at $C_0^\star = 400$, the regular customer market size at $\Lambda_1 = 825$ per day, and the free trial market size at $\Lambda_2 = 20$ per day. Furthermore, we assume that the free trial duration is $\tau = 30$ days and that $\delta_t = 1$ (i.e., all free trial customers convert to regular customers) and we let $\mu = q = 1$. We make two important observations from Figure 1: First, we see that the first order approximation performs remarkably well and closely captures the behavior of the exact revenue, thanks to the $o(1)$ optimality established in Theorem 2. Second, we see that revenue rate increases towards an upper bound also as suggested by the first order fluid approximation.

Effect of Offering Free Trials (What to Expect). Having verified the performance of the first order approximation, we now turn to the focal point of this section, namely the effect of offering free trials. To do so, we use the ex ante optimized system as our benchmark and assess the benefit of offering free trials by comparing the revenue rate of after offering free trials to that of the ex

ante optimized system. To start, we recall from the analysis in Section 3.1 that the revenue rate *prior* to offering free trials can be expresses as

$$R_0^\star \approx p_0^\star \left[C_0^\star \mu - \beta^\star \sqrt{C_0^\star \mu} \right]. \tag{19}$$

Note that the term in the parentheses on the right-hand-side of (19) is the equilibrium arrival rate which is the rate revenue is generated. This rate is smaller than the system capacity $C_0^\star \mu$ by a term that is proportional to the square root of system capacity $\sqrt{C_0^\star \mu}$ which is due to the inherent stochasticity of the system. After incorporating free trials, the revenue rate in day t becomes

$$R_t(C_0^\star, p_0^\star) \approx p_0^\star \left[(C_0^\star \mu - \Lambda_2) - 1/\bar{d}_t \right]. \tag{20}$$

From (19), we see that the primary effect of incorporating free trials to an ex ante optimized system is decrease in the effective capacity by a magnitude that is equal to the free trial market size Λ_2. Furthermore, we see that there is a secondary effect due to the now non-negligible delay that fades away in time as free trials are offered.

The next step in understanding the effect of offering free trials is to compare the revenue rate with free trials given in (20) to the benchmark revenue rate in (19). Consequently, we see that the relevant trade-off is between the decrease in effective capacity and the gain in additional throughput, which, in turn, depends on the relative magnitude of the free trial market size and the square root of the ex ante optimal capacity as follows:

Case 1 Small Free Trial Market Sizes: When magnitude of the free trial market size Λ_2 is small compared to the $\sqrt{C_0^\star}$, the square root of the ex ante optimal capacity, offering free trials can indeed yield a benefit. However, it follows from (19) that the alluded benefit can be at most be of order of $\sqrt{C_0^\star}$. Consequently, the relative benefit of offering free trials will be of order of $\sqrt{C_0^\star}$, and thus negligible in larger systems.

Case 2 Large Free Trial Market Sizes: When the magnitude of the free trial market size Λ_2 is larger than the $\sqrt{C_0^\star}$, offering free trials will in fact decrease the revenue rate thereby hurting the company. This is because the loss in revenue due to the decrease in the effective capacity dominates the gain from the increased market size.

In summary, we find that it is the relative magnitude of the free trial market size that determines whether the firm is better off offering a free trial service. This happens when the magnitude of free trial market size is small compared to the square root of the ex ante optimal capacity in which case the relative benefit of offering free trials will be small. Hence, we conclude that a service firm cannot expect to incur substantial benefits without further action. It is also interesting to note that the conversion rate does not play a significant role in this which is counter to the emphasis on

improving conversion rates in the marketing literature (see, for example, Foubert and Gijsbrechts (2016))

Effect of Offering Free Trials (What Really Happens). In this section we numerically verify that our asymptotic intuition for the benefit of free trials is indeed accurate for large systems of finite size. To do so, we plot percentage change in exact revenue rate (with respect to the revenue rate in our benchmark system without free trials) as a function of time by varying free trial market sizes in Figure 3.2 and by varying conversion rates in Figure 3.2. We set all other parameters as in Figure 1 and also start plotting revenue rates once conversions start to happen at $t = \tau$.

In Figure 3.2, we plot the exact percentage change in exact revenue for three different free trial market sizes: $\Lambda_2 = 20$, $\Lambda_2 = 25$, and $\Lambda_2 = 30$, respectively. Recall that the ex ante optimal capacity which is kept fixed is $C_0^\star = 400$. Then, our insights based on asymptotic analysis suggest that offering free trials should be beneficial only for small free trial market sizes that are of order of $\Lambda_2 = \sqrt{400} = 20$. Indeed, it is only the case with $\Lambda_2 = 20$ among the three that yields some long-run benefit; but as expected, the relative benefit is not substantial (around 0.5%). For the two other free trial market sizes that are only slightly higher, the service firm is in fact worse off offering free trials. Interestingly, and in line with our asymptotic results, we see that a smaller free trial market size is preferable.

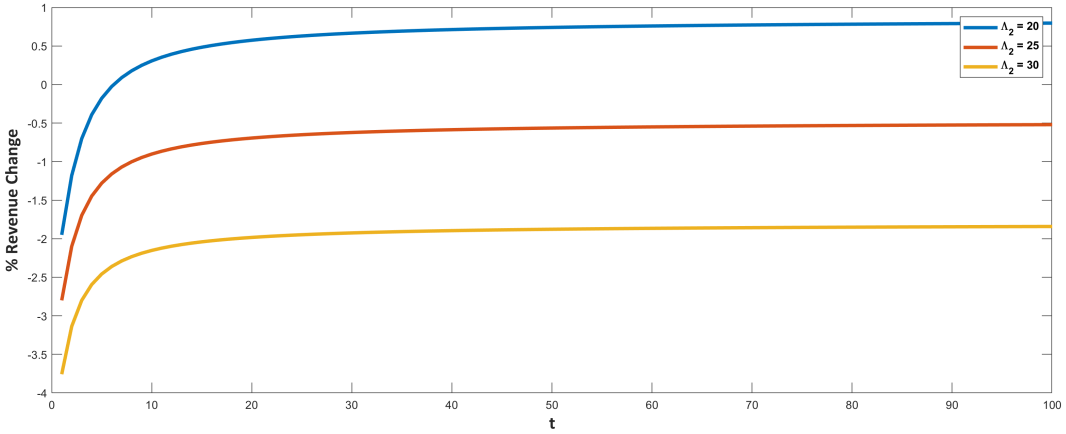

Figure 2 % **Effect of Offering Free Trials under Free Trial Market Rates**

Next, we examine the effect of conversion rates on the benefit of offering free trials in Figure 3.2, where we plot the exact percentage change in exact revenue for three different conversion rates: $\delta_t = 0.50$, $\delta_t = 0.75$, and $\delta_t = 1.00$, respectively. Our asymptotic analysis suggests that the conversion rate does not play a significant role in determining the long run benefit of offering free

trials. Figure 3.2 corroborates this insight: We see that the long run revenue rate is converging towards the same rate. On the other hand, a higher conversion rate is still preferable in the short run as it yields a higher the revenue rate.

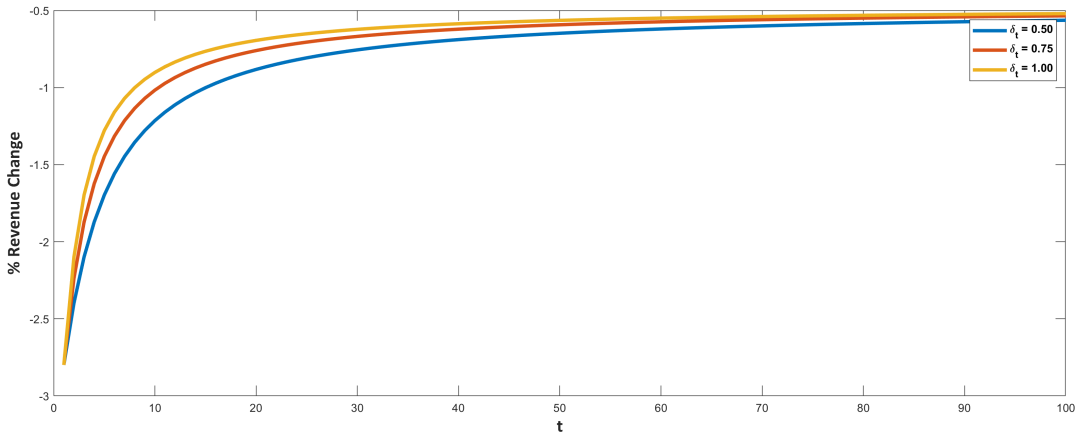

Figure 3 % **Effect of Offering Free Trials under Free Trial Market Rates**

Discussion Our analytical and numerical results shed light onto the operational effect of offering free trials and yield the following new and important insights: We find that a service firm that ex ante optimizes its capacity and price and offers free trials without any further control will not be able to enjoy the expected benefits of free trials. Specifically the firm will be either worse off offering free trials; or, when the firms is better off, the associated benefits will be small and thus negligible in larger systems. This also suggests, all else being the same, that a service firm will prefer a smaller free trial market size which sharply contradicts the marketing intuition that a larger market size can make up for smaller conversion rates (see, for example, Kumar (2014)). Another surprising result we get is that larger conversion rates do not have a significant impact on the revenues in the long run. This suggests that the emphasis on achieving higher conversion rates in the marketing literature (see, for example, Foubert and Gijsbrechts (2016)) is not necessarily warranted and that the service firm should opt for a different solution to unlock the expected benefits of free trials. This is the focus of our next section.

4. Capacity Optimization with Free Trials: Unlocking the Benefits of Free Trials

The major takeaway from the analysis in Section is that free trials cannot be converted into throughput which is the primary reason why the expected benefits of an increased market size are not realized. Another potential issue is that the quality of service as measured by the expected delay deteriorates as free trials are offered. Hence, it might be desirable to mitigate this effect. One

potential solution to is increasing the system capacity on the go as free trials are offered. This leads us to consider the capacity optimization problem with free trials where the firm can optimally set the capacity at the beginning of each day (while keeping the price fixed), to maximize the log run average profit per day (recall from Section that the problem decouples and that maximizing the average profit per day is equivalent to maximizing profit for a given planning horizon).

We devise an asymptotically optimal capacity sizing policy which shows that offering free trials is always beneficial for large scale systems provided that the linear capacity cost is below a certain threshold; see the condition (22) below.

Our first step is to solve for the first-order approximation of the optimal profit given by

$$\max_{C\geq 0} \hat{P}(C) := (C\mu - \Lambda_2)\,\bar{p}\,(C) - w\mu C,$$

and let $\hat{C}$ be the corresponding maximizer. Also, set the first order price term $\hat{\bar{p}} := \bar{p}(\hat{C})$. This is justified because $\hat{\bar{p}} := \bar{p}(\hat{C}) \approx \bar{p}(C^\star)$, and we can show the system still operates under QED with capacity $\hat{C}$.

Indexing. Our first and main result of this section establishes that increasing the capacity optimally puts the optimally designed system back into the QED heavy traffic regime.

THEOREM 3. (OPTIMALITY OF QED) *Suppose the regular and free market sizes scale proportionally. Then, the first order optimal capacity $\hat{C}^n$ is asymptotically optimal, i.e,*

$$\frac{P^n(\hat{C}^n, p_0^{\star,n})}{P^n(C^{\star,n}, p_0^{\star,n})} \to 1 \ \textit{as}\ n\to\infty,$$

and, under the exact optimal capacity $C^{\star,n}$, the expected delay is negligible for large scale systems, i.e.,

$$d_t^{\star,n} \to 0 \ \textit{as}\ n\to\infty.$$

Furthermore, the QED heavy traffic regime is optimal in the sense that the optimal expected delay $d_t^{\star,n}$ scales as

$$\sqrt{n} d_t^{\star,n} \to d_t^\star \in (0,\infty) \ \textit{as}\ n\to\infty.$$

Note that Theorem 3 is important on its own as it shows that the optimal system operates in the QED regime. As a consequence we see that expected delays are negligible gain, and that the impact of offering free trials on the QoS is mitigated. On the other hand, Theorem 3 does not provide an explicit expression for the optimal profit or capacity. This is important because we would like to understand the effect of offering free trials when capacity is optimized. To do so, we will first derive an asymptotically optimal capacity which will lead to an approximate yet tractable expression for

the optimal profit in large systems. Then, we will use the aforementioned expression to shed light onto the effect of offering free trials.

Our approach to devise an asymptotically optimal capacity is as follows: First we optimize the first order approximation of the optimal profit under joint capacity and price optimization with free trials. Specifically, we solve

$$\max_{C_t^n \geq 0} \hat{P}_t(C_t^n) := (C_t^n \mu - \Lambda_2)\, \bar{p}_t\left(C^n\right) - w\mu C_t^n = \left(\Lambda_1^n + \Lambda_2^n\right) \left[\bar{p}_t\left(C^n\right) \bar{F}\left(\bar{p}_t\left(C^n\right)\right) - w\left(\bar{F}\left(\bar{p}_t\left(C^n\right)\right) - \kappa_2\right)\right],$$

and let $\hat{C}$ be the corresponding maximizer. We also set the first order price term $\hat{\bar{p}} := \bar{p}_t(\hat{C}_t)$, which only depends on the valuation distribution F and capacity cost w.

Proposition 2. (Characterization of the Optimal First Order Price and Capacity) *There exists a unique $\hat{C}_t$ that solves* (EC.26) *and a corresponding $\hat{\bar{p}}_t := \bar{p}_t(\hat{C}_t)$ which can be further characterized as follows:*
(i) If $v_{min} - \frac{\bar{F}(v_{min})}{f(v_{min})} < w$, then $\hat{\bar{p}}_t \in (v_{min}, v_{max})$ and solves the first order condition given in (EC.28) or equivalently in (EC.29).
(ii) If $v_{min} - \frac{\bar{F}(v_{min})}{f(v_{min})} \geq w$, then $\hat{\bar{p}}_t = v_{min}$.

It follows that the sequence of optimal capacity and price pairs yields a scaled limiting profit of $\hat{\bar{p}}\bar{F}(\hat{\bar{p}}) - w(\bar{F}(\hat{\bar{p}}) + \kappa_2)$, which is identical to that of the proposed capacity $\hat{C}$ along with an appropriately set price is asymptotically optimal for the joint problem as it achieves the same limiting profit as the exact optimal policy.

Then, the question is if the proposed capacity $\hat{C}$ can achieve the same limiting scaled profit under capacity optimization where the price is set to $p_0^{\star,n}$. Since the optimal profit under capacity optimization is bounded by the optimal profit under joint capacity and price optimization, this would imply that the proposed capacity $\hat{C}$ is also asymptotically optimal under capacity optimization. It turns out this is indeed the case and that

$$\frac{P_t^n(\hat{C}_t^n, p_0^{\star,n})}{\Lambda_1^n + \sigma_t \Lambda_2^n} \to \hat{\bar{p}}\bar{F}(\hat{\bar{p}}) - w(\bar{F}(\hat{\bar{p}}) + \kappa_2),$$

which leads us to our next result.

Proposition 3. *Suppose the regular and free market sizes scale proportionally. Then, the first order optimal capacity $\hat{C}^n$ is asymptotically optimal, i.e,*

$$\frac{P^n(\hat{C}_t^n, p_0^{\star,n})}{P^n(C_t^{\star,n}, p_0^{\star,n})} \to 1 \text{ as } n \to \infty,$$

It follows from Proposition 3 that

$$\frac{P_t^n(C_t^{\star,n}, p_0^{\star,n})}{\Lambda_1^n + \sigma_t \Lambda_2^n} \to \hat{\bar{p}}\bar{F}(\hat{\bar{p}}) - w(\bar{F}(\hat{\bar{p}}) + \kappa_2). \tag{21}$$

The importance of (21) is that it enables us to approximate the optimal profit which in turn provides a condition under which a large scale service firm is better off by offering the free trials. We present the said approximation along with the condition it yields next.

PROPOSITION 4. *Suppose the regular and free market sizes scale proportionally. Then,*

$$P(C^\star) \approx P(\hat{C}) = \Lambda_1 \left[\hat{\bar{p}}\bar{F}\left(\hat{\bar{p}}\right) - w\right] + \Lambda_2 \left[(\hat{\bar{p}} - w)\sigma_t \bar{F}\left(\hat{\bar{p}}\right) - w\right]$$

Therefore, for large scale systems, offering free trials is beneficial in day t if and only if

$$w < (\hat{\bar{p}} - w)\sigma_t \bar{F}\left(\hat{\bar{p}}\right). \tag{22}$$

It is worthwhile to note that whether offering free trials is beneficial in day t depends *only* the model parameters F, σ_t, and w. It is also clear that once condition (22) holds in day t, it continues to hold since σ_t is increasing in t. Hence, it is the valuation distribution, capacity cost, and conversion rates that determine when free trials start to become beneficial and remains so thereafter. Furthermore, the condition (22) can be explicitly determined. For example, we can assume a free trial period of 30 days and set $\sigma_t = 1$ to determine the condition for free trials to be beneficial 2 months into offering free trials (recall the first 30 days will not generate any revenue yet). Then under $U[0,\theta]$ valuations, we require $w < \theta/(3+2\sqrt{2})$ with $\sigma_t = 1$. It is also interesting to notice that the earned benefit is (approximately) linear in t if δ_t is constant, i.e., when the conversion rate does not change from day to day.

Numerical Example. Next, we illustrate our asymptotic insights for a finite size system numerically. We let the regular market size $\Lambda_1 = 900$ and assume that customer valuations follow a uniform distribution where $v \sim U[0,4]$. The service rate μ, customer delay cost q, and conversion rate δ_t is set to 1. Given these parameters, initial system optimization yields and optimal capacity of $C_0^\star = 400$ when the capacity cost is $w = 0.34$, which will be the fixed capacity that we assume for this study. Then, our insights based on asymptotic analysis suggest that offering free trials should not be beneficial for free trial market sizes that are significantly larger than $\Lambda_2 = \sqrt{400} = 20$. Hence, we plot the percentage change in revenue rate with respect to the benchmark case $\Lambda_2 = 0$ as a function of time for three different Λ_2 values in Figure 3.2. Figure 4 illustrates the said point.

Discussion of Some Key Modelling Choices. We end this section with a discussion of our modelling choices and possible extensions.

- Quality of Service Dependent Conversion Rates: It is reasonable that customers make their conversion decisions based on the quality of service as measured by the expected delay experienced during the free trial. Particularly, it is possible that the conversion rate of a customer starting her free trial in period t is in fact a function that depends on all the expected delays experienced in an

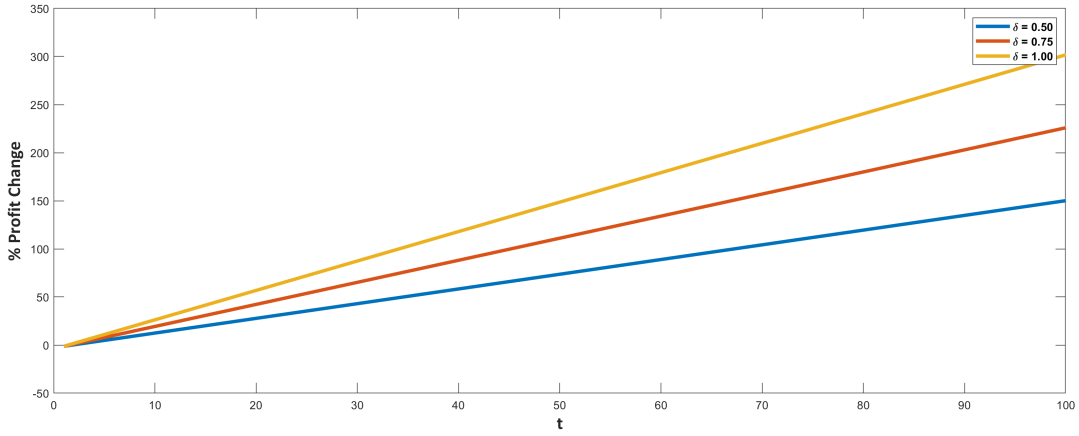

Figure 4 % **Effect of Offering Free Trials under Free Trial Market Rates**

arbitrary way and given by $\delta_t(\mathbf{d_t})$, where $\mathbf{d_t} := [d_t, d_{t+1}, \cdots, d_{t+\tau}]$. It is also reasonable that $\delta(\cdot)$ is a decreasing function of its arguments. While a thorough analysis of this extension is cumbersome, we believe that the essence of our results extend to this setting. To see why, first notice that our analysis thus far correspond to the ideal case with a conversion rate of $\delta_t(\mathbf{0})$. However, we know that in reality the system with free trials but no control operates in the ED heavy traffic regime with non-negligible delays. This implies that the conversion rates are even lower and thus the benefit of offering free trials is even further limited. On the other hand, when capacity is optimized as free trials are offered, the systems moves back to the QED heavy traffic regime where delays are negligible and conversions are close to the assumed ideal level. Thus, optimizing capacity plays a dual role with QoS dependent conversions: It relaxes the constraint on the capacity while simultaneous making sure that delays are kept at bay so that conversions are occurring as effectively as possible.

- Heterogeneous Valuation Distributions: It is also plausible that free trial customers have a different valuation distribution. Specifically, the free trial customers might have a valuation distribution that is stochastically smaller than the regular customers. Again, we believe that such an extension does not change the main takeaway of our paper. In the case of free trials with no control, we would expect a lower arrival rate from converted free trial customers resulting an even smaller gain from offering free trials.
- Stationarity: We have assumed that the service company is able to attract the same free trial market size from which a certain percentage converts. It is also possible that the free trial market size is time-dependent. As long as the service firm can dynamically adjust capacity, our main results continue to hold.

5. Conclusions

Despite the pervasive practice of free trial services in high tech information, retail, and entertainment services, there is limited research as to its operational impact on improving firms' profits. This paper offers a stochastic model and examines whether offering free trials is really beneficial to the service providers. Our analysis reveals the following intriguing structural and actionable insights that are relevant to large scale service systems:

1. Prior to offering free trials service, when the service provider ex ante optimizes its price and capacity, it is always optimal to choose the capacity and price such that the system operates in the QED heavy traffic regime. This implies the initial optimized system experiences small expected delays while enjoys a high utilization with ample capacity of order of square root of service demand.

2. When the regular market size and free trial market size scale proportionally, free trials decrease the revenue rate and are not beneficial. On the other hand, when the free trial market scales as the order of square root of the regular market size, free trials do increase the revenue rate and are beneficial. However, it can only be beneficial up to the square root order of the ex ante optimal capacity and as such the benefit is relatively limited in large scale systems.

3. The aforementioned benefit will even be further limited if conversion rates are QoS (quality of service) dependent, and we see that increasing conversion rates doesn't benefit the system much. Therefore, it is advisable that the service firm should opt to increase capacity to turn the converted customers into throughput and to fully unlock the benefits of free trials.

There are several interesting directions for future research. First, we would like to point out that we assumed that the parameters of our problem are fully known to the service provider. Thus, an interesting extension is to model the Bayesian learning problem where the service provider either learns consumer characteristics as in Afèche and Ata (2013) or the market size as in Araman and Caldentey (2009). Second, under our model, both paying customers and current free trial customers receive the same level of service. To optimally design the system so as to attain maximum profit rate, the system manager could consider differentiating service levels such that the paying customers receive a higher service level as in Maglaras and Zeevi (2005). Note also that we only consider the case where customers pay a price each time they use the service (i.e., pay-as-you-go), but not the case where customers pay a membership or subscription fee independent of their actual usage. Hence, an interesting future research topic is to take an approach similar to Randhawa and Kumar (2008) and to refine our asymptotic analysis by incorporating the subscription and the pay-as-you-go pricing schemes in the face of free trial markets, and to compare their operational benefits. Finally, we note that there is an emerging business model that is similar to free trials: In the so-called *freemium* business model, customers can access a basic version of a service for free and indefinitely while a premium version of the service with enhanced features can be accessed

by paying a fee (see, for example, Mishra et al. (2018)). A possible future research direction is to compare the operational benefits of free trials and freemium services, and to determine conditions under which one is preferred over the other.

6. Appendix: Proofs of Main Results

In what follows, the notation $f^n = O(g^n)$ means that $\limsup_{n\to\infty} f^n/g^n < \infty$ and the notation $f^n = o(g^n)$ means that $\lim_{n\to\infty} f^n/g^n = 0$.

Proof of Proposition 1: In this proof only, we let the superscript e denote an equilibrium quantity for a given price. Given $\xi > 0$, we also let $E[D(\xi)]$ denote the expected delay in an $M/M/C$ queue facing the arrival rate $\lambda(p+q\xi)$. Then, the equilibrium delay and equilibrium arrival rate satisfy

$$\lambda(p+q\xi^e) = (\Lambda_1+\Lambda_2)\,P(v > p+q\xi^e) + \Lambda_2 \text{ and } \xi^e = E[D(\xi^e)].$$

Define $h(\xi) := \xi - E[D(\xi)]$ and observe that $h(0) < 0$ and $h(\infty) > 0$. Also, note that $E[D]$ is a continuously differentiable function of the arrival rate λ and the arrival rate $\lambda(\cdot)$ is a continuously differentiable function of ξ. Furthermore, $\lambda(\cdot)$ is a decreasing function of ξ and expected delay $E[D]$ is an increasing function of the arrival rate λ. Thus, it follows from the chain rule that $E[D(\xi)]$ is continuously differentiable and that

$$h'(\xi) = 1 - \frac{\partial}{\partial\xi} E[D(\xi)] > 0.$$

Hence, $h(\xi) = 0$ has a unique solution $0 < \xi^e < \infty$, which is the equilibrium expected delay. The equilibrium arrival rate is $\lambda^e(p) = (\Lambda_1+\Lambda_2)\,P(v > p+q\xi^e) + \Lambda_2$, and since ξ^e is finite, the equilibrium traffic intensity satisfies $\rho^e = \frac{\lambda^e(p)}{C\mu} < 1$. ■

Proof of Theorem 1: We first show that the optimal system operates under the QED heavy traffic regime. It follows from the analysis in Section EC.1.1 that this is equivalent to showing that the elasticity condition in Condition 1 holds under the sequence of optimal capacities $C^{\star,n}$ for large enough n. Suppose this is not the case. Then, there must exist a subsequence along which the elasticity assumption does not hold. With an abuse of notation we let $C^{\star,n}$ also denote this subsequence. Then, Theorem EC.3 implies that the system operates under the QD heavy traffic regime along this subsequence and the optimal revenue rate is given by

$$R^{\star,n} = (\Lambda_1^n + \Lambda_2^n)\tilde{p}\bar{F}(\tilde{p}) + o(n).$$

Now we define $\tilde{C}^n := (\Lambda_1^n + \Lambda_2^n)\bar{F}(\tilde{p}) + \Lambda_2^n$ and note that $\bar{p}(\tilde{C}^n) = \tilde{p}$. We differentiate between two cases: $\varepsilon(v_{min}) \le 1$ and $\varepsilon(v_{min}) > 1$. We start with the former, and assume $\varepsilon(v_{min}) \le 1$ first. Then,

since we know that the elasticity condition does not hold, it must be true that $C^{\star,n} \geq \tilde{C}^n$ for all n along this subsequence. Hence, the optimal profit rate satisfies

$$P^{\star,n} \leq (\Lambda_1^n + \Lambda_2^n)\tilde{p}\bar{F}(\tilde{p}) - w\tilde{C}^n\mu + o(n), \tag{23}$$

for all n along this subsequence. Note also that $\bar{p}(\tilde{C}^n) = \tilde{p}$ is a feasible solution to (EC.26), and since $\varepsilon(v_{min}) \leq 1$, we have that $\varepsilon(\tilde{p}) = 1$. However, it follows from Proposition 2 that there is the unique $\hat{\tilde{p}} := \bar{p}(\hat{C}^n)$ that satisfies the first order condition (EC.29), and thus $\varepsilon(\hat{\tilde{p}}) > 1$. Hence, $\bar{p}(\tilde{C}^n) = \tilde{p}$ is a strictly suboptimal solution to (EC.26), which implies that

$$(\tilde{C}^n\mu - \Lambda_2^n)\tilde{p} - w\tilde{C}^n\mu < (\hat{C}^n\mu - \Lambda_2^n)\hat{\tilde{p}} - w\hat{C}^n\mu,$$

or equivalently that

$$(\Lambda_1^n + \Lambda_2^n)\tilde{p}\bar{F}(\tilde{p}) - w\tilde{C}^n\mu < (\Lambda_1^n + \Lambda_2^n)\hat{\tilde{p}}\bar{F}(\hat{\tilde{p}}) - w\hat{C}^n\mu. \tag{24}$$

Next, we recall from Proposition 2 that $\hat{C}^n$ satisfies the elasticity assumption. Thus, it follows from the analysis in Section EC.1.1 that

$$\hat{P}^n = (\Lambda_1^n + \Lambda_2^n)\hat{\tilde{p}}\bar{F}(\hat{\tilde{p}}) - w\hat{C}\mu + o(n). \tag{25}$$

Then, combining (23) and (25) with (24) yields that

$$\limsup_{n\to\infty} \frac{\hat{P}^n}{P^{\star,n}} = \frac{P^n\left(\hat{C}^n, \hat{p}^n\right)}{P^n\left(C^{\star,n}, p^{\star,n}\right)} > 1, \tag{26}$$

and contradicts the optimality of the subsequence of capacities $C^{\star,n}$. Hence, we conclude that such a subsequence cannot exist when $\varepsilon(v_{min}) \leq 1$.

Next, we assume $\varepsilon(v_{min}) > 1$. Then, $\tilde{p} = v_{\min}$, and since the elasticity assumption does not hold, we see that $C^{\star,n} < \tilde{C}^n$ and that

$$P^{\star,n} < (\Lambda_1^n + \Lambda_2^n)\tilde{p}\bar{F}(\tilde{p}) - w\tilde{C}\mu + o(n).$$

Again, it follows from Proposition 2 and the analysis in Section EC.1.1 that

$$\hat{P}^n = (\Lambda_1^n + \Lambda_2^n)\hat{\tilde{p}}\bar{F}(\hat{\tilde{p}}) - w\hat{C}\mu + o(n).$$

Essentially the same arguments as before also yield that

$$(\Lambda_1^n + \Lambda_2^n)\tilde{p}\bar{F}(\tilde{p}) - w\tilde{C}^n\mu \leq (\Lambda_1^n + \Lambda_2^n)\hat{\tilde{p}}\bar{F}(\hat{\tilde{p}}) - w\hat{C}^n\mu,$$

where the equality applies only if $\hat{\tilde{p}} = v_{min}$. In either case, we arrive at (26), which leads to the same contradiction, and shows that there cannot exist a subsequence of optimal system capacities that

do not satisfy the elasticity assumption. Thus, we conclude that elasticity assumption is satisfied for large enough n.

Now, we prove the asymptotic optimality of the proposed policy. Recall that Λ_1^n and Λ_2^n grow large proportionally according to (EC.30) and define the constant $\eta := \frac{\Lambda_2^n}{\Lambda_1^n+\Lambda_2^n} = \frac{\eta_2}{\eta_1+\eta_2}$. We also let $\hat{\kappa} := \frac{\hat{C}^n\mu-\Lambda_2^n}{\Lambda_1^n+\Lambda_2^n}$ so that $\bar{F}(\hat{\bar{p}}) = \hat{\kappa}$. Then, since $\hat{C}^n$ satisfies the elasticity assumption, it follows from the analysis in Section EC.1.1 that

$$\frac{\hat{P}^n}{\Lambda_1^n+\Lambda_2^n} = \frac{P^n\left(\hat{C}^n, \hat{p}^n\right)}{\Lambda_1^n+\Lambda_2^n} \to \hat{\kappa}\hat{\bar{p}} - w(\hat{\kappa}+\eta). \tag{27}$$

Next, we let $\kappa^{\star,n} := \frac{C^{\star,n}-\Lambda_2^n}{\Lambda_1^n+\Lambda_2^n}$ and suppose that $\kappa^{\star,n} \to \kappa^\star \in [0,1]$ as $n \to \infty$. Since the elasticity assumption holds for large n, it also follows from the analysis in Section EC.1.1 that the profit rate scales as $P^{\star,n} = (C^{\star,n}\mu - \Lambda_2^n)\bar{p}(C^{\star,n}) + O(\sqrt{C^{\star,n}})$, and thus

$$\frac{P^{\star,n}}{\Lambda_1^n+\Lambda_2^n} = \frac{P^n\left(C^{\star,n}, p^{\star,n}\right)}{\Lambda_1^n+\Lambda_2^n} \to \kappa^\star\bar{p}(\kappa^\star) - w(\kappa^\star+\eta), \tag{28}$$

where $\bar{p}(\kappa^\star) := \bar{F}^{-1}(\kappa^\star)$. Note that the limit on the right-hand-side of (28) is the scaled asymptotic profit rate of a sequence of optimal policies, and so $\kappa^\star\bar{p}(\kappa^\star) - w(\kappa^\star+\kappa_2) \geq \hat{\kappa}\hat{\bar{p}} - w(\hat{\kappa}+\kappa_2)$. On the other hand, $\hat{C}^n$ is the unique maximizer of (EC.26) for all n, which implies $\kappa^\star\bar{p}(\kappa^\star) - w(\kappa^\star+\kappa_2) \leq \hat{\kappa}\hat{\bar{p}} - w(\hat{\kappa}+\kappa_2)$. This shows that $\kappa^\star = \hat{\kappa}$, and completes the proof. ■

Proof of Theorem 2: We first prove that the system with no control operates under the Efficiency Driven (ED) heavy traffic regime, and then we establish the scaling relationships.

Fix t, and suppose that $\rho_t^n \to \rho_t < 1$, possibly along a subsequence. Then it follows from the Erlang-C delay formula $d_t^n = \frac{\rho\alpha}{\lambda(1-\rho)}$ that $d_t^n \to 0$. Then, applying a Taylor expansion we get

$$\rho_t^n = \frac{(\Lambda_1^n + \Lambda_2^n\sigma_t)\bar{F}(\bar{p}_0) + \Lambda_2^n + o(n)}{C^{\star,0}\mu} \to \rho_t > 1,$$

which yields a contradiction. Hence, it must be that $\rho_t^n \to 1$ and

$$\rho_t^n \to 1 \text{ and } d_t^n \to \bar{d}_t,$$

where

$$\bar{d}_t := \frac{\bar{p}_t - \bar{p}_0}{q}$$

and $\bar{p}_t$ is the first order price that sets the capacity equal to demand in period t, i.e.,

$$(\Lambda_1^n + \Lambda_2^n\sigma_t)\bar{F}(\bar{p}_t) + \Lambda_2^n = C\mu.$$

Next, we let $d_t^n := \bar{d}_t + \delta_t^n$ and we observe from the Erlang-C delay formula that

$$\bar{d}_t + \delta_t^n = \frac{\alpha_t^n}{C\mu - \lambda_t^n},$$

and after applying a Taylor's expansion we get

$$\bar{d}_t + \delta_t^n = \frac{\alpha_t^n}{(\Lambda_1^n + \Lambda_2^n \sigma_t) f(\bar{p}_t + q\bar{d}_t) + o(n\delta_t^n)},$$

which along with the fact that α_t^n yields that

$$\delta_t^n = \frac{1}{(\Lambda_1^n + \Lambda_2^n \sigma_t)\bar{d}_t f(\bar{p}_t + q\bar{d}_t)} + o(1/n).$$

Hence, it follows that the exact delay scales as

$$d_t^n = \bar{d}_t + \frac{1}{(\Lambda_1^n + \Lambda_2^n \sigma_t)\bar{d}_t f(\bar{p}_t + q\bar{d}_t)} + o(1/n)$$

Finally, we recall that the revenue is given by $R_t^n = p_0^{\star,n}(\Lambda_1^n + \Lambda_2^n \sigma_t)\bar{F}(p_0^{\star,n} + qd_t^n)$ and applying a Taylor expansion we get

$$\begin{aligned} R_t^n &= p_0^{\star,n}(\Lambda_1^n + \Lambda_2^n \sigma_t)\bar{F}(\bar{p}_0 + \bar{d}_t) - p_0^{\star,n}(\Lambda_1^n + \Lambda_2^n \sigma_t) f(\bar{p}_0 + \bar{d}_t)\delta_t^n + o(n\delta_t^n) \\ &= p_0^{\star,n}\left[(C\mu - \Lambda_2) - 1/\bar{d}_t\right] + o(1) \end{aligned}$$

■

Proof of Theorem 3: Let $C_t^{\star,n}$ denote the optimal capacity in period t. Similarly, let us suppose that capacity and price can be jointly optimized along and let $C_t^{opt,n}$ and $p_t^{opt,n}$ denote the optimal capacity and price under joint capacity and price optimization. Recall that the system manager sets the initial price at

$$p_0^{opt,n} = \bar{p}_0(C_0^{opt,n}) + \frac{\pi_0^{opt}(C_0^{opt,n})}{\sqrt{n}} + o(\frac{1}{\sqrt{n}}),$$

Note also it follows [CITE HERE] that $p_0^{opt,n} \to \hat{\bar{p}}$. Our claim is that

$$p_0^{opt,n} = \hat{\bar{p}} + \frac{\hat{\pi}_0}{\sqrt{n}} + o(\frac{1}{\sqrt{n}}).$$

where $\hat{\bar{p}}$ is as defined in [CITE HERE]. To prove our claim, it suffices to show that $C_0^{opt,n} \leq \hat{C}_0^n + O(\sqrt{n})$, which, in turn, will follow if we show that the optimal second order revenue term $\sqrt{C}\Pi^\star(C)$ is decreasing in capacity C where $\Pi^\star(C) := (\bar{p}(C)\beta(\pi^\star) - \pi^\star)$

PROPOSITION 5. *$C_0^{opt,n} \leq \hat{C}_0^n + O(\sqrt{n})$, i.e., $\frac{C_0^{opt,n} - \hat{C}_0^n}{\sqrt{n}} \to M$ for some constant $M \geq 0$.*

To prove the proposition [CITE HERE], we show that the second order revenue loss is a decreasing function of the capacity C. Recall $C_0^{opt,n}$ maximizes profit and $\hat{C}_0^n$ maximizes the first order profit (and disregards the second order profit loss due to delay) for given n. Then, the second order revenue loss of former must be smaller, and thus, it must be that $C_0^{opt,n} \geq \hat{C}_0^n$ for large n. Finally, noting that the second order revenue loss is $\mathcal{O}(\sqrt{n})$ proves that the optimal capacity cannot be more than $\mathcal{O}(\sqrt{n})$ larger, and establishes the result.

Recall that the second order revenue term for a fixed capacity is given by $\sqrt{C}\Pi^\star(C)$ where

$$\Pi^\star(C) := \sqrt{C}\mu(\bar{p}(C)\beta(\pi^\star(C)) - \pi^\star(C)) \text{ for } \pi^\star := \arg\max \bar{p}(C)\beta(\pi) - \pi$$

To show that $\sqrt{C}\Pi^\star(C)$ is decreasing in C, we first differentiate it to get

$$\frac{d}{dC}\left[\sqrt{C}\Pi^\star(C)\right] = \sqrt{C}\frac{d}{dC}\left[\Pi^\star(C)\right] + \frac{1}{2\sqrt{C}}\Pi^\star(C).$$

Then, we observe that it suffices to show that $\frac{d}{dC}\left[\Pi^\star(C)\right] < 0$. To show this, we first fix C and recall that, given the fixed C, the optimal $\pi^\star$ satisfies the first order condition

$$\frac{d\Pi(\pi)}{d\pi} = \bar{p}(C)\frac{d\beta(\pi)}{d\pi} - 1 = 0.$$

Hence, it follows that $\frac{d\beta(\pi)}{d\pi}\Big|_{\pi=\pi^\star} = \frac{1}{\bar{p}(C)}$. Finally, we differentiate $\Pi^\star(C)$ to get

$$\begin{aligned}\frac{d\Pi^\star(C)}{dC} &= \frac{d\bar{p}(C)}{dC}\beta(\pi^\star(C)) + \bar{p}(C)\frac{d\beta}{d\pi}\Big|_{\pi=\pi^\star}\frac{d\pi^\star(C)}{dC} - \frac{d\pi^\star(C)}{dC} \\ &= \frac{d\bar{p}(C)}{dC}\beta(\pi^\star(C)) > 0,\end{aligned}$$

which establishes the desired result.

Recall from Theorem [CITE HERE] that under joint price and capacity optimization the optimal capacity and price pair $(C_t^{opt,n}, p_t^{opt,n})$ will satisfy

$$\frac{P^n(C_t^{opt,n}, p_t^{opt,n})}{\Lambda_1^n + \Lambda_2^n} \to \hat{\bar{p}}\bar{F}(\hat{\bar{p}}) - w(\bar{F}(\hat{\bar{p}}) + \kappa_2).$$

Next, recalling that

$$p_0^{opt,n} = \hat{\bar{p}} + \frac{\hat{\pi}_0}{\sqrt{n}} + o(\frac{1}{\sqrt{n}}),$$

we see that the profit under $(\hat{C}_t^n, p_0^{opt,n})$ will also satisfy

$$\frac{P^n(\hat{C}_t^n, p_0^{opt,n})}{\Lambda_1^n + \Lambda_2^n} \to \hat{\bar{p}}\bar{F}(\hat{\bar{p}}) - w(\bar{F}(\hat{\bar{p}}) + \kappa_2), \tag{29}$$

which implies that under the optimal capacity $C_t^{\star,n}$ capacity optimization it must also be the case that

$$\frac{P^n(C_t^{\star,n}, p_0^{opt,n})}{\Lambda_1^n + \Lambda_2^n} \to \hat{\bar{p}}\bar{F}(\hat{\bar{p}}) - w(\bar{F}(\hat{\bar{p}}) + \kappa_2). \tag{30}$$

Hence, we see from (29) and (30) that $\hat{C}_t^n$ is indeed asymptotically optimal. Next, we prove that the optimal capacity $C_t^{\star,n}$ along with the initially optimized price $p_0^{opt,n}$ indeed yields the QED heavy traffic regime.

To this end, we first let $C_t^{\star,n} = \hat{C}_t^n + \beta_t^n$. In what follows, we will establish that $\beta_t^n = \mathcal{O}(\sqrt{n})$, which will allow us to establish that the optimal system operates under QED. First, we see that β_t^n cannot be of higher order than n, otherwise the scaled cost will not converge as in (30). Similarly, it is also not possible to have $\beta_t^n > 0$ and $\beta_t^n = \mathcal{O}(n)$ since this would increase the limiting cost without increasing the limiting revenue.

Next, we show that $\beta_t^n < 0$ and $\beta_t^n = \mathcal{O}(n)$ is not possible either. Suppose first that $\beta_t^n < 0$ and $\beta_t^n = \mathcal{O}(n)$. Then, the limiting cost will be smaller than the right side of (30), which requires that the limiting revenue is also smaller. This, in turn, requires that the expected delay converges to a nonzero constant, i.e., $d_t^{\star,n} \to d_t^\star$ so that

$$\hat{\bar{p}}\bar{F}(\hat{\bar{p}}) = \hat{\bar{p}}\bar{F}(\hat{\bar{p}} + qd^\star) - w\beta\mu, \tag{31}$$

where $\beta := \lim_{n\to\infty} \frac{\beta^n}{\Lambda_1^n + \Lambda_2^n}$. However, this implies that

$$\begin{aligned}
\rho^{\star,n} = \frac{\lambda^{\star,n}}{C^{\star,n}\mu} &= \frac{(\Lambda_1^n + \Lambda_2^n)\bar{F}(\hat{\bar{p}} + qd^\star) + \Lambda_2^n + o(n)}{\hat{C}^n + \beta^n} \\
&= \frac{(\Lambda_1^n + \Lambda_2^n)\bar{F}(\hat{\bar{p}} + qd^\star) + \Lambda_2^n + o(n)}{(\Lambda_1^n + \Lambda_2^n)\bar{F}(\hat{\bar{p}}) + \Lambda_2^n + \beta^n} \\
&\to \frac{\bar{F}(\hat{\bar{p}} + qd^\star) + \kappa_2}{\bar{F}(\hat{\bar{p}}) + w\beta\mu + \kappa_2} > 1,
\end{aligned}$$

which is not possible. Therefore, we conclude that $\beta^n = o(n)$. It follows that $d^{\star,n} \to 0$ and $\rho^{\star,n} \to 1$

In the last step of the proof, we show that the $o(n)$ term β^n is in fact at most of order $\mathcal{O}(\sqrt{n})$ which will establish that QED is indeed optimal. To see this, let us assume that β^n is indeed at most of order $\mathcal{O}(\sqrt{n})$. Then, we note that the optimal arrival rate scales as

$$\begin{aligned}
\lambda^{\star,n} &= (\Lambda_1^n + \Lambda_2^n)\bar{F}(\hat{\bar{p}}) + \Lambda_2^n - (\Lambda_1^n + \Lambda_2^n)f(\hat{\bar{p}})(\frac{\hat{\pi}}{\sqrt{n}} + qd^{\star,n}) + o(n(\frac{\hat{\pi}}{\sqrt{n}} + qd^{\star,n})) \\
&= \hat{C}^n\mu - (\Lambda_1^n + \Lambda_2^n)f(\hat{\bar{p}})(\frac{\hat{\pi}}{\sqrt{n}} + qd^{\star,n}) + o(n(\frac{\hat{\pi}}{\sqrt{n}} + qd^{\star,n}))
\end{aligned} \tag{32}$$

First, let us suppose $\sqrt{n}d^{\star,n} \to 0$, which implies that $\sqrt{n}(1 - \rho^{\star,n}) \to \infty$. However, since β^n is at most of order $\mathcal{O}(\sqrt{n})$, it follows from (32) that $\sqrt{n}(1 - \rho^{\star,n}) \to 0$ and yields a contradiction.

Next, suppose that $\sqrt{n}d^{\star,n} \to \infty$. However, the optimal delay must satisfy

$$d^{\star,n} = \frac{\alpha^{\star,n}}{C^{\star,n}\mu - \lambda^{\star,n}} = \frac{\alpha^{\star,n}}{\beta^n\mu + (\Lambda_1^n + \Lambda_2^n)f(\hat{\bar{p}})(\frac{\hat{\pi}}{\sqrt{n}} + qd^{\star,n}) + o(n(\frac{\hat{\pi}}{\sqrt{n}} + qd^{\star,n}))}, \tag{33}$$

which is not possible, since $\alpha^{\star,n} \to 1$ when $\sqrt{n}d^{\star,n} \to \infty$.

The optimal profit under capacity optimization satisfies

$$\begin{aligned}
P^n(C^{\star,n}) &= p_0^{\star,n}(\Lambda_1^n+\Lambda_2^n)\bar{F}(p_0^{\star,n}+qd^{\star,n}) - C^{\star,n}\mu \\
&= (\hat{\bar{p}}+\frac{\hat{\pi}}{\sqrt{n}}+o(\frac{1}{\sqrt{n}}))(\Lambda_1^n+\Lambda_2^n)\left[\bar{F}(\hat{\bar{p}}) - f(\hat{\bar{p}})(\frac{\hat{\pi}}{\sqrt{n}}+qd^{\star,n})+o(\frac{\hat{\pi}}{\sqrt{n}}+qd^{\star,n})\right] \\
&- (\hat{C}^n\mu+\beta^n\mu)w \\
&= \hat{\bar{p}}(\Lambda_1^n+\Lambda_2^n)\bar{F}(\hat{\bar{p}}) - \hat{\bar{p}}(\Lambda_1^n+\Lambda_2^n)f(\hat{\bar{p}})(\frac{\hat{\pi}}{\sqrt{n}}+qd^{\star,n}) + \frac{\hat{\pi}}{\sqrt{n}}(\Lambda_1^n+\Lambda_2^n)\bar{F}(\hat{\bar{p}}) \\
&- (\hat{C}^n\mu+\beta^n\mu)w + o(n(\frac{\hat{\pi}}{\sqrt{n}}+qd^{\star,n})) \\
&= \hat{\bar{p}}(\Lambda_1^n+\Lambda_2^n)\bar{F}(\hat{\bar{p}}) - w((\Lambda_1^n+\Lambda_2^n)\bar{F}(\hat{\bar{p}})+\Lambda_2^n) \\
&- \left[\hat{\bar{p}}(\Lambda_1^n+\Lambda_2^n)f(\hat{\bar{p}})qd^{\star,n}+w\beta^n\mu\right] \\
&- \hat{\bar{p}}(\Lambda_1^n+\Lambda_2^n)f(\hat{\bar{p}})\frac{\hat{\pi}}{\sqrt{n}} + \frac{\hat{\pi}}{\sqrt{n}}(\Lambda_1^n+\Lambda_2^n)\bar{F}(\hat{\bar{p}}) + o(n(\frac{\hat{\pi}}{\sqrt{n}}+qd^{\star,n}))
\end{aligned} \tag{34}$$

Note first that the trade-off is between all the terms that include $d^{\star,n}$ and β^n since all else is independent of the capacity sizing decision. Hence, the question is if we can argue if β^n is of order of or of smaller order than $\sqrt{n}$.

First, let us suppose that $\frac{\beta^n}{\sqrt{n}} \to \infty$. Clearly, this case cannot be optimal. Next, suppose that $\frac{\beta^n}{\sqrt{n}} \to -\infty$. Recall that $C^{\star,n}\mu = \hat{C}^n\mu + \beta^n\mu$ and from (32) that

$$\lambda^{\star,n} = \hat{C}^n\mu - (\Lambda_1^n+\Lambda_2^n)f(\hat{\bar{p}})(\frac{\hat{\pi}}{\sqrt{n}}+qd^{\star,n}) + o(n(\frac{\hat{\pi}}{\sqrt{n}}+qd^{\star,n})).$$

Since it is necessary for stability that $\lambda^{\star,n} < C^{\star,n}\mu$, it follows that

$$(\Lambda_1^n+\Lambda_2^n)f(\hat{\bar{p}})(\frac{\hat{\pi}}{\sqrt{n}}+qd^{\star,n}) + \beta^n > 0, \tag{35}$$

for large n, which implies that $n\beta^n$ is at least of order of $\mathcal{O}(\beta^n)$. Therefore, it must also be that $\sqrt{n}d^{\star,n} \to \infty$ which implies that $\alpha^{\star,n} \to 1$ and that $\sqrt{n}(1-\rho^{\star,n}) \to 0$ (i.e., the ED heavy traffic regime). Then, from the latter we get that

$$\beta^n\mu + (\Lambda_1^n+\Lambda_2^n)f(\hat{\bar{p}})(\frac{\hat{\pi}}{\sqrt{n}}+qd^{\star,n}) + o(n(\frac{\hat{\pi}}{\sqrt{n}}+qd^{\star,n})) = o(\sqrt{n}). \tag{36}$$

However, for (36) to hold, it must be true that $\lim_{n\to\infty} \frac{(\Lambda_1^n+\Lambda_2^n)f(\hat{\bar{p}})qd^{\star,n}}{-\beta^n\mu} = 1$. Now, recall that $\hat{\bar{p}} > w$ by assumption. Then, it follows that $\lim_{n\to\infty} \frac{\hat{\bar{p}}(\Lambda_1^n+\Lambda_2^n)f(\hat{\bar{p}})qd^{\star,n}}{-w\beta^n\mu} > 1$, and thus

$$\frac{\hat{\bar{p}}(\Lambda_1^n+\Lambda_2^n)f(\hat{\bar{p}})qd^{\star,n}+w\beta^n\mu}{\sqrt{n}} \to \infty \tag{37}$$

i.e., the second order profit loss in (34) is of larger order than $\sqrt{n}$. Since this term is only of order of $\sqrt{n}$ under the QED regime (i.e., when $-\infty < \lim_{n\to\infty} \frac{\beta^n}{\sqrt{n}} < \infty$), we conclude that it cannot be optimal to have $\frac{\beta^n}{\sqrt{n}} \to -\infty$, which shows that QED is indeed the optimal operating regime. ■

Edited up to here.

References

Afèche P, Ata B (2013) Bayesian dynamic pricing in queueing systems with unknown delay cost characteristics. *Manufacturing & Service Operations Management* 15(2):292–304.

Araman VF, Caldentey R (2009) Dynamic pricing for nonperishable products with demand learning. *Operations Research* 57(5):1169–1188.

Armony M, Maglaras C (2004a) Contact centers with a call-back option and real-time delay information. *Operations Research* 52(4):527–545.

Armony M, Maglaras C (2004b) On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Operations Research* 52(2):271–292.

Dai J, He S (2012) Many-server queues with customer abandonment: A survey of diffusion and fluid approximations. *Journal of Systems Science and Systems Engineering* 21(1):1–36.

Datta H, Foubert B, Van Heerde HJ (2015) The challenge of retaining customers acquired with free trials. *Journal of Marketing Research* 52(2):217–234.

Foubert B, Gijsbrechts E (2016) Try it, you'll like it–or will you? The perils of early free-trial promotions for high-tech service adoption. *Marketing Science* 35(5):810–826.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2):79–141.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing & Service Operations Management* 4(3):208–227.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29(3):567–588.

Hassin R, Haviv M (2003) *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59 (Springer Science & Business Media).

Kumar S, Randhawa RS (2010) Exploiting market size in service systems. *Manufacturing & Service Operations Management* 12(3):511–526.

Kumar V (2014) Making" freemium" work. *Harvard business review* 92(5):27–29.

Lariviere MA (2006) A note on probability distributions with increasing generalized failure rates. *Operations Research* 54(3):602–604.

Lian Z, Gu X, Wu J (2016) A re-examination of experience service offering and regular service pricing under profit maximization. *European Journal of Operational Research* 254(3):907–915.

Maglaras C, Yao J, Zeevi A (2017) Optimal price and delay differentiation in large-scale queueing systems. *Management Science* 64(5):2427–2444.

Maglaras C, Zeevi A (2003) Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science* 49(8):1018–1038.

Maglaras C, Zeevi A (2005) Pricing and design of differentiated services: Approximate analysis and structural insights. *Operations Research* 53(2):242–262.

Mendelson H (1985) Pricing computer services: Queueing effects. *Communications of the ACM* 28(3):312–321.

Mendelson H, Whang S (1990) Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Operations Research* 38(5):870–883.

Mishra N, Najafi S, Najafi Asadolahi S, Tsay A (2018) How Freemium Gets Consumers to Pay a Premium: The Role of Loss-Aversion. Available at SSRN: https://ssrn.com/abstract=2961548.

Nair J, Wierman A, Zwart B (2016) Provisioning of large-scale systems: The interplay between network effects and strategic behavior in the user base. *Management Science* 62(6):1830–1841.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica: Journal of the Econometric Society* 37(1):15–24.

Plambeck EL, Ward AR (2006) Optimal control of a high-volume assemble-to-order system. *Mathematics of Operations Research* 31(3):453–477.

Randhawa RS, Kumar S (2008) Usage restriction and subscription services: Operational benefits with rational users. *Manufacturing & Service Operations Management* 10(3):429–447.

Stidham S (1985) Optimal control of admission to a queueing system. *IEEE Transactions on Automatic Control* 30(8):705–713.

Stidham S (2009) *Optimal design of queueing systems* (Chapman and Hall/CRC).

van Leeuwaarden JS, Mathijsen BW, Zwart B (2017) Economies-of-scale in resource sharing systems: Tutorial and partial review of the QED heavy-traffic regime. *arXiv preprint arXiv:1706.05397* .

Van Mieghem JA (2000) Price and service discrimination in queuing systems: Incentive compatibility of G$c\mu$ scheduling. *Management Science* 46(9):1249–1267.

Ward AR (2012) Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys in Operations Research and Management Science* 17(1):1–14.

Whitt W (2003) How multiserver queues scale with growing congestion-dependent demand. *Operations Research* 51(4):531–542.

Zhou W, Lian Z, Wu J (2014) When should service firms provide free experience service? *European Journal of Operational Research* 234(3):830–838.

Ziya S, Ayhan H, Foley RD (2004) Relationships among three assumptions in revenue management. *Operations Research* 52(5):804–809.

Electronic Companion to "Operational Perils and Benefits of Free Trials in Large Scale Service Systems"

In this electronic companion we provide supporting results and their proofs. In Section EC.1, we provide supporting results that were used in the proof of Theorem 1.

EC.1. Price Optimization in Large Scale Systems

In this section, we assume that the service provider can only change the per usage price charged while offering free trials. Our objective is to determine a pricing policy that is near optimal for large scale systems. Thus, we follow the same asymptotic approach and consider a sequence of systems where the capacity, regular market size, and free trial market size grow large. Then, we study two separate cases depending on whether the Elasticity condition given in 1 holds or not. Specifically, in Section EC.1.1, we assume the said condition holds, and we show that under pricing optimization the optimal system indeed operates in the QED heavy traffic regime. In Section EC.1.2, we consider the case where the elasticity condition is not satisfied, and show that the QD heavy traffic regime arises as the optimal operating regime. Next, we present the alluded elasticity condition which can equivalently be interpreted as a resource scarceness assumption.

CONDITION 1. (ELASTICITY CONDITION) Define the demand function $\Lambda_t(p) := (\Lambda_1 + \sigma_t\Lambda_2)\bar{F}(p)$. We assume that the first order price $\bar{p}_t$ is well-defined and that $\Lambda_t(p)$ is elastic[2] for all $p \geq \bar{p}_t$, or equivalently the optimization problem (that disregards the effects of congestion)

$$\begin{aligned} \max_{p\geq 0} \quad & (\Lambda_1 + \sigma_t\Lambda_2)p\bar{F}(p) \\ s.t. \quad & (\Lambda_1 + \sigma_t\Lambda_2)\bar{F}(p) + \Lambda_2 \leq C\mu \end{aligned}$$

is maximized at $\bar{p}_t > \tilde{p} := \arg\max p\bar{F}(p)$.

REMARK EC.1. (RELATION TO EARLIER LITERATURE) We note that, when $t = 0$, the elasticity condition in Condition 1 is equivalent to Assumption 1 in Maglaras and Zeevi (2003). Furthermore, when the elasticity condition holds for $t = 0$, it also holds for any $t > 0$.

EC.1.1. A System where the Elasticity Condition Holds: Optimality of QED

In this section, we study the pricing problem of the service firm when the customer valuation distribution or equivalently the demand function satisfies the elasticity condition, which we introduce next. As in the statement of the elasticity condition in Condition 1, we prove our results for an arbitrary $t \geq 0$ for theoretical completeness. We also remind the reader that we we consider sequences of systems indexed by n that appears in the superscript and that we refer to optimal quantities (under price optimization now) along this sequence with an additional $\star$ in the superscript. Our

[2] $\lambda(p)$ is elastic at price p if $\varepsilon(p) > 1$ where $\varepsilon(p) := -\frac{d\lambda(p)}{dp}\frac{p}{\lambda(p)}$

first result, which is also the main result in this section, and establishes that the QED regime is indeed the optimal regime under pricing optimization provided that the demand function satisfies the elasticity condition (and extends Maglaras and Zeevi (2003) to include free trial demand).

THEOREM EC.1. (OPTIMALITY OF QED) *Suppose that elasticity condition in Condition 1 holds and that the regular market size and the free trial market size scale with capacity according to* (14)*. Then, under the optimal pricing policy, the system operates in the QED heavy traffic regime. Furthermore, the optimal price scales as*

$$p_t^{\star,n} = \bar{p}_t + \frac{\pi_t^\star}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{EC.1}$$

where $\bar{p}_t$ is the first order price defined as in (17) and $\pi_t^\star$ is the optimal second order price correction defined as

$$\pi_t^\star := \arg\min_{\pi\in\mathbb{R}} \{\bar{p}_t \beta(\pi) - \pi(1-\kappa_2)\}. \tag{EC.2}$$

To prove Theorem EC.1, we need to auxiliary results which we present and prove next. First, we show that, $\rho_t^{\star,n}$, the optimal utilization (that corresponds to the optimal price) in a system with size n, approaches unity.

PROPOSITION EC.1. *Suppose that elasticity condition in Condition 1 holds and that the regular market size and the free trial market size scale with capacity according to* (14)*. Then, the utilization under the optimal pricing policy approaches unity, i.e.,*

$$\rho_t^{\star,n} \to 1 \text{ as } n \to \infty.$$

Proof of Proposition EC.1: We prove the result by contradiction. First, note that it follows from Proposition 1 that $\rho^{\star,n} = \frac{\lambda^{\star,n}}{n\mu} < 1$, and so

$$\limsup_{n\to\infty} \rho^{\star,n} \leq 1.$$

Now, let us suppose that $\liminf_{n\to\infty} \rho^{\star,n} \leq (1-\delta)$ for some $\delta \in (0,1)$. Then, since $E[D^{\star,n}] = \frac{\rho^{\star,n}\alpha^{\star,n}}{\lambda^{\star,n}(1-\rho^{\star,n})} \leq \frac{(1-\delta)}{n\delta}$, it follows that $\liminf_{n\to\infty} E[D^{\star,n}] = 0$. Then, there exists a subsequence where the liminf is attained as a limit and, with an abuse of notation, we also index this subsequence by n. Next, we apply a Taylor's expansion along this sequence to get

$$\begin{aligned}
\rho^{\star,n} &= \frac{(\Lambda_1^n + \Lambda_2^n)\bar{F}^{-1}(p^{\star,n} + qE[D^{\star,n}]) + \Lambda_2^n}{n\mu} \\
&= (\kappa_1+\kappa_2)\bar{F}^{-1}(p^{\star,n} + qE[D^{\star,n}]) + \kappa_2 \\
&= (\kappa_1+\kappa_2)\bar{F}^{-1}(p^{\star,n}) - (\kappa_1+\kappa_2) f(p^{\star,n}) qE[D^{\star,n}] + o(E[D^{\star,n}]) + \kappa_2.
\end{aligned} \tag{EC.3}$$

Consequently, it follows from (EC.3) that

$$\liminf_{n\to\infty} \bar{F}^{-1}\left(p^{\star,n}\right) \leq \frac{1-\kappa_2-\delta}{(\kappa_1+\kappa_2)} = \bar{F}^{-1}\left(\bar{p}\right) - \frac{\delta}{(\kappa_1+\kappa_2)} < \bar{F}^{-1}\left(\bar{p}\right). \tag{EC.4}$$

Next, we define the price sequence be defined as $\bar{p}^n := \bar{p}$, and let $\bar{d}^n$ denote the associated expected delay. Recall from Theorem EC.2 that $d^n \to 0$. Then, it follows from the elasticity condition in Assumption 1 and (EC.4) that

$$\limsup_{n\to\infty} \frac{R^n\left(\bar{p}^n\right)}{R^n\left(p^{\star,n}\right)} \geq (1+\delta'),$$

which contradicts the optimality of the sequence $p^{\star,n}$. Thus, we conclude that $\liminf_{n\to\infty} \rho^{\star,n} \leq (1-\delta)$ for all $\delta \in (0,1)$, which implies that $\liminf_{n\to\infty} \rho^{\star,n} = 1$ and this completes the proof. ■

Note that Proposition EC.1 does not directly imply Theorem EC.1 since the utilization might also approach unity in other heavy traffic regimes. Thus, we need to show that the utilization converges to one in a specific way that corresponds to the QED heavy traffic regime, which, in turn, requires characterizing the scaling relationships of the QED regime. We establish this in the next result.

THEOREM EC.2. (QED SCALING RELATIONSHIPS) *Assume that the regular market size and the free trial market size scale with capacity according to* (14) *for some fixed* $\kappa_1 > 0$ *and* $0 \leq \kappa_2 < 1$ *and let* $\alpha_t^n(\rho^n, C^n)$ *denote the Erlang-C delay probability in free trial cycle* t *along this sequence. Then,*

$$\alpha_t^n(\rho^n, C^n) \to \alpha_t \in (0,1) \text{ as } n \to \infty \text{ if and only if} \tag{EC.5}$$

(i) equilibrium arrival rate scales as

$$\lambda_t^n = n\mu - \beta_t\sqrt{n}\mu + o(\sqrt{n}), \tag{EC.6}$$

or equivalently, the equilibrium traffic intensity scales as

$$\rho_t^n = 1 - \frac{\beta_t}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{EC.7}$$

where $\beta_t \in (0,\infty)$ *is uniquely defined via* α_t*;*

(ii) The expected delay scales as

$$d_t^n := E\left[D_t^n\right] = \frac{d_t}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{EC.8}$$

where $d_t \in (0,\infty)$ *is a function of* β_t*;*

(iii) The price p_t^n *scales as*

$$p^n = \bar{p}_t + \frac{\pi_t}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{EC.9}$$

where $\bar{p}_t$ *is as defined in (17), and* $\pi_t \in (-\infty,\infty)$ *is a function of* β_t*.*

Proof of Theorem EC.2: To prove part (i), we observe that, for a given price, the system is an $M/M/C$ queue with an arrival rate that corresponds to the equilibrium arrival rate determined by (1). Then, we can apply Proposition 1 in Halfin and Whitt (1981) to the sequence of systems that scale according to (14). Specifically, it follows from Proposition 1 in Halfin and Whitt (1981) that

$$\lim_{n\to\infty} \alpha^n(\rho^n, C^n) = \alpha \in (0,1) \text{ if and only if } \lim_{n\to\infty} \sqrt{n}\,(1-\rho^n) = \beta \in (0,\infty),$$

which is equivalent to

$$\lambda^n = n\mu - \beta\sqrt{n}\mu + o(\sqrt{n}).$$

It also follows from Proposition 1 in Halfin and Whitt (1981) that $\alpha \in (0,1)$ and $\beta > 0$ uniquely determine each other via

$$\alpha := \alpha(\beta) = \frac{\phi(\beta)}{\beta\Phi(\beta) + \phi(\beta)}, \tag{EC.10}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal probability density function and cumulative distribution function, respectively.

For part (ii), recall that the expected delay is given by

$$d^n := E[D^n] = \frac{\rho^n \alpha^n(\rho^n, C^n)}{\lambda^n (1-\rho^n)}.$$

Recall also from part (i) that $\sqrt{n}\,(1-\rho^n) \to \beta$, and that $\rho^n \to 1$ as $n \to \infty$. Hence, it follows that

$$\sqrt{n}E[D^n] = \sqrt{n}d^n \to d(\beta) := \frac{\alpha(\beta)}{\beta} = \frac{\phi(\beta)}{\beta[\beta\Phi(\beta) + \phi(\beta)]}.$$

For part (iii), observe that part (ii) implies that $d^n = d/\sqrt{n} + o(1/\sqrt{n})$. Then, applying a Taylor's expansion around p^n we get

$$\begin{aligned}
\lambda(p) &= (\Lambda_1^n + \Lambda_2^n)\,\bar{F}(p^n + qE[D^n]) + \Lambda_2^n \\
&= (\Lambda_1^n + \Lambda_2^n)\,\bar{F}(p^n) - (\Lambda_1^n + \Lambda_2^n)\,f(p^n)q\frac{d}{\sqrt{n}} + o\left(\sqrt{n}\right) + \Lambda_2^n \\
&= n\mu - \beta\sqrt{n}\mu + o\left(\sqrt{n}\right),
\end{aligned}$$

where the last equality follows from part (ii). Then, it follows that

$$(\Lambda_1^n + \Lambda_2^n)\,\bar{F}(p^n) + \Lambda_2^n = n\mu + \delta\sqrt{n}\mu + o\left(\sqrt{n}\right),$$

for some constant δ. Recall that $\bar{F}(\bar{p}) = \dfrac{n\mu - \Lambda_2^n}{\Lambda_1^n + \Lambda_2^n}$. Hence, the equality above implies that

$$p^n = \bar{p} + \frac{\pi}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

which establishes part (iii) and completes the proof. ■

In the analysis that follows, we will find it more convenient to view the second order arrival rate correction β_t (also known as the *service grade*) as a function of the second order price correction term π_t. Our next result establishes this connection.

PROPOSITION EC.2. *Suppose the assumption in Theorem EC.2 holds. Then, for fixed* $\pi_t \in (-\infty, \infty)$, $\beta_t(\pi_t)$ *is the unique solution of*

$$\frac{1}{q_t}\left(\frac{\bar{F}\left(\bar{p}_t\right)}{f\left(\bar{p}_t\right)\left(1-\kappa_2\right)}\beta_t - \pi_t\right) = \frac{\phi\left(\beta_t\right)}{\beta_t\left(\beta_t\Phi\left(\beta_t\right)+\phi\left(\beta_t\right)\right)}, \tag{EC.11}$$

where $\phi(\cdot)$ *and* $\Phi(\cdot)$ *denote the standard normal probability density function and cumulative distribution function, respectively.*

Proof of Proposition EC.2: Recall from the proof of Theorem EC.2 that

$$\sqrt{n}E\left[D^n\right] = \sqrt{n}d^n \to d(\beta) := \frac{\alpha(\beta)}{\beta} = \frac{\phi\left(\beta\right)}{\beta\left[\beta\Phi\left(\beta\right)+\phi\left(\beta\right)\right]}. \tag{EC.12}$$

Recall also from Theorem EC.2 that the arrival rate λ^n scales as

$$\lambda^n = n\mu - \beta\sqrt{n}\mu + o(\sqrt{n}). \tag{EC.13}$$

Then, we can invoke (EC.8) and (EC.9) and apply a Taylor's expansion in (1) to get

$$\begin{aligned}\lambda^n &= \left(\Lambda_1^n + \Lambda_2^n\right)\bar{F}\left(p^n + qE\left[D^n\right]\right) + \Lambda_2^n \\ &= \left(\Lambda_1^n + \Lambda_2^n\right)\bar{F}\left(\bar{p}\right) - \left(\Lambda_1^n + \Lambda_2^n\right)\frac{f(\bar{p})}{\sqrt{n}}\left(\pi + qd\right) + o\left(\sqrt{n}\right) + \Lambda_2^n.\end{aligned} \tag{EC.14}$$

Noting that $\bar{F}(\bar{p}) = \frac{n\mu - \Lambda_2^n}{\Lambda_1^n + \Lambda_2^n}$ and combining (EC.13) and (EC.14) we get

$$\beta = \left(\pi + qd\left(\beta\right)\right) f\left(\bar{p}\right)\left(1-\kappa_2\right)/\bar{F}\left(\bar{p}\right). \tag{EC.15}$$

Finally, we use the definition of $d(\beta)$ given in (EC.12) and rearrange (EC.15) to get

$$\frac{1}{q}\left(\frac{\bar{F}\left(\bar{p}\right)}{f\left(\bar{p}\right)\left(1-\kappa_2\right)}\beta - \pi\right) = \frac{\phi\left(\beta\right)}{\beta\left(\beta\Phi\left(\beta\right)+\phi\left(\beta\right)\right)}. \tag{EC.16}$$

It remains to show that $\beta(\pi)$ defined through (EC.16) is unique. To see this, we first rearrange (EC.16) to get

$$h(\beta) := \frac{\bar{F}\left(\bar{p}\right)}{qf\left(\bar{p}\right)\left(1-\kappa_2\right)}\beta - \frac{\phi\left(\beta\right)}{\beta\left(\beta\Phi\left(\beta\right)+\phi\left(\beta\right)\right)} = \frac{\pi}{q}.$$

Next, we observe that $h(\beta)$ is a continuous increasing function for $\beta > 0$. Furthermore, $\lim_{\beta\to 0} h(\beta) = -\infty$ and $\lim_{\beta\to\infty} h(\beta) = \infty$. Hence, it follows that $h(\beta) = \frac{\pi}{q}$ has a unique solution establishing the desired result. ■

Now, we are in a position to prove Theorem EC.1.

Proof of Theorem EC.1: Let $\{p^{\star,n}\}$ denote a revenue maximizing price sequence. Also, let $d^{\star,n} :=$

$\frac{\rho^{\star,n}\alpha^{\star,n}}{\lambda^{\star,n}(1-\rho^{\star,n})}$ denote the optimal expected delay, where $\alpha^{\star,n} := \alpha^{\star,n}(\rho^{\star,n}, n)$ is the optimal Erlang-C delay probability. Recall that $\rho^{\star,n} \to 1$ as $n \to \infty$. Now, suppose that $\liminf_{n\to\infty} n(1-\rho^{\star,n}) \leq M$ for some positive constant $M > q/\bar{p}$. Then, $\liminf_{n\to\infty} \sqrt{n}(1-\rho^{\star,n}) = 0$, and it follows from Proposition 1 in Halfin and Whitt (1981) that $\alpha^{\star,n} \to 1$ and so $\limsup_{n\to\infty} d^{\star,n} \geq 1/M$. Also, note that

$$\rho^{\star,n} = \frac{\lambda^{\star,n}}{n\mu} = \frac{(\Lambda_1^n + \Lambda_2^n)\bar{F}(p^{\star,n} + qd^{\star,n}) + \Lambda_2^n}{n\mu} \to 1,$$

which yields that $p^{\star,n} + qd^{\star,n} \to \bar{p}$. Since $\limsup_{n\to\infty} d^{\star,n} \geq 1/M$, it follows that

$$\liminf_{n\to\infty} p^{\star,n} \leq \bar{p} - \frac{q}{M},$$

which in turn implies

$$\liminf_{n\to\infty} \frac{R^n(p^{\star,n})}{R^n(\bar{p})} \leq 1 - \frac{q}{\bar{p}M} < 1,$$

and contradicts the optimality of $p^{\star,n}$. Hence, we conclude that $\liminf_{n\to\infty} n(1-\rho^{\star,n}) \geq M$ for any positive constant. Noting that M is arbitrary, it follows that $n(1-\rho^{\star,n}) \to \infty$. Thus, $d^{\star,n} = o(1)$ and $p^{\star,n} \to \bar{p}$.

Next, we let $\beta^{\star,n} := 1 - \rho^{\star,n}$ and $\pi^{\star,n} := p^{\star,n} - \bar{p}$. Then, we apply a Taylor's expansion to get

$$\lambda^{\star,n} = n\mu - n\mu f(\bar{p})(\kappa_1 + \kappa_2)(\pi^{\star,n} + qd^{\star,n}) + O\left(n(\pi^{\star,n} + qd^{\star,n})^2\right).$$

We can also express the revenue rate as

$$R(p^{\star,n}, n) = \lambda^{\star,n} p^{\star,n} = (n\mu - \Lambda_2^n)\bar{p} - \underbrace{n\mu\left[\bar{p}\beta^{\star,n} - \pi^{\star,n}(1-\kappa_2)\right]}_{\psi(\beta^{\star,n})} + O(n\beta^{\star,n}\pi^{\star,n}),$$

where the last term is of lower order since $\beta^{\star,n} \to 0$ and $\pi^{\star,n} \to 0$ as $n \to \infty$. Next, we take a closer look at $\psi(\beta^{\star,n})$ which captures the second-order effects of stochasticity:

$$\begin{aligned}
\psi(\beta^{\star,n}) &= n\mu\left(\bar{p}\beta^{\star,n} - \pi^{\star,n}(1-\kappa_2)\right) \\
&= n\mu\left(\bar{p}\beta^{\star,n} - \frac{(1-\kappa_2)}{f(\bar{p})(\kappa_1+\kappa_2)}\beta^{\star,n} + qd^{\star,n}(1-\kappa_2)\right) \\
&= \sqrt{n}\mu\left(\sqrt{n}\beta^{\star,n}\left(\bar{p} - \frac{\bar{F}(\bar{p})}{f(\bar{p})}\right) + q\sqrt{n}d^{\star,n}(1-\kappa_2)\right). \qquad \text{(EC.17)}
\end{aligned}$$

Now, let us suppose $\sqrt{n}\beta^{\star,n} = o(1)$. Then, it follows from Proposition 1 in Halfin and Whitt (1981) that $\alpha^{\star,n} \to 1$. Consequently, $\sqrt{n}d^{\star,n} \to \infty$ as $n \to \infty$. Hence, $\frac{\psi(\beta^{\star,n})}{\sqrt{n}} \to \infty$. Next, suppose $\sqrt{n}\beta^{\star,n} \to \infty$. Then, it follows from Proposition 1 in Halfin and Whitt (1981) that $\limsup_{n\to\infty} \sqrt{n}d^{\star,n} < \infty$. Furthermore, from the elasticity condition in Assumption 1, we get

$$\varepsilon(p) = -\frac{d(\Lambda_1^n + \Lambda_2^n)\bar{F}(p)}{dp}\frac{p}{(\Lambda_1^n + \Lambda_2^n)\bar{F}(p)} = f(p)\frac{p}{\bar{F}(p)} > 1,$$

or equivalently that $\bar{F}(p) < pf(p)$. Thus, we see that $\frac{\psi(\beta^{\star,n})}{\sqrt{n}} \to \infty$, again. Finally, suppose that $\sqrt{n}\beta^{\star,n} \to \beta > 0$. Then, it follows from Theorem 1 that $\frac{\psi(\beta^{\star,n})}{\sqrt{n}} \to \psi < \infty$ and we conclude that the lost revenue due to the second order effect of stochasticty is minimized.

To complete the proof, we first observe from Theorem 1 that the optimal price should scale as

$$p^{\star,n} = \bar{p} + \frac{\pi^\star}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{EC.18}$$

So, we let

$$\pi^\star := \arg\min_{\pi \in \Re} \{\bar{p}\beta(\pi) - \pi\left(\frac{n\mu - \Lambda_2^n}{\Lambda_2^n}\right)\}, \tag{EC.19}$$

which minimizes the second-order effects, and optimally balances lost revenues with delay costs. ■

Note that we eventually would like to understand the behavior of the optimal revenue rate. Hence, it is important to observe that Theorem EC.2 and Theorem EC.1 taken together imply that the optimal arrival rate $\lambda^{\star,n}$ scales as

$$\lambda_t^{\star,n} = n\mu - \beta(\pi_t^\star)\sqrt{n}\mu + o(\sqrt{n}), \tag{EC.20}$$

where $\beta(\pi)$ and $\pi_t^\star$ are as defined in (EC.11) and (EC.2), respectively. This, in turn, paves the way for a two-part representation for the optimal revenue rate. Specifically, we can combine (EC.1) with (EC.20) and write the optimal revenue rate in a system of size n as

$$R_t^{\star,n} = p_t^{\star,n}(\lambda_t^{\star,n}(p_t) - \Lambda_2^n) = (n\mu - \Lambda_2^n)\bar{p}_t - \sqrt{n}\mu\left[\bar{p}_t\beta(\pi_t^\star) - (1-\kappa_2)\pi_t^\star\right] + o(\sqrt{n}). \tag{EC.21}$$

Hence, we see that the optimal revenue rate also admits a two-part decomposition similar to (EC.1). Finally, we observe that the optimal delay $d^{\star,n}$ scales analogously, i.e., we have

$$d_t^{\star,n} := E\left[D_t^{\star,n}\right] = \frac{d_t(\beta(\pi_t^\star))}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{EC.22}$$

EC.1.2. A System where the Elasticity Condition doesn't Hold: Optimality of QD

In this section we consider the case where the elasticity condition in Condition 1 does not hold. Then, the system is over-capacitated and either $\tilde{p}_t \geq \bar{p}_t$ for $\tilde{p}_t$, or $\bar{p}_t$ is not well-defined since $\frac{C\mu - \Lambda_2}{\Lambda_1 + \Lambda_{2,t}} > 1$. Again, we pursue an asymptotic analysis by letting the capacity and the regular and free trials market sizes grow large proportionally according to (14). Noting that the elasticity condition that underpins the QED regime implies a capacity constrained system, it is intuitive to think that the optimal operating regime when this condition does not hold should be the quality-driven (QD) heavy traffic regime where the Erlang-C delay probability approaches zero. Our next result proves that this intuition is indeed correct.

THEOREM EC.3. (OPTIMALITY OF QD) *Suppose the elasticity condition in Assumption 1 does not hold and that capacity scales proportionally to demand according to* (14) *for some fixed* $\kappa_1 > 0$ *and* $0 \le \kappa_2 < 1$. *Then, under the optimal pricing policy, the system operates in the QD heavy traffic regime. Furthermore, the optimal price satisfies*

$$p_t^{\star,n} \to \tilde{p}_t \ \ as \ \ n \to \infty, \tag{EC.23}$$

where $\tilde{p}_t := \arg\max p\bar{F}(p)$.

Proof of Theorem EC.3: Recall that the elasticity assumption does not hold if $\varepsilon(v_{min}) \le 1$ and $\bar{p} \le \tilde{p}$ or if $\bar{p}$ is undefined. We start with the case $\varepsilon(v_{min}) \le 1$ and consider the case $\bar{p} = \tilde{p}$ first. Observe that Proposition EC.1 continues to hold in this case. Hence, $\rho^{\star,n} \to 1$ and we observe that the proof of Theorem EC.1 also holds up to the arguments leading to (EC.17). Thus, $p^{\star,n} \to \bar{p} = \tilde{p}$. Next, from (EC.17), we see that $\bar{p} - \frac{\bar{F}(\bar{p})}{f(\bar{p})} = 0$ since $\varepsilon(\bar{p}) = \varepsilon(\tilde{p}) = 1$. Hence, the first term in the parenthesis in (EC.17) is zero for all n. Then, to minimize the second order revenue loss, the optimal policy must be such that $\sqrt{n}\beta^{\star,n} \to \infty$, which proves that the system operates in QD.

Next, we consider the case $\bar{p} < \tilde{p}$. The idea of the proof is as follows: For a sequence of prices such that $\tilde{p}^n \to \tilde{p}$ as $n \to \infty$, we establish that the associated expected delay $\tilde{d}^n \to 0$ as $n \to \infty$. This allows us to show that such a policy achieves a first order revenue rate of $(\Lambda_1^n + \Lambda_2^n)\tilde{p}\bar{F}(\tilde{p})$, which also shows that $p^{\star,n} \to \tilde{p}$ as $n \to \infty$.

Define $\tilde{d}^n := E[\tilde{D}^n]$ as the expected delay along a sequence of systems with corresponding prices such that $p^n \to \tilde{p}$ as $n \to \infty$. Also, let $\bar{d}^n$ be the expected delay along a sequence of systems with pricing policy $\bar{p}^n = \bar{p}$ for all n. Since $p^n \to \tilde{p} > \bar{p}$, it must be true that $p^n > \bar{p}$ for all n large enough. Then, it also follows that $\tilde{d}^n < \bar{d}^n$ for n large enough. To see this, we define $h_p(\xi)$ as in the proof of Proposition 1 and we make the dependence on the fixed priced p explicit. For $\xi = \bar{d}^n$, $h_{\bar{p}^n}(\bar{d}^n) = 0$, and since $p^n > \bar{p} = \bar{p}^n$, we see that $h_{p^n}(\xi) > 0$ for all n large enough. Noting that $h_{p^n}(\tilde{d}^n) = 0$ and that $h_{p^n}(\xi)$ is increasing, we conclude that $\tilde{d}^n < \bar{d}^n$ for all n large enough. Finally, from Theorem EC.2, we know that $\bar{d}^n \to 0$ as $n \to \infty$ since $\bar{p}^n = \bar{p}$. Thus, it also follows that $\tilde{d}^n \to 0$ since $\tilde{d}^n$ is positive and $\tilde{d}^n < \bar{d}^n$ for all n large enough.

When $\bar{p}$ is not well-defined (i.e., when the system is over-capacitated), the result follows similarly by comparing the system to one where $\bar{p}$ is well-defined. Since the latter has less capacity and delay converges to zero, it must also converge to zero in the former case.

Next, we let $\tilde{R}^n$ denote corresponding revenue rate in the sequence of systems where $p^n \to \tilde{p}$ as $n \to \infty$. Noting that $\tilde{d}^n \to 0$, we apply a Taylor expansion to get the first order revenue rate as

$$\tilde{R}^n = (\Lambda_1^n + \Lambda_2^n)p^n\bar{F}(p^n + qd^n) = (\Lambda_1^n + \Lambda_2^n)\tilde{p}\bar{F}(\tilde{p}) + o(n),$$

which yields

$$\limsup_{n\to\infty}\frac{\tilde{R}^n}{R^{\star,n}}=\limsup_{n\to\infty}\frac{(\Lambda_1^n+\Lambda_2^n)p^n\bar{F}(p^n+qd^n)}{(\Lambda_1^n+\Lambda_2^n)p^{\star,n}\bar{F}(p^{\star,n}+qd^{\star,n})}$$
$$=\limsup_{n\to\infty}\frac{\tilde{p}\bar{F}(\tilde{p})}{p^{\star,n}\bar{F}(p^{\star,n}+qd^{\star,n})}\geq 1,$$

since $\tilde{p}:=\arg\max_p p\bar{F}(p)$. Hence, for $p^{\star,n}$ to be optimal, it must also be that $p^{\star,n}\to\tilde{p}$.

It remains to show that the delay probability converges to zero in order to show that the system operates in the QD heavy traffic regime. We do so by showing that $\rho^{\star,n}\to\rho<1$, and the result follows from Proposition 1 in Halfin and Whitt (1981). Since $p^{\star,n}\to\tilde{p}$ and $d^{\star,n}\to 0$, it follows that

$$\lim_{n\to\infty}\rho^{\star,n}=\lim_{n\to\infty}\frac{(\Lambda_1^n+\Lambda_2^n)\bar{F}(p^{\star,n}+qd^{\star,n})+\Lambda_2^n}{C^n\mu}$$
$$=\lim_{n\to\infty}\frac{(\Lambda_1^n+\Lambda_2^n)\bar{F}(\tilde{p})+o(n)+\Lambda_2^n}{C^n\mu}=\rho<1,$$

and completes the proof. ■

EC.2. Proofs of Other Results

Now we consider the joint problem where the system manager can choose both the capacity and price to maximize the long run average profit rate. In addition to the model parameters described in Section 2, we also assume a linear capacity cost given as $\$w\mu>0$ per unit of capacity per unit time. Then, the objective of the system manager is

$$\max_{p_t,C_t>0}P_t(p_t,C_t):=\sum_{t=1}^{T}R_t(p_t,C_t)-w\mu C_t. \tag{EC.24}$$

We let $p_t^\star$ and $C_t^\star$ be the associated maximizers of (EC.24) and define the maximum profit rate as $P_t^\star:=P_t(p_t^\star,C_t^\star)$. Observe that one might compute $p_t^\star$ and $C_t^\star$, by first optimizing $P_t(p_t,C_t)$ over p_t for fixed C_t, and then optimizing over C_t. However, this approach does not yield any analytically tractable expressions for $p_t^\star$ and $C_t^\star$ and thus does not offer any insight. Hence, we build upon the results in Section EC.1, and consider approximate solutions that are appropriate for large scale systems. Next, we describe the intuition behind our approach.

Recall from Section EC.1 that, under the elasticity assumption, the revenue rate of a large scale system for a given price and capacity is given by

$$R_t(p_t,C_t)\approx(C_t\mu-\Lambda_2)\,\bar{p}_t\,(C_t)-\sqrt{C}\mu\left[\bar{p}_t\,(C)\,\beta\,(\pi_t)-\pi_t\,(1-\kappa_2)\right].$$

The above equation suggest that, optimal profit rate of the joint profit maximization problem in (EC.24) can be approximated by the *second order approximation* as

$$P_t^\star\approx\max_{C_t>0,\pi_t\in\Re}\left[(C_t\mu-\Lambda_2)\,\bar{p}_t\,(C_t)-w\mu C_t\right]-\sqrt{C_t}\mu\left[\bar{p}_t\,(C_t)\,\beta\,(\pi_t)-\pi_t\,(1-\kappa_2)\right]. \tag{EC.25}$$

Note that for large enough capacities the effect of the second order square-root term in (EC.25) will be negligible, and thus, the capacity sizing and pricing problems decouple for large scale systems. Hence, we propose the following approximate solution:

Step 1 Capacity Sizing: Optimize the *first-order approximation* for optimal profit rate given by

$$\max_{C_t>0} \hat{P}_t(C_t) := (C_t\mu - \Lambda_2)\,\bar{p}_t\,(C_t) - w\mu C_t, \tag{EC.26}$$

and set the capacity to $\hat{C}_t$, the corresponding maximizer. Also, define the corresponding first order price $\hat{\bar{p}}_t := \bar{p}_t(\hat{C}_t)$ and the optimal first order profit rate $\hat{P}_t^\star := \hat{P}_t(\hat{C}_t)$.

Step 2 Pricing: Given $\hat{C}_t$ and $\hat{\bar{p}}_t$, minimize the second order profit correction term in square brackets in (EC.25) over π_t. Let $\hat{\pi}_t$ be the corresponding maximizer. Set the price to $\hat{p}_t := \hat{\bar{p}}_t + \hat{\pi}_t/\sqrt{\hat{C}_t}$.

In order to get a better understanding of the optimal $\hat{C}_t$ that solves (EC.26), we first recall that $\bar{p}_t\,(C_t) := \bar{F}^{-1}\left(\frac{C_t\mu-\Lambda_2}{\Lambda_1+\Lambda_{2,t}}\right)$ and differentiate (EC.26) with respect to C_t which yields the first order condition (FOC)

$$\mu\bar{F}^{-1}\left(\frac{C_t\mu-\Lambda_2}{\Lambda_1+\Lambda_{2,t}}\right) + (C_t\mu-\Lambda_2)\,\frac{1}{-f\left(\bar{F}^{-1}\left(\frac{C_t\mu-\Lambda_2}{\Lambda_1+\Lambda_{2,t}}\right)\right)}\left(\frac{\mu}{\Lambda_1+\Lambda_2}\right) - w\mu = 0. \tag{EC.27}$$

We can also express the FOC in terms of $\bar{p}_t(C_t)$ as

$$\bar{p}_t\,(C_t) - \frac{\bar{F}\,(\bar{p}_t\,(C_t))}{f\,(\bar{p}_t\,(C_t))} = w, \tag{EC.28}$$

or equivalently as

$$\bar{p}_t\,\frac{\varepsilon\,(\bar{p}_t)-1}{\varepsilon\,(\bar{p}_t)} = w. \tag{EC.29}$$

It turns out that the problem (EC.26) indeed has a unique solution which follows from our IGFR assumption on the valuation distribution. To see this, recall that the IGFR assumption is equivalent to increasing price elasticity, which implies that the left-hand-side of (EC.29) is increasing. Following this observation, we can explicitly characterize the solution of (EC.26) as follows.

Remark EC.2. It follows from Proposition 2 that the first order optimal price $\hat{\bar{p}}_t := \bar{p}_t(\hat{C}_t)$ is independent of the regular market size Λ_1 and the free trial market size Λ_2 (and only depends on the valuation distribution and w). Since $\bar{F}(\hat{\bar{p}}_t) = \frac{\hat{C}_t\mu-\Lambda_2}{\Lambda_1+\Lambda_{2,t}}$, the ratio $\frac{\hat{C}_t\mu-\Lambda_2}{\Lambda_1+\Lambda_{2,t}}$ is also independent of the free trial market size Λ_2, and consequently, the first order optimal capacity $\hat{C}_t$ increases as the free trial market size Λ_2 increases.

It remains to justify (EC.25) and to establish the theoretical performance of our proposed approximate solution for large scale systems. To do this, we again consider a sequence of systems indexed by n; but since the capacity is now a decision variable, we only scale the arrival rates, and let Λ_1^n and Λ_2^n increase according to

$$\Lambda_1^n := \eta_1 n, \text{ and } \Lambda_2^n := \eta_2 n, \tag{EC.30}$$

where $\eta_1 := \frac{\Lambda_1}{\Lambda_1+\Lambda_2}$ and $\eta_2 := \frac{\Lambda_2}{\Lambda_1+\Lambda_2}$. Note that the original system is still embedded in this sequence and can be recovered by setting $n = \Lambda_1 + \Lambda_2$. Again, we do not scale the individual customer characteristics and let the mean service requirement $1/\mu$, the valuation cdf $F(\cdot)$, and the delay cost q remain fixed throughout the sequence. Let $\hat{P}_t^n := P_t^n(\hat{C}_t^n, \hat{p}_t^n)$ and $P_t^{\star,n} := P_t^n(C_t^{\star,n}, p_t^{\star,n})$ denote the profit rate under our proposed solution and the optimal profit rate for a system with size n, respectively. The following result establishes that our proposed policy achieves the performance of an optimal policy asymptotically, and that the optimal system operates under the QED heavy traffic regime. It is worthwhile to point out that the elasticity condition in Assumption 1 is no longer assumed, rather it is shown that condition is to be satisfied by optimal systems that are large in scale.

Theorem EC.4. (Optimality of QED under Joint Optimization) *Suppose that the regular and free trial market sizes scale according to* (EC.30)*. Then, under joint price and capacity optimization, the optimal system operates under the quality-efficiency driven (QED) heavy traffic regime and the capacity and price pair* $\left(\hat{C}_t^n, \hat{p}_t^n\right)$ *is asymptotically optimal, i.e.,*

$$\frac{\hat{P}_t^n}{P_t^{\star,n}} = \frac{P_t^n(\hat{C}_t^n, \hat{p}_t^n)}{P_t^n(C_t^{\star,n}, p_t^{\star,n})} \to 1 \ \ as \ \ n \to \infty. \tag{EC.31}$$

Remark EC.3. (Relation to the Value of Offering Free Trials) Note that Theorem EC.4 also implies that QED is the optimal regime for a large service system *without* free trials, i.e., when $\Lambda_2 = 0$. As a result, Theorem 4 also yields an important insight for service firms that set the initial price and capacity optimally, and then, have the option of offering free trials without the ability to change the capacity further: Since the optimal capacity is chosen such that the system operates in QED, it is the analysis of Section EC.1.1 that is relevant, and the benefit of offering free trials is indeed limited. Equivalently, we expect a service firm to benefit significantly from offering free trials as observed in Section EC.1.2 only when the initial capacity is not chosen optimally.

Proof of Proposition 2: Let $\hat{P}'(\bar{p}) := \hat{P}'(\bar{p}(C))$ denote the derivative of $\hat{P}(C)$ with respect to C

expressed as a function of $\bar{p}$. Since F is IGFR, $\varepsilon(\bar{p})$ is an increasing function of $\bar{p}$. Then, it follows that

$$\hat{P}'(\bar{p}) = \bar{p}\frac{\varepsilon(\bar{p})-1}{\varepsilon(\bar{p})} - w$$

is also increasing in $\bar{p}$, or equivalently $\hat{P}'(C)$ is decreasing in C. Hence, there exists a unique $\hat{C}$ and a corresponding $\bar{p}(C)$ that solves (EC.26).

Next, suppose that $v_{min} - \frac{\bar{F}(v_{min})}{f(v_{min})} < w$. If $v_{max} = \infty$, it follows from Theorem 2 in Lariviere (2006) that $\lim_{\bar{p}\to\infty} \varepsilon(\bar{p}) > 1$, and so

$$\hat{P}'(\bar{p}) = \bar{p}\frac{\varepsilon(\bar{p})-1}{\varepsilon(\bar{p})} - w \to \infty \text{ as } \bar{p} \to \infty.$$

If $v_{max} < \infty$, it follows that $\limsup_{\bar{p}\to v_{max}} \frac{f(\bar{p})}{\bar{F}(\bar{p})} = \infty$, and so $\lim_{\bar{p}\to v_{max}} \varepsilon(\bar{p}) = \infty$. Then, we have

$$\lim_{\bar{p}\to v_{max}} \hat{P}'(\bar{p}) = \lim_{\bar{p}\to v_{max}} \bar{p}\frac{\varepsilon(\bar{p})-1}{\varepsilon(\bar{p})} - w = v_{max} - w > 0.$$

Finally, at $\bar{p} = v_{min}$,

$$\hat{P}'(v_{min}) = v_{min} - \frac{\bar{F}(v_{min})}{f(v_{min})} - w < 0.$$

Thus, $\hat{P}'(\bar{p})$ changes sign and the monotonicity of $\hat{P}'$ shows that there exists a unique $\hat{\bar{p}} \in (v_{min}, v_{max})$ such that $\hat{P}'(\hat{\bar{p}}) = 0$ and corresponding $\hat{C}$.

To prove part (ii), we suppose $v_{min} - \frac{\bar{F}(v_{min})}{f(v_{min})} \geq w$. Then, $\hat{P}'(v_{min}) = v_{min} - \frac{\bar{F}(v_{min})}{f(v_{min})} - w \geq 0$ and the monotonicity of $\hat{P}'$ implies that

$$\hat{P}'(\bar{p}) = \bar{p} - \frac{\bar{F}(\bar{p})}{f(\bar{p})} - w > 0,$$

for all $\bar{p} > v_{min}$. Thus, $\hat{P}'(C)$ is increasing in C, and consequently $\hat{\bar{p}} = v_{min}$. ■