Revenue optimization for a make-to-order queue in an uncertain market environment

Omar Besbes* Constantinos Maglaras†

Submitted: 11/2006, Revised: 08/2007, 02/2008
To appear in *Operations Research*

Abstract

We consider a revenue maximizing make-to-order manufacturer that serves a market of price and delay sensitive customers and operates in an environment in which the market size varies stochastically over time. A key feature of our analysis is that no model is assumed for the evolution of the market size. We analyze two main settings: i) the size of the market is observable at any point in time; and ii) the size of the market is not observable and hence cannot be used for decision-making. We focus on high-volume systems that are characterized by large processing capacities and market sizes, and where the latter fluctuate on a slower time scale than that of the underlying production system dynamics. We develop an approach to tackle such problems that is based on an asymptotic analysis and that yields near-optimal policy recommendations for the original system via the solution of a stochastic fluid model.

Keywords: Revenue management, dynamic pricing, market uncertainty, queueing, state dependent queues, asymptotic analysis.

1 Introduction

An information service provider -such as a firm offering a software application or video content on demand- serves a market of price and delay sensitive consumers. Similarly, a make-to-order manufacturer sells customized products with specified delivery leadtimes that themselves affect the overall demand faced by the firm. In both examples, and assuming that the respective firms enjoy some degree of pricing power, the goal is to select the product pricing strategy to maximize their profitability in a way that optimally balances their revenues with the congestion externalities that arise when the corresponding service delivery or production systems become overloaded. This classic problem has been studied extensively in the literature. In most cases, this body of work has assumed that the firm has perfect information of the market and demand characteristics. For example, it is typical to assume that orders arrive according to a non-homogeneous Poisson process with instantaneous rate at time t given by $\Lambda f(p(t), d(t))$, where Λ is the total market size,

*The Wharton School, JMHH, 3730 Walnut Str., Philadelphia, PA 19104. (obesbes@wharton.upenn.edu)
†Columbia Business School, 409 Uris Hall, 3022 Broadway, NY, NY 10027. (c.maglaras@gsb.columbia.edu)

which is assumed constant, and $f(p(t), d(t))$ is a known market response function that gives the fraction of all potential buyers that choose to place an order when the posted price is $p(t)$ and the target leadtime is $d(t)$ time units. (See, e.g., [4] and references therein.)

In many practical settings, however, firms have imperfect knowledge of the market response functions and market sizes they face. Typically, the former is addressed by fitting a particular demand function to past demand observations at different (price, leadtime) offered combinations; this can be achieved with a fairly high degree of accuracy, at least locally around a price range of interest. The latter is captured through the firm's forecast. Since forecasts often neglect various exogenous and uncontrollable factors such as unpredicted competitive actions, weather phenomena, political changes, etc., the end result is that the actual market size may exhibit significant stochastic fluctuations around the nominal forecasted trend. This may result in severe exposure for the firm both in terms of financial and operational performance. The objective of this paper is to propose a general approach for deriving pricing policies that yield near-optimal performance in the presence of such stochastic, inter-temporal market fluctuations. In particular, the recommendations aimed at are simple, intuitive and "model free" with respect to the market size evolution. In addition, we seek to understand the extent and significance of this market size time variability on congestion and profitability.

We adopt the stylized $M/M/1$ queue as a model of a make-to-order firm that serves a market of price and delay sensitive customers, whose size $\Lambda(t)$ varies stochastically over time, noting that the analytical framework that we will propose in the sequel is directly applicable to more complex production and service systems. We assume that customers make their purchase decisions based on the prevailing price and leadtime at the time of their arrival to the system, and that the firm selects its price in a dynamic (or static) manner, to optimize its expected profitability.

To clarify the settings analyzed in the paper, it is worth putting them into perspective. In particular, one could define a hierarchy of settings as follows: a) $\Lambda(\cdot)$ is constant and known; b) $\Lambda(\cdot)$ is an observable continuous time Markov chain, with, say, k possible states $\{\Lambda_1, ..., \Lambda_k\}$; c) $\Lambda(\cdot)$ is an observable Markov process with a continuum of states; d) $\Lambda(\cdot)$ is observable but no model is available for its underlying dynamics; and e) $\Lambda(\cdot)$ is *not* observable and no model is available for its underlying dynamics. In cases a) and b), one could formulate the problem as a classical Markov Decision Process (MDP), that would be one-dimensional for a) and two-dimensional for b). Then, the MDP could be analyzed to derive structural results of optimal solutions (see, e.g., [4]). Note that as the number of possible market size states (k) increases, the model parameters would be hard to estimate. This motivates the consideration of continuous time models (c)), addressed through appropriate continuous time control problems. The primary focus of this paper are cases d) and e), i.e., when $\Lambda(\cdot)$ is an exogenous process with unknown dynamics and is not necessarily observable. Without a precise model for the evolution of the market size process $\Lambda(\cdot)$ it is unclear what the appropriate modeling approach and objective criterion should

be. The main contributions of this paper relate to this problem. Specifically, we propose an analytical approach for this problem and quantify the structural differences between near-optimal policies that may or may not have real-time market information. The optimal pricing policies are dynamic. Our analysis yields simple and practical insights, as well as provably good policy prescriptions. Finally, our analysis also provides insights for the simpler cases a)-c) that could potentially be addressed via an MDP formulation.

Our study is motivated by applications that involve high volumes of demand and production, and where the market size fluctuations, while potentially significant, vary slowly relative to the underlying production dynamics. For example, a typical make-to-order automobile manufacturer[1] may produce in total thousands of cars for all of its models per day, have a backlog of existing orders that is equivalent to a few weeks worth of production, and observe major changes in the underlying market for automobiles every several months or so. Focusing on systems that are described by the above two conditions, we develop an asymptotic characterization of their behavior as the production capacity and the potential market size grow large that highlights the "macro" time scale of the market size process. This builds on the recent idea of multi-scale fluid limits of Bassamboo et al. [6, 7]. Intuitively, viewed on the time-scale of the market size fluctuations, the fast dynamics of the queue will always settle into an "equilibrium" state, and the overall system behavior will be one that moves from one equilibrium state to another as the market size process evolves, giving rise to a stochastic fluid limit model. The precise setting in [6] differs from ours in two important aspects: first, [6] considers a many-server limit, whereas we focus on a setting where the number of servers is fixed (and equal to one); and second, having a handle on the integrated queue length is not sufficient in our setting to analyze the system dynamics and ultimately the objective, which stems in part from the non-linearity present in the current problem. To the best of our efforts, these differences do not allow one to generalize or adapt the methodology developed in [6]. As a result, we develop a proof technique that relies on an extension of a powerful framework advanced by Bramson [9], which is applicable to general models of multiclass queueing networks; this alternate approach than that taken in [6], that appears necessary in our setting, is one of the methodological contributions of our paper.

Adopting the limiting stochastic fluid model as a representation of the system dynamics, we pursue to characterize the dynamic pricing policies that maximize the system's revenues. The optimal fluid dynamic pricing policy admits a simple and intuitive structure (Proposition 2): at any point in time the firm should apply a price that is given by the pointwise maximum between the price that would optimize the instantaneous revenue rate in the absence of the capacity constraint, and the price that would induce full resource utilization; this is reminiscent of the static pricing

[1]The make-to-order model has gained significant ground in the automobile industry. For example, BMW claims that 80% of the cars sold in Europe and 30% of those sold in the US are built to order. When a dealer inputs a potential order to BMW's web ordering service, a target leadtime is generated within five seconds. This is typically 11 to 12 days in Europe and about double that amount in the US [14].

heuristic of Gallego and van Ryzin [15]. The first of these components is constant, while the second one will vary over time in accordance to the market size fluctuations. Since the market evolves on a slow time scale, the corresponding price adjustments will be relatively infrequent, which is practically appealing. We subsequently apply this policy in the case of observable market size and show that this pricing heuristic is near-optimal in the regime considered (Proposition 4).

The intuition gained from the stochastic fluid model and the associated optimal policy can be used in the case of unobservable market size. In particular, we derive policies in the latter setting that indirectly track changes in the market size through changes in congestion. The latter is done by adjusting the posted price dynamically (either continuously or at discrete points in time) in response to the current level of congestion; if the queue length is positive, the price increases linearly, and if the queue is empty, the price decreases linearly until it reaches the capacity-unconstrained revenue maximizing price. We show that the derived policies achieve the same asymptotic performance as the best possible one when the market size is observable (Proposition 5). An extensive set of numerical results provides an illustration of the above findings.

The remainder of the paper is organized as follows: this section concludes with a brief literature review. §2 describes the model, and its stochastic fluid model approximation is derived in §3. §4 studies the problem of profit maximization under dynamic pricing in the case where the market size is observable. §5 focuses on the case of unobservable market size. §6 reports on an extensive set of numerical experiments and offers some concluding remarks. The proofs of the main results are collected in Appendix A, while Appendix B contains the proofs of auxiliary lemmas.

Literature Survey: This paper builds on an extensive literature on the economic analysis of queues that originated with the seminal work of Naor [30]. The economic framework that we adopt here was first proposed by Mendelson [28], and further extended by Mendelson and Whang [29]. An extensive review of known results on customer behavior in queueing environments can be found in the book by Hassin and Haviv [19].

In contrast to the above papers that focused on social welfare optimization, our work looks at the revenue maximization objective. There is a growing set of papers that focus on the latter. The closest one to ours is the paper by Maglaras and Zeevi [26] that showed that the revenue maximizing price in a high-volume single product system induces the so-called heavy-traffic. Afèche [1] studied the revenue maximization for an $M/M/1$ queue for a market with two types of price and delay sensitive customers. Both [26, 1], as well as most of the references therein, assume that customers respond to steady-state delay estimates (as in [28, 29]), instead of real-time delay quotations considered in our work.

Papers that have studied dynamic pricing in queues include Low [24] and Chen and Frank [12] using exact analysis, and Kachani and Perakis [20], Kleywegt [21], Maglaras [25] and Çelik and Maglaras [11] using fluid and diffusion model approximations; all of these references except [11] assume that congestion costs are incurred by the firm and not the customers. Finally, Ata and

Schneorson [4] studied using an MDP formulation the problem of dynamic arrival and capacity control in an $M/M/1$ queue.

All of the papers listed above assume perfect knowledge of the demand function and market size, which is taken to be constant over time. The effect of a stochastically varying market size process, almost exclusively under the assumption of a Markov modulated process, has been studied substantially, but in all cases that we know this did not involve any pricing decisions. (See, e.g., [32] for an inventory context and [16] for a production setting.) The specification of our market size uncertainty is closer to the one adopted in Bassamboo et al. [7] and Steckley et al. [33]. Both looked at staffing and routing problems in call centers that face uncertain demands.

In terms of analysis, our work adopts the multi-scale fluid limit framework proposed by Bassamboo et al. [7], which is related to the stationary approximation heuristic proposed by Green and Kolesar [17] for analyzing multiserver systems facing Poisson arrivals with time-varying intensity. Our proof technique relies on Bramson's [9] framework for proving state-space collapse results in multiclass queueing networks; see also Dai and Tezcan [13] for a recent application of these tools. The real time delay notification results in a queue with state-dependent parameters, which was studied by Mandelbaum and Pats [27].

An early reference that focuses on learning parameters in a non-stationary environment is Little [23] in the context of promotional spending. Finally, a related issue that we do not address in this paper is that of estimating the demand function itself. There is a growing literature on revenue management problems with or without queueing effects that involve such considerations; see, e.g., Aviv and Pazgal [5], Afèche and Ata [2], Araman and Caldentey [3] and Besbes and Zeevi [8], as well as the references therein.

2 Model

We will first describe a dynamic pricing problem setting for a queue that operates in an environment with a stochastically varying market size.

2.1 General setting

We model a make-to-order firm offering a single product or service as an $M/M/1$ queue. We assume that the firm initially sets its service rate at μ and can dynamically adjust the price of the product $p(t)$ at time t. Potential customers arrive according to a non-homogeneous Poisson process with instantaneous rate $\Lambda(t)$, which can be interpreted as the market size at time t. It is typical to assume that $\Lambda(t)$ is constant for all times t (see, e.g., [4] and references therein). In contrast, our model will allow this process to vary stochastically over time in an effort to capture the uncertainty that results from the imperfections of forecasting techniques as well as the presence of exogenous uncontrollable factors that may affect the market size at different

points in time. More precisely, we assume that the market size $\{\Lambda(t), t \geq 0\}$ is a positive, almost everywhere continuous stochastic process, independent of the firm's controls and the state or the evolution of the system in question. The focus of this paper is to study the effect of this market size uncertainty on the behavior of the system, the structure of its optimal pricing policy, and its overall profitability. Examples of models for $\Lambda(t)$ are detailed in §6.1.1.

Each customer arrives with a private valuation (or willingness-to-pay) for the product, v, which is an independent draw from a distribution function F, which is known by the firm. Upon arrival to the system at time t, a customer is told the prevailing price $p(t)$ for the product and the expected time it will take until his service starts $Q(t)/\mu$, where $Q(t)$ denotes the number of customers already in the system at time t. Based on this information, the customer computes his perceived cost as $p(t) + cQ(t)/\mu$. We assume that c, the sensitivity of customers to congestion is positive and constant across customers[2]. A customer with valuation v will purchase the product if and only if $v \geq p(t) + cQ(t)/\mu$. Customers join a single queue and are processed in the order of their arrival (FIFO discipline). Let $\bar{F}(\cdot) := 1 - F(\cdot)$. Given the above discussion, the instantaneous arrival rate at time t, when the price is $p(t)$ is

$$\lambda(t) = \Lambda(t)\mathbb{P}\left(v \geq p(t) + \frac{c}{\mu}Q(t)\right) = \Lambda(t)\bar{F}\left(p(t) + \frac{c}{\mu}Q(t)\right).$$

Assumption The willingness-to-pay has support $[0, v_{max}]$, where $0 < v_{max} < \infty$, and $\bar{F}$ is Lipschitz continuous and decreasing on $[0, v_{max}]$.

System dynamics. The arrival process to the firm is denoted by $\{A(t), t \geq 0\}$, and is a non-homogeneous Poisson process with instantaneous rate $\lambda(t), t \geq 0$; i.e., $A(t)$ is equal to the number of arrivals in $[0, t]$. Let $D(t)$ denote the number of service completions in $[0, t]$, $\mathscr{T}(t)$ the cumulative time allocated in processing jobs up to time t and $Y(t)$ the cumulative idleness up to time t. Let $S(t)$ be a Poisson process with rate μ. The dynamics of the queue length process $Q(t)$ are governed by the following equations:

$$Q(t) = Q(0) + A(t) - D(t), \tag{1}$$

$$D(t) = S(\mathscr{T}(t)), \tag{2}$$

$$\mathscr{T}(t) + Y(t) = t, \tag{3}$$

$$Y(t) \text{ cannot increase unless } Q(t) = 0. \tag{4}$$

[2]The latter assumption can be relaxed, cf. discussion following Theorem 1.

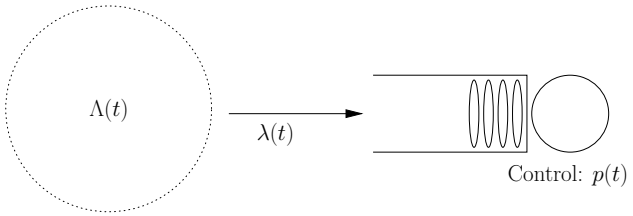

Figure 1: Model sketch

2.2 The economic objective

We first consider a baseline model where the market size $\Lambda(\cdot)$ is an observable and fully characterized stochastic process.

Baseline model: [$\Lambda(\cdot)$ observable, known model for $\Lambda(\cdot)$ dynamics] The firm has two controls. The first is whether to idle the server when there are orders waiting in queue to be processed; the second is to dynamically select its posted price in an effort to control congestion, adjust to changes in market size, and optimize its revenues.

For purposes of our analysis we will restrict attention to pricing policies that are functions of $Q(t)/\mu$ and $\Lambda(t)/\mu$, where $Q(t)/\mu$ is the expected workload embodied in all orders in queue at time t, and $\Lambda(t)/\mu$ is the normalized market size: $p(\cdot,\cdot) : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$. The market size information, $\Lambda(t)$, is important since the firm has limited capacity, and needs to adjust its pricing in order for its target market share (i.e., its potential demand in the absence of congestion) to be close to its processing capabilities. We come back to this point shortly. We will also make the following technical assumption:

Assumption $p(\cdot,\cdot)$ is non-decreasing with respect to its first argument and Lipschitz.

We denote $\mathcal{P}$ the set of pricing policies satisfying the above.

In this setting, the firm's revenue optimization problem is to select a dynamic pricing strategy $p \in \mathcal{P}$ and server allocation policy $\mathscr{T}(t)$ to maximize

$$\mathbb{E}\left[\int_0^T p\Big(\frac{Q(s)}{\mu}, \frac{\Lambda(s)}{\mu}\Big)\, dA(s)\right],$$

where T denotes the time horizon. It is easy to show that in our single-product setting it is optimal to adopt a server allocation policy $\mathscr{T}(t)$ that is non-idling. Using results for intensity control problems (see [10]) we can rewrite the dynamic pricing control problem as

$$\max_{p\in\mathcal{P}} \mathbb{E}\left[\int_0^T p\Big(\frac{Q(s)}{\mu}, \frac{\Lambda(s)}{\mu}\Big)\, \Lambda(s)\, \bar{F}\left(p\Big(\frac{Q(s)}{\mu}, \frac{\Lambda(s)}{\mu}\Big) + \frac{c}{\mu}Q(s)\right) ds\right]. \tag{5}$$

The policies considered above are implicitly allowed to use the dynamics of $\Lambda(t)$ in deciding how

to price at any given time. From a practical viewpoint, it is not clear how a firm would have access to such a model and how much it would decide to rely on it. With this in mind, the focus is on investigating settings where no such description is available and the firm only observes the queue length and potentially the instantaneous market size.

Variant 1: [$\Lambda(\cdot)$ observable, no model for $\Lambda(\cdot)$ dynamics] We investigate in §4 what strategies can the firm adopt when it has access to the value of $\Lambda(\cdot)$ at any point in time but no access to a model of the market size dynamics.

Variant 2: [$\Lambda(\cdot)$ unobservable, no model for $\Lambda(\cdot)$ dynamics] As it will turn out later on, our proposed solution to the dynamic pricing problem for variant 1 does indeed make use of the information on the market size values. From a practical viewpoint, however, the firm may not be able to observe the prevailing market size in real time so as to price accordingly. This case is studied in §5 where the pricing decision at time t is now only allowed to be a function of the queue length $Q(t)$ and the prevailing price $p(t^-)$. In particular, we seek to understand what performance is achievable in such cases.

Comments on modeling assumptions and proposed approach. The assumption that $v_{max} < \infty$ is only made to simplify some arguments and is not necessary for the analysis. In the current study, we restrict attention to policies in $\mathcal{P}$. The main reason for that is to obtain some form of tractability that can ultimately lead to a stochastic fluid model. The restriction on $\mathcal{P}$ is only a sufficient condition for the analysis to come to hold and it is likely that the results we derive could be appropriately extended to a broader set of policies, but at the expense of significant additional technical complexity.

3 Fluid approximation

It seems difficult to tackle the variants presented above through exact analysis. This section develops an asymptotic approximation for the dynamics of the system described above in settings where: a) the processing capacity and market size are large, and b) the market size process varies slowly relative to the production dynamics of the queue. The latter assumption essentially says that the time it takes the server to process one order or even to clear a modest backlog of orders is small relative to the time it takes for the market size to change significantly. For example, if the market size evolved according to a bi-level continuous time Markov chain, then a set of parameter values that would be in line with these assumptions would be a processing rate of 100 orders per day, two market size levels that are 200 and 600 orders per day, respectively, and equal transition rates between these two levels that are equal to .1 transitions per day; it roughly takes 2 weeks for the market to switch between the high and low demand regimes.

The resulting approximation given in §3.2.2 is a stochastic fluid model, which we will use in §4 and §5 for purposes of performance analysis and profit optimization.

3.1 Setup

To quantify the characteristics mentioned above we will study a sequence of systems, indexed by r, with increasing processing capacity and market size, which highlights the time scale separation between the market size and the production dynamics. In particular, we let $\{a_r, r \geq 1\}$ be an increasing sequence such that $\lim_{r\to\infty} a_r = \infty$ and $\lim_{r\to\infty} a_r/r = 0$ and consider the following set of parameters:

$$F^r(\cdot) = F(\cdot), \quad T^r = a_r T, \quad \Lambda^r(\cdot) = r\Lambda\left(\cdot/a_r\right) \quad \text{and} \quad \mu^r = r\mu. \tag{6}$$

That is, the underlying willingness-to-pay distribution F (i.e., the customer characteristics) remains unchanged, the processing capacity and market size grow proportionally large as a function of the scaling parameter r, and the market size dynamics are slowed down by the factor a_r. In particular, when the system has evolved over t units of time, the clock of the market size process changed only by t/a_r and hence as r increases, the market size will appear as varying on a slower time scale from the perspective of the clock associated with the evolution of the queue. Conversely, in order for the market size process to evolve over t time units over which some fluctuations may be observed, the queue will have evolved over the much larger $a_r t$ time interval. This regime is motivated by the work of Bassamboo et al. [6], where for the intuitive reasons given above, they refer to the resulting limiting model as a Pointwise-Stationary-Approximation (PSA). For example, the parameters of the continuous time Markov chain model given earlier can be embedded in the sequence of systems with $a_r = \sqrt{r}$, $\mu^r = r$, two levels $\Lambda_1^r = 2 \cdot r$, $\Lambda_2^r = 6 \cdot r$, and a transition rates of $\nu^r = 1$; $r = 100$ recovers the parameters of the system of original interest.

We will study the behavior of the system for large r for any dynamic pricing policy $p \in \mathcal{P}$, which is a function of the workload and normalized market size at time t, $p(Q(t)/\mu, \Lambda(t)/\mu)$. To highlight the effect of the market size evolution on the structure of the optimal pricing policy and the resulting system behavior, we will focus on the time scale on which the market size process exhibits significant fluctuations, and on the expected revenue criterion over $[0, a_r T]$ given by

$$\begin{aligned} R^r(T) &= \mathbb{E}\left[\int_0^{a_r T} p\left(\frac{Q^r(s)}{\mu^r}, \frac{\Lambda^r(s)}{\mu^r}\right) \Lambda^r(s)\, \bar{F}\left(p\left(\frac{Q^r(s)}{\mu^r}, \frac{\Lambda^r(s)}{\mu^r}\right) + c\frac{Q^r(s)}{\mu^r}\right) ds\right] \\ &= r a_r \mathbb{E}\left[\int_0^{T} p\left(\frac{Q^r(a_r u)}{r\mu}, \frac{\Lambda(u)}{\mu}\right) \Lambda(u)\, \bar{F}\left(p\left(\frac{Q^r(a_r u)}{r\mu}, \frac{\Lambda(u)}{\mu}\right) + c\frac{Q^r(a_r u)}{r\mu}\right) du\right]. \end{aligned}$$

Defining for all $t \geq 0$, the normalized queue length

$$\tilde{Q}^r(t) = \frac{Q^r(a_r t)}{r}, \tag{7}$$

the expected revenues in the r^{th} system can be written as

$$R^r(T) = ra_r \, \mathbb{E}\left[\int_0^T p\left(\frac{\tilde{Q}^r(u)}{\mu}, \frac{\Lambda(u)}{\mu}\right) \Lambda(u) \, \bar{F}\left(p\left(\frac{\tilde{Q}^r(u)}{\mu}, \frac{\Lambda(u)}{\mu}\right) + c\frac{\tilde{Q}^r(u)}{\mu}\right) du\right]. \quad (8)$$

That is, $\tilde{Q}^r(t)$ is the queue length process evaluated on the slower time scale of the market size fluctuations driven by $\Lambda(t)$. In terms of the physical queueing system, the transition from $\tilde{Q}^r(t)$ to $\tilde{Q}^r(t+\delta t)$ corresponds to an evolution of the queue over $a_r\delta t$ time units. The remainder of this section is dedicated to constructing an asymptotic approximation for $\tilde{Q}^r(t)$ (cf. Theorem 1).

3.2 General analysis

Broad overview. The analysis of the behavior of $\tilde{Q}^r(t)$ as r grows large relies on a framework developed by Bramson [9] to establish state space collapse results in stochastic networks operating in heavy traffic. While our setting falls outside the class of problems studied in [9], the main steps of his analysis can be extended to address our problem as well. The rough idea is that the horizon $[0, a_r t]$, which is relevant in studying the evolution of $\tilde{Q}^r(t)$, can be "constructed" by piecing together many shorter time intervals of length Δ. Each such interval is sufficiently long to observe significant fluctuations in the queue length, but too short in the time scale of the market size process ($\Delta/a_r \approx 0$) to observe any changes in the latter. This motivates the analysis of §3.2.1 that focuses on a fluid system where $\Lambda(t)$ is constant. The latter will suggest that on each short interval the queue length will reach a target configuration, and, moreover, this transient evolution will appear instantaneous in the longer time scale of $\Lambda(t)$. In §3.2.2, we use this result to "piece together" the queue length evolution over the longer time interval over which the market size process evolves.

3.2.1 Intermediate analysis on the "fast" time-scale of the queue

In this subsection we will assume that $\Lambda(t) = \bar{\Lambda}$ for all $t \geq 0$ for some constant $\bar{\Lambda} > 0$, and consider the following set of deterministic fluid equations:

$$q(t) = q(0) + \int_0^t \bar{\Lambda}\bar{F}\Big(p(q(s)/\mu, \bar{\Lambda}/\mu) + cq(s)/\mu\Big)ds - d(t), \quad (9)$$

$$d(t) = \mu\tau(t), \quad (10)$$

$$y(t) + \tau(t) = t, \quad (11)$$

$$y(t) \text{ cannot increase unless } q(t) = 0, \quad (12)$$

where q, d, τ, y are the fluid analogues of $Q, D, \mathscr{T}, Y$; the system of equations (9)-(12) is guaranteed to have a solution and the latter is unique under the assumptions on $\bar{F}$ and $p(\cdot,\cdot)$ (see [27, Proposition 3.1]). The asymptotic justification of this model is offered as part of the proof of the

main theorem of the next subsection.

Let $p_0 = p(0, \bar{\Lambda}/\mu)$ be the price under a candidate strategy $p(\cdot, \bar{\Lambda}/\mu)$ when the queue length is empty. The next proposition establishes the following intuitive result: a) if $\bar{\Lambda}\bar{F}(p_0)/\mu \leq 1$ (i.e., the system is over-capacitated), then $q(t) \to 0$ as $t \to \infty$; and b) if $\bar{\Lambda}\bar{F}(p_0)/\mu > 1$ (i.e., the system is under-capacitated), then $q(t) \to q^*$ as $t \to \infty$, where q^* is such that $\bar{\Lambda}\bar{F}(p(q^*/\mu, \bar{\Lambda}/\mu)+cq^*/\mu) = \mu$. Note that under the current assumptions that $\bar{F}$ is decreasing on $[0, v_{max}]$ and that $p(\cdot, \bar{\Lambda}/\mu)$ is non-increasing, q^* is unique. In the first case, the system drains its initial queue length and stays empty thereafter, whereas in the second case, the queue monotonically converges to a state that results in an arrival rate that exactly matches the system's processing capacity. In both cases, the limiting state can be expressed as a function of the market size $\bar{\Lambda}$. In particular, one can write the limit queue length configuration mentioned above as

$$q^* = h^{(p)}(\bar{\Lambda}) = \inf\left\{x \geq 0 : \ \bar{F}\big(p(x/\mu; \bar{\Lambda}/\mu) + (c/\mu)x\big) = \min\left\{\mu/\bar{\Lambda}, \bar{F}(p_0)\right\}\right\} \quad \text{for all } \bar{\Lambda} > 0, \tag{13}$$

where the superscript (p) emphasizes the dependence of the limit queue length configuration on the pricing policy $p(\cdot, \cdot)$.

Proposition 1 *Let q be the solution to (9)-(12) with $q(0) \leq M$ for some constant M. Then:*

i.) *Convergence:* $q(s) \to h^{(p)}(\bar{\Lambda})$ *as* $s \to \infty$.

ii.) *For all $\delta > 0$, there is a continuous function of $\bar{\Lambda}$, $s(\delta, M, \bar{\Lambda}) < \infty$ such that* $\left|q(s) - h^{(p)}(\bar{\Lambda})\right| < \delta$ *for all* $s \geq s(\delta, M, \bar{\Lambda})$.

3.2.2 Stochastic fluid model approximation on the $\Lambda(t)$ time-scale

We are now in position to present the main result of this section, which essentially establishes that the $\tilde{Q}^r(t)$ process appears as if it is always in its target state configuration $h^{(p)}(\Lambda(t))$, which is a result of "piecing" together teh evolution of $\tilde{Q}^r(t)$ over many "short" intervals.

In the remainder of the text, for any event $\mathcal{A}$, $\mathbb{P}_\Lambda(\mathcal{A})$ will denote $\mathbb{P}(\mathcal{A} \,|\, \{\Lambda(t) : 0 \leq t \leq T\})$ and for any random variable X, $\mathbb{E}_\Lambda[X]$ will denote the conditional expectation $\mathbb{E}[X|\{\Lambda(t) : 0 \leq t \leq T\}]$.

Theorem 1 *Fix a policy $p(\cdot, \cdot) \in \mathcal{P}$. Suppose $Q^r(0)/r \leq M$ for some $M > 0$. Then for any positive and continuous market size trajectory $\{\Lambda(t) : 0 \leq t \leq T\}$ and for all $\epsilon > 0$, we have*

$$\mathbb{P}_\Lambda\Big\{\sup_{0<t\leq T}\left|\tilde{Q}^r(t) - h^{(p)}(\Lambda(t))\right| > \epsilon\Big\} \to 0 \qquad \text{as } r \to \infty, \tag{14}$$

where $h^{(p)}(\cdot)$ is defined in (13).

In the limit, the dynamics of the queue degenerate to a simple function of the market size $\Lambda(t)$. Conversely, knowing the queue length position in large scale systems provides partial information

about the current market size value. We come back to implications of this point when designing pricing policies in §5.

Remarks: i.) In Theorem 1, we assumed that $\Lambda(\cdot)$ is continuous. A corollary is that when $\Lambda(\cdot)$ has a finite number of discontinuities on $[0,T]$, say at $t_1,...,t_k$, then $\sup_{t\in(0,T]\setminus\{t_1,...,t_k\}}\left|\tilde{Q}^r(t)-h^{(p)}(\Lambda(t))\right|$ converges to zero in $\mathbb{P}_\Lambda$-probability as $r\to\infty$. ii.) Our results extend to the case where customers are differentiated both in terms of their valuation and delay sensitivity parameter, c; i.e., (v,c) is a random variable characterizing each arriving customer. The proportion of customers joining the system now depends on the joint distribution of (v,c).

4 Case 1: $\Lambda(\cdot)$ observable

The problem where the market size is observable will act as a benchmark for our subsequent analysis in §5 where the $\Lambda(\cdot)$ process is unobservable and the firm has no model for its potential evolution. The stochastic limit derived above for $\tilde{Q}^r(t)$ gives rise to a stochastic fluid model whose dynamics are driven by the market size process $\Lambda(t)$. The queue length position, $h^{(p)}(\Lambda(t))$, is characterized through the flow balance equation

$$\Lambda(t)\bar{F}\left(p\Big(\frac{h^{(p)}(\Lambda(t))}{\mu},\frac{\Lambda(t)}{\mu}\Big)+c\frac{h^{(p)}(\Lambda(t))}{\mu}\right)=\min\left\{\mu,\Lambda(t)\bar{F}\left(p\Big(0,\frac{\Lambda(t)}{\mu}\Big)\right)\right\};\tag{15}$$

i.e., whenever $\Lambda(t)\bar{F}(p(0,\Lambda/\mu))\le\mu$, the server is under-utilized and $h^{(p)}(\Lambda(t))=0$, whereas in all other cases, the queue length increases such that the resulting congestion effect will reduce the demand to equal the processing capacity μ.

4.1 The dynamic pricing fluid heuristic

Expression (15) yields that the instantaneous arrival rate is $\lambda(t)=\min\{\mu,\Lambda(t)\bar{F}(p(0,\Lambda(t)/\mu))\}$, which, in turn, results in the following approximating revenue maximization problem:

$$\max_{p\in\mathcal{P}}\mathbb{E}\left[\int_0^T p\Big(\frac{h^{(p)}(\Lambda(t))}{\mu},\frac{\Lambda(t)}{\mu}\Big)\min\left\{\mu,\Lambda(t)\bar{F}\left(p\Big(0,\frac{\Lambda(t)}{\mu}\Big)\right)\right\}du\right],\tag{16}$$

where $\mathcal{P}$ is the set of admissible policies defined in §2.2. In the remainder of the text, we suppose that $x\mapsto x\bar{F}(x)$ is a unimodal function with maximizer p^*.

Proposition 2 *For all $l>0$, Let $p^\mu(l)$ be the unique solution of $\bar{F}(\cdot)=\min\{1/l,1\}$ on $[0,v_{max}]$. Then, the optimal solution of (16) is given by*

$$\hat{p}(q/\mu,\Lambda(t)/\mu)=\max\{p^*,p^\mu(\Lambda(t)/\mu)\}.\tag{17}$$

In addition, the corresponding queue length is given by $h^{(\hat{p})}(\Lambda(t))=0$ for all $t>0$.

The dynamic pricing heuristic given above is intuitive: if the capacity is ample, the firm charges a static (monopoly) price that maximizes revenues per unit time, p^*; if the capacity is scarce, i.e., the capacity constraint is violated under the monopoly price, then the optimal price $p^\mu(\Lambda(t)/\mu)$ is higher and ensures that the capacity constraint is met while customers do not experience any congestion. This structure resembles the pricing policy derived by Gallego and van Ryzin [15] in their single product dynamic pricing problem. The monopoly price only depends on the valuation distribution F, while the price that induces full resource utilization $p^\mu(\Lambda(t)/\mu)$ also depends on the market size process $\Lambda(t)$. An important insight of the above result is that when $\Lambda(\cdot)$ is observable, the firm would not use information that can be extracted from the model that describes the market size dynamics in its pricing policy. Such a dependence would manifest itself in the solution of (16) if the optimal price at time t was not solely a function of $\Lambda(t)$, but depended on its future values. This is a direct consequence of the firm's ability to change its price dynamically in response to the changing market conditions.

In the sequel, we will characterize the asymptotic performance of the pricing heuristic $\hat{p}$ derived above, and then modify it in §5 so as to construct a policy that achieves the same asymptotic performance without observing the market size $\Lambda(t)$.

4.2 Asymptotic performance guarantees

We first derive an upper bound on the performance of any dynamic pricing policy in $\mathcal{P}$. Recall the expression of the expected revenues generated by a policy p, $R^r(T)$, given in (8). In what follows, we assume that $\{\Lambda(s), 0 \le s \le T\}$ is positive and continuous. Define

$$R^r_\Lambda(T) := ra_r \mathbb{E}_\Lambda\Big[\int_0^T p^r(u)\Lambda(u)\bar{F}\big(p^r(u) + \frac{c}{\mu}\tilde{Q}^r(u)\big)du\Big] \quad \text{where} \quad p^r(u) := p\Big(\frac{\tilde{Q}^r(u)}{\mu}, \frac{\Lambda(u)}{\mu}\Big). \tag{18}$$

The superscript in $p^r(u)$ is only used to emphasize the dependence of the current price on $\tilde{Q}^r(u)$ and one should note that the mapping $p(\cdot,\cdot)$ is not changing with r. Note also that $R^r(T) = \mathbb{E}[R^r_\Lambda(T)]$, where the outer expectation is with respect to $\Lambda(\cdot)$.

Proposition 3 *For any policy $p(\cdot,\cdot) \in \mathcal{P}$,*

$$\limsup_{r\to\infty} \frac{R^r_\Lambda(T)}{ra_r} \le \int_0^T \sup_x \big\{x\ \min\{\mu, \Lambda(u)\bar{F}(x)\}\big\}du. \tag{19}$$

The next result shows that the policy $\hat{p}$ specified in (17) achieves the upper bound in (19).

Proposition 4 *Let $\hat{p}^r(u) := \hat{p}(\tilde{Q}^r(u)/\mu, \Lambda(u)/\mu)$ for all $u \ge 0$, where $\hat{p}$ was defined in (17). Then*

$$\lim_{r\to\infty} \mathbb{E}_\Lambda\Big[\int_0^T \hat{p}^r(u)\Lambda(u)\bar{F}\big(\hat{p}^r(u) + \frac{c}{\mu}\tilde{Q}^r(u)\big)du\Big] = \int_0^T \sup_x \big\{x\ \min\{\Lambda(u)\bar{F}(x), \mu\}\big\}du.$$

Together with the result of Proposition 3, this establishes that $\hat{p}$ asymptotically achieves the best possible performance in the parameter regime outlined in §3.

5 Case 2: $\Lambda(t)$ unobservable

We already pointed out that an important element of the problem studied in §4 above and of the resulting pricing heuristic $\hat{p}$ is that the firm can accurately observe the market size process $\Lambda(t)$ at every point in time. From a practical viewpoint, in addition to the difficulty the firm may have in trying to form a probabilistic model of the market size process, it may not be possible to detect in real time the prevailing market size. As we illustrate below, however, it is possible to construct a policy that can induce the market size by focusing on controlling the real time congestion in the system and exploiting the intuitive structure of the pricing rule $\hat{p}$.

5.1 The dynamic pricing fluid heuristic

Broadly speaking, whenever the queue length is positive, the firm should increase its price to reduce congestion and essentially find the point $p^{\mu}(\Lambda(t)/\mu)$ that matches the demand with the system's processing capacity. When the queue length is zero, then the firm should lower its price down to p^* or to the level that the queue length starts to grow again. In both cases, the firm never prices below p^*, which is the price that maximizes the revenue rate irrespective of the system's capacity constraint; this is consistent with (17). Recall that p^* is independent of $\Lambda(t)$, and is known to the firm.

In more detail, we will assume that the firm increases the price linearly at a rate of $\$\beta^+$ per unit time whenever the queue length is positive, and that it decreases the price at a rate of $\$\beta^-$ per unit time whenever the queue is empty and the prevailing price is above p^*. The actual price at any time t will depend on the cumulative effect of the aforementioned price changes, which can be succinctly expressed as follows:

$$\tilde{p}^r(t) = p^* + \delta^r(t), \qquad \tilde{p}^r(0) = p^*, \tag{20}$$

where we note that the initial condition $\tilde{p}^r(0)$ can be any price point greater or equal than p^* and

$$\delta^r(t) := x^r(t) - \inf_{0\le s\le t} x^r(s), \quad \text{for} \quad x^r(t) = -\beta^- Y^r(t) + \beta^+(t - Y^r(t)); \tag{21}$$

i.e., $x^r(t)$ increases whenever the server is busy and decreases whenever the server is idling, and the calculation of $\delta^r(t)$ simply corrects for the fact that the price is never allowed to drop below p^*. Note that $\delta^r = \Phi(x^r)$, where $\Phi(\cdot)$ is the one-sided regulator mapping (see Harrison [18, §2]).

5.2 Asymptotic performance guarantees

The next result characterizes the asymptotic performance of the policy $\tilde{p}$ in the parameter regime defined by (6).

Proposition 5 *Suppose one applies the pricing policy $\tilde{p}^r(\cdot)$ defined in (20)-(21). Then,*

$$\lim_{r\to\infty} \mathbb{E}_\Lambda\Big[\int_0^T \tilde{p}^r(u)\Lambda(u)\bar{F}(\tilde{p}^r(u) + \frac{c}{\mu}\tilde{Q}^r(u))du\Big] = \int_0^T \sup_x\big\{x\ \min\{\Lambda(u)\bar{F}(x), \mu\}\big\}du.$$

That is, the policy $\tilde{p}$ defined through (20)-(21) achieves the same asymptotic revenues as the ones achieved by $\hat{p}$ that assumed that the market size $\Lambda(t)$ was observable. This is accomplished by focusing the pricing decisions on controlling congestion together with some degree of price experimentation when the queue is empty; both of these, however, imply that the price changes occur at the "faster" time scale of the queue length dynamics. The practical implications of the latter are that the manager may now be changing prices daily rather than monthly, if s/he had access to the value of the market size at all times.

Remark. A question of interest pertaining to the policy $\tilde{p}$ is the choice of the parameters β^- and β^+. It is first important to note that for the policy $\tilde{p}$ to track indirectly the market size changes, it should only exhibit significant changes on the slow time scale. If β^+ and/or β^- were to be chosen too large, then the policy $\tilde{p}$ would be reacting to random queueing fluctuations rather than underlying changes in the market size. This implies that the price responsiveness parameters β^+ and β^- should be roughly of the same order than $1/a_r$. This allows to focus on the first order effects associated with the changes in the market size. We investigate numerically the impact of the parameters β^+ and/or β^- on performance in §6.

6 Numerical Examples and Concluding Remarks

This section reports a set of numerical experiments that compare the candidate policies described thus far and strive to offer some insights on a) the effect of market size variability on pricing, delays, and revenues; and b) the impact of not knowing the uncertain market size when making the various pricing decisions. Overall, our findings will highlight the value of the asymptotic approximation considered.

6.1 Setting of the experiments

6.1.1 Market size models

Throughout this section we will illustrate our results with two natural mathematical models for the market size process, which we briefly describe below. Λ_∞ will be used to denote a random variable with the steady-state distribution of the process (when it exists).

Bi-level (high-low) demand: A simple model for the market size uncertainty is to assume that it evolves as a continuous time Markov Chain (CTMC) over a predetermined discrete set of potential values. As an example, this approach was used in [32] to represent the evolution of an underlying state-of-the world influencing the level of the demand in a discrete time setting for an inventory problem. In our setting we could model the process $\Lambda(t)$ as a CTMC with state space $\{\Lambda_1, \Lambda_2\}$, transition rate ν, and transition probability matrix $P = [p_{ij}]$. This process has a steady state distribution: if the vector q is the solution to $qP = q$ and $q_1 + q_2 = 1$, then the limiting probabilities are given by (q_1, q_2).

One could generalize such a model by allowing the market size to take more than 2 levels. However, as the number of levels increases, the informational requirements of such a CTMC model grow substantially, and in those cases it may be more appropriate to use a more aggregated model.

A continuous model: We consider a model that allows the market size process to exhibit random variations around a nominal trend. This property in conjunction with the fact that the process should stay positive is typically required for interest rate models in the area of mathematical finance. In the vast literature on such models (see, e.g., [22]), a celebrated example is the Cox-Ingersoll-Ross (CIR) process that satisfies the following stochastic differential equation

$$d\Lambda(t) = \alpha(b(t) - \Lambda(t))dt + \sigma\sqrt{\Lambda(t)}dB(t), \tag{22}$$

where $\{B(t), t \geq 0\}$ is a standard Brownian motion, α and σ are positive constants and $b(\cdot)$ is a positive function. This is an example of an affine diffusion process: $b(\cdot)$ represents the anticipated "trend" (e.g., due to seasonal effects); α characterizes the speed at which the process reverts toward $b(\cdot)$, while σ quantifies the stochastic variability of the process. The term $\sqrt{\Lambda(t)}$ prevents the process from becoming negative. When $b(\cdot)$ is constant and $\Lambda(0) > 0$, the process stays positive under the condition that $2\alpha b \geq \sigma^2$. If $b(\cdot)$ is constant over time, then $\{\Lambda(t), t \geq 0\}$ has a steady-state distribution and Λ_∞ is distributed as $\frac{\sigma^2}{4\alpha}\chi^2_\nu$, where $\nu = 4b\alpha/\sigma^2$ and χ^2_ν denotes a chi-square random variable with ν degrees of freedom.

The parameters of both of these models can be estimated using past data through, e.g., maximum likelihood estimation since the likelihood ratios associated with these two models are available (see, e.g., [31]).

For the the purpose of numerical results, for the CTMC, we focus on bi-level models (with state-space $\{\Lambda_1, \Lambda_2\}$) and fix throughout the transition rate to $\nu = 1$. We consider the following transition matrix parametrized by a single parameter: $p_{11} = p_{22} = 1 - p_{12} = 1 - p_{21}$. With such a structure, p_{11} is a proxy for the "inertia" associated with the process and the limiting probabilities are just given by $P_1 = P_2 = 1/2$. Note that $\mathbb{E}[\Lambda_\infty] = (\Lambda_1 + \Lambda_2)/2$ and $\text{var}[\Lambda_\infty] = (\Lambda_2/2 - \Lambda_1/2)^2$ and hence $\Lambda_2 - \Lambda_1$ is a proxy for the variability of the process. For the CIR model, we will restrict attention to cases where the trend $b(\cdot)$ is constant and analyze the performance of the policies

introduced in the paper for various combinations of the parameters α, b, σ.

6.1.2 Policies

We introduce below the pricing policies that we will analyze in the current section. They are presented in ascending order of informational requirements.

i.) The *dynamic* pricing policy $\tilde{p}(\cdot)$ defined in (20)-(21). we will use the generic price responsiveness parameters $\beta^- = 0.5$ and $\beta^+ = 0.1$. We also analyze how this choice influences the performance of $\tilde{p}$ in Table 5. Recall that this policy does not require any description of the market size process dynamics.

ii.) The *dynamic* pricing policy $\hat{p}(\cdot)$ defined in (17). This policy only requires the current value of the market size.

iii.) The *static* pricing policy p^s that employs the optimal static price when $\Lambda(\cdot)$ is constant and equal to $\mathbb{E}[\Lambda_\infty]$; this is obtained via simulation. This price ignores variations in $\Lambda(\cdot)$ but has knowledge of the *true* mean of the market size.

iv.) The *static* pricing policy p^f obtained through the fluid analysis. This price solves the following revenue optimization problem:

$$\max_{p \geq 0} \quad p\, \mathbb{E}\left[\int_0^T \Lambda(s)\bar{F}\left(p + \frac{c}{\mu}Q(s)\right) ds\right]. \tag{23}$$

This static problem (23) could be readily solved through simulation given a statistical description of $\{\Lambda(t), t \geq 0\}$, but such a solution does not typically provide insights on the structure of the optimal price, as well as the effect of the market size variability. An alternative approach would use the stochastic fluid model approximation introduced earlier. Specifically, the instantaneous demand rate at time t is given by $\lambda(t) = \min(\mu, \Lambda(t)\bar{F}(p))$, i.e., the rate is equal to μ at time when the system is under-capacitated ($\mu \leq \Lambda(t)\bar{F}(p)$), and it is equal to the available demand $\Lambda(t)\bar{F}(p)$ when the system is over-capacitated. The above expression leads to the following revenue optimization problem:

$$\max_{p} \quad p\, \mathbb{E}\left[\int_0^T \min(\mu, \Lambda(s)\bar{F}(p))ds\right]. \tag{24}$$

Suppose further that $\Lambda(0)$ is drawn from the steady-state distribution[3], then $\Lambda(s)$ is distributed according to the steady-state distribution for all $s \geq 0$, and the static pricing

[3]To avoid this assumption, one can alternatively look at a long run average objective.

problem in (24) reduces to

$$\max_p \quad pT\,\mathbb{E}\left[\min(\mu, \Lambda_\infty \bar{F}(p))\right]. \tag{25}$$

Note that p^f requires the full description of the steady-state distribution of $\Lambda(\cdot)$.

High-low CTMC model: Using the limiting probabilities in the objective function (25) reduces to:

$$\max_p \quad p\left[P_1 \min(\Lambda_1 \bar{F}(p), \mu) + P_2 \min(\Lambda_2 \bar{F}(p), \mu)\right],$$

which is readily solvable without any simulation.

Continuous CIR model: The static optimization problem can now be written as

$$\max_p \quad \left\{p\bar{F}(p)h(\mu/\bar{F}(p))\right\} \quad \text{where} \quad h(y) = \mathbb{E}[\min(y, \frac{\sigma^2}{4\alpha}\chi^2_\nu)] \qquad \text{for all } y > 0,$$

and be solved numerically.

Remark: We have compared the value of p^f and that of the best static price (that solves (23) for various instances and observed that the two prices were very close and their performances were not statistically distinguishable.

6.1.3 Capacity and valuations

Throughout, we fix $\mu = 1$. We assume that the customers' willingness-to-pay is exponentially distributed, $\bar{F}(x) = \exp(-x/2)$, $x > 0$, for which case the maximizer of $p\bar{F}(p)$ is given by $p^* = 2$.

6.2 General behavior

Figure 2 depicts a simulated sample path of the system behavior under the four candidate pricing policies considered in our study. The market size is a realization of a CIR process with parameters $\alpha = 0.1$, $b = e$ and $\sigma = 0.2$. In Figure 2(a), in addition to the sample path of $\Lambda(t)$, we also show the critical levels $\mu/\bar{F}(p^*)$ and $\mu/\bar{F}(p^f)$, and recall that $\hat{p}(t) = p^*$ whenever $\Lambda(t) \leq \mu/\bar{F}(p^*)$. The latter is illustrated for example in Figure 2(d) for times around $t = 5$. On the other hand, whenever $\Lambda(t) \leq \mu/\bar{F}(p^f)$, $h^{(p^f)}(\Lambda(t))$, the limiting queue length under p^f is equal to zero. We observe this in Figure 2(b) from time $t = 0$ to time $t = 6$.

Figure 2(b) also illustrates how the queue length under p^f "tracks" (with some latency) the limiting queue length $h^{(p^f)}(\Lambda(t))$ (as predicted by the convergence result of Theorem 1). Figure 2(c) illustrates that the dynamic pricing policies $\hat{p}$ and $\tilde{p}$ allow the system manager to maintain lower queue length levels than those observed under the static price p^f. Turning to the pricing policies themselves, Figure 2(d) illustrates how $\tilde{p}$, the dynamic pricing policy that does not use information about the market size, "tracks" $\hat{p}$. In addition, the underlying structure of $\hat{p}$ is

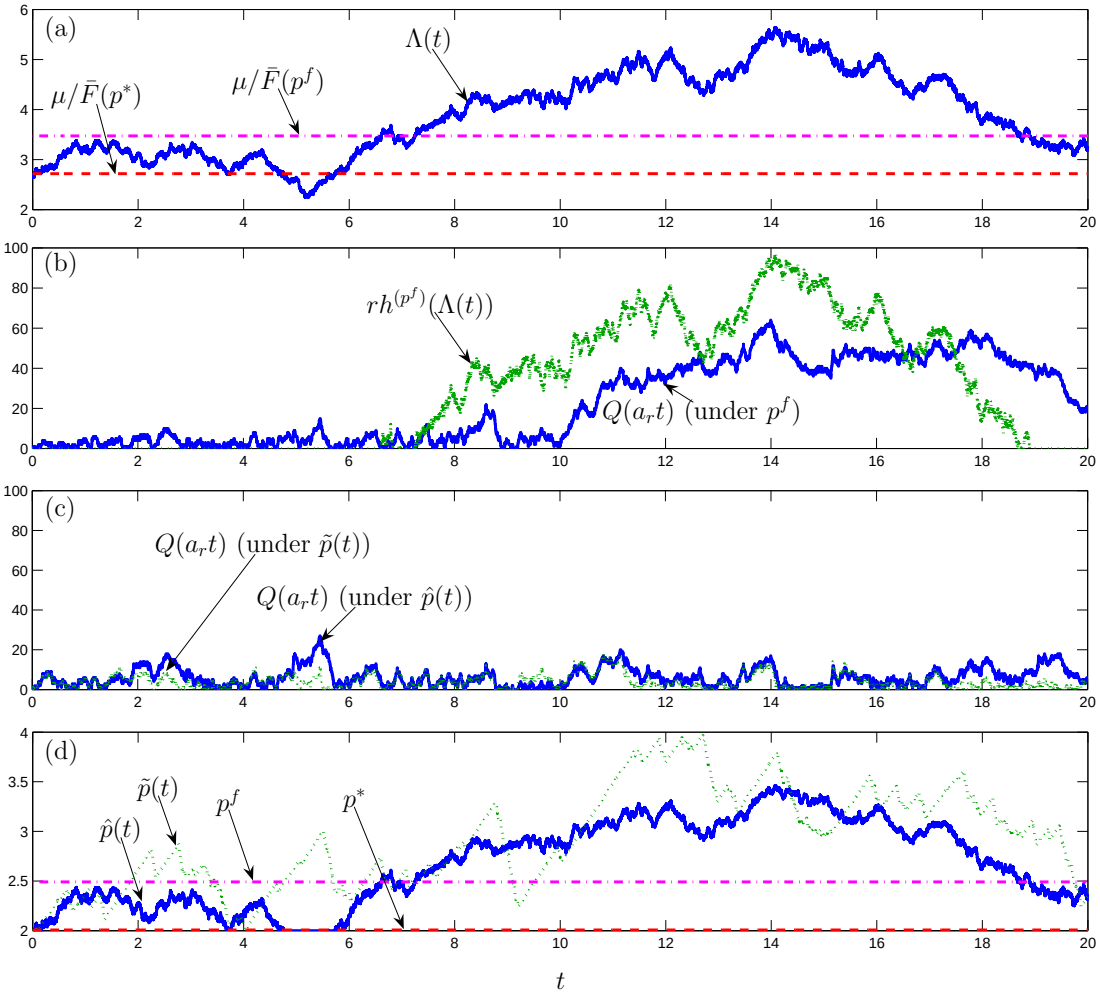

Figure 2: Illustration of the system behavior for a given sample path of market size for $r = 10$ and $a_r = 8$: (a) depicts the market size process $\Lambda(t)$, the level under which the system is congested under p^f and the level under which the system is congested when using p^*; (b) depicts the queue length under p^f as well as the limiting queue length; (c) represents the queue length processes under the two dynamic policies $\hat{p}$ and $\tilde{p}$; (d) shows the pricing policies p^f, $\hat{p}$ and $\tilde{p}$.

highlighted. Whenever the queue length is positive, the price increases linearly and whenever the system is empty, the price decreases linearly until it reaches p^*.

6.3 Performance study

We focus on a problem of moderate size with $r = 20$ and $a_r = 14$; i.e., using the scaling relations in (6) we will study a system for which $\mu^r = 20$ and $\Lambda^r(t) = 10\Lambda(t/14)$ for a various market size models $\Lambda(\cdot)$. Again, the main consequence of the expression $\Lambda^r(t) = 20\Lambda(t/14)$ is that the time scale within which $\Lambda^r(\cdot)$ is fluctuating is 14 times slower than that of the production dynamics.

Let R^r_π denote the expected revenues associated with a pricing policy π. Focusing on the four policies introduced earlier, we treat $\hat{p}(\cdot)$ as a benchmark and for every other policy π, compute its relative performance as $\theta(\pi) := 100 \times (1 - R^r_\pi / R^r_{\hat{p}})$. Note that $\theta(\pi)$ measures by how much $\hat{p}$ outperforms the policy π, in percentages of the performance of $\hat{p}$.

Next, we analyze the impact of various market size model parameters on the performance of the pricing policies. Throughout, we fix the horizon to be $T = 10$ and the estimated expected revenues were derived by averaging over 10^3 replications; the standard error never exceeded 1.2%.

Impact of variability: σ and $\Lambda_2 - \Lambda_1$ are proxies for the variability associated with the CIR and CTMC models, respectively. Table 1 reports the changes in performance associated with changes in these two parameters. We observe that the price based on the fluid approximation

Pricing policy	CIR model			CTMC model		
	σ			$\Lambda_2 - \Lambda_1$		
	0.2	0.3	0.4	1	2	4
p^s	0.5	2.6	4.5	0.5	4.7	15.7
p^f	0.1	1.3	2.6	0.1	1.2	1.6
$\tilde{p}$	–0.2	0.3	0.7	0.2	1.3	4.2

Table 1: Relative performance $\theta(\pi)$ as variability increases. $c = 0.2$; CIR model: $\alpha = 0.2$, $b = e$; CTMC model: $(\Lambda_1 + \Lambda_2)/2 = e$, $\nu = 1$, $p_{11} = 0.7$.

p^f yields in general an improvement in expected revenues when compared with the price p^s. For these two static prices, relative performance deteriorates as variability increases. This illustrates how the policies $\hat{p}(\cdot)$ and $\tilde{p}(\cdot)$ are able to effectively operate in highly variable environments.

Impact of "inertia": Table 2 reports the changes in performance associated with changes in the inertia of the market size process, which is related to α and p_{11} in the CIR and CTMC models, respectively. We observe that the value of dynamic pricing is higher in settings where

Pricing policy	CIR model			CTMC model		
	α			p_{11}		
	0.1	0.2	0.3	0.9	0.7	0.5
p^s	5.8	2.6	1.4	5.6	4.7	4.3
p^f	3.4	1.3	0.6	1.2	1.2	1.2
$\tilde{p}$	0.9	0.3	0.0	1.2	1.3	1.4

Table 2: Relative performance $\theta(\pi)$ as inertia varies. $c = 0.2$; CIR model: $b = e$, $\sigma = 0.3$; CTMC model: $\Lambda_1 = e - 1$ $\Lambda_2 = e + 1$, $\nu = 1$.

inertia is high. This observation confirms the following intuition: when inertia is high, the market size changes on an even slower time scale and the system has more time to reach the "equilibrium" queue length and it is on the latter that the dynamic pricing policies are based. When inertia is small, the system might not be able to reach the "equilibrium" sufficiently fast and the system

would always be tracking with some delay the fluid prediction, making the control less effective.

Impact of the average level of the market size: Table 3 reports the changes in performance associated with changes in the average market size, which is related to b and $(\Lambda_1+\Lambda_2)/2$ in the CIR and CTMC models, respectively. For small market sizes, all policies perform comparably

Pricing policy	CIR model			CTMC model		
	b			$(\Lambda_1+\Lambda_2)/2$		
	2	e	4	2	e	4
p^s	1.3	2.6	3.6	1.7	4.7	3.9
p^f	0.6	1.3	2.9	0.1	1.2	3.0
$\tilde{p}$	−0.2	0.3	2.5	0.0	1.3	2.6

Table 3: Relative performance $\theta(\pi)$ as the average market size varies. $c = 0.2$; CIR model: $\alpha = 0.2$, $\sigma = 0.3$; CTMC model: $\Lambda_2 - \Lambda_1 = 1$, $p_{11} = 0.7$, $\nu = 1$.

since the system will almost never be congested and pricing at (or close to) p^* is optimal. As the average market size increases, the policies $\hat{p}$ and $\tilde{p}$ start outperforming p^s and p^f significantly.

Impact of delay sensitivity: We report in Table 4 the impact of varying c. An interesting

Pricing policy	CIR model			CTMC model		
	c			c		
	0.1	0.2	0.3	0.1	0.2	0.3
p^s	3.2	2.6	2.2	5.5	4.7	4.2
p^f	1.7	1.3	1.1	1.2	1.2	1.1
$\tilde{p}$	0.8	0.3	0.0	1.7	1.3	1.1

Table 4: Relative performance $\theta(\pi)$ as the delay sensitivity varies. CIR model: $b = e$, $\alpha = 0.2$, $\sigma = 0.3$; CTMC model: $\Lambda_1 = e - 1$ $\Lambda_2 = e + 1$, $p_{11} = 0.7$, $\nu = 1$

point to note here is that the relative advantage of $\hat{p}$ over p^s and $\tilde{p}$ decreases as the delay sensitivity increases. This is due to the fact that $\hat{p}$ does not depend on delay sensitivity, whereas p^s and $\tilde{p}$ take it into account ($\tilde{p}$ does that indirectly through congestion feedback).

Impact of price responsiveness parameters (β^-, β^+) of the policy $\tilde{p}$: We report in Table 5 the relative performance of $\tilde{p}$ as (β^-, β^+) varies. We observe that in general, β^- should be chosen higher than β^+. The intuition associated with this is that the policy $\tilde{p}$ should not "overreact" when the queue length is positive, as this could be the result of normal queue fluctuations rather than changes in the underlying market size. We would also like to point out here that one could refine such policies by imposing that the price only starts to increase at rate β^+ when the queue length has been busy for a predetermined amount of time. This would allow to avoid reacting too fast to fluctuations of the queuing system on the fast time scale and only react to congestion that builds up due to changes in the market conditions.

		β^+			
		0.05	0.1	0.3	0.5
β^-	0.05	(7.2 ; 7.1)	(18.5 ; 18.3)	(44.6 ; 44.4)	(56.8 ; 56.8)
	0.1	(2.7 ; 3.6)	(9.6 ; 9.7)	(32.4 ; 32.4)	(45.5 ; 45.5)
	0.3	(0.1 ; 1.4)	(1.5 ; 2.6)	(11.8 ; 12.4)	(21.6 ; 22.0)
	0.5	(0.1 ; 1.2)	(0.3 ; 1.3)	(5.8 ; 6.9)	(12.4 ; 13.3)

Table 5: Relative performance $\theta(\pi)$ of $\tilde{p}$ as (β^-, β^+) varies. The first argument represents the relative performance under the CIR model while the second one reports the performance under the CTMC model. $c = 0.2$; CIR model: $b = e$, $\alpha = 0.2$, $\sigma = 0.3$; CTMC model: $\Lambda_1 = e - 1$ $\Lambda_2 = e + 1$, $p_{11} = 0.7$, $\nu = 1$

6.4 Concluding remarks

In this paper, we have developed pricing recommendations for a make-to-order manufacturer operating in an environment with stochastically varying market size. In particular, we have focused on large scale systems where the dynamics of the market size are slow relative to the dynamics of the production queue. Two main informational settings were studied: i.) one in which the decision maker observes the underlying market size at any point time and ii.) a setting in which the market size is not observable. The paper derived near-optimal policies for both cases, based on the analysis of a stochastic fluid model. A key point is that the decision maker is able to take advantage of the feedback present in the system to design pricing policies where the market size is not observable. From a general perspective, the paper has highlighted how one can incorporate market size time variations in designing pricing policies and the value associated with doing so. The broad set of numerical results illustrated this value in a variety of market conditions. From a methodological point view, the proof technique we have used, that is based on [9], should generalize to more general pricing problems than the single product application considered here.

Acknowledgments

The authors are grateful to three anonymous referees for their comments that lead to significant improvements of the original manuscript.

References

[1] P. Afèche. Incentive-compatible revenue management in queueing systems: Optimal strategic idleness and other delaying tactics. 2004. Working paper, Kellogg School of Management, Northwestern University.

[2] P. Afèche and B. Ata. Revenue management in queueing systems with unknown demand characteristics. 2005. working paper, Kellogg School of Management.

[3] V. F. Araman and R. A. Caldentey. Dynamic pricing for non-perishable products with demand learning. 2005. working paper, Stern School of Business, New York University.

[4] B. Ata and S. Schneorson. Dynamic control of an M/M/1 service system with adjustable arrival and service rates. 2006. forthcoming in Management Science.

[5] Y. Aviv and A. Pazgal. Pricing of short life-cycle products through active learning. 2005. under review.

[6] A. Bassambo, J. M. Harrison, and A. Zeevi. Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems*, 51:249–285, 2005.

[7] A. Bassamboo, J. Harrison, and A. Zeevi. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Oper. Res.*, 54(3):419–435, 2006.

[8] O. Besbes and A. Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *working paper, Columbia University*, 2007.

[9] M. Bramson. State space collapse with applications to heavy-traffic limits for multiclass queueing networks. *Queueing Systems*, 30:89–148, 1998.

[10] P. Brémaud. *Point Processes and Queues: Martingale Dynamics*. Springer-Verlag, 1980.

[11] S. Çelik and C. Maglaras. Dynamic pricing and leadtime quotation for a multi-class make-to-order queue. 2005. Working paper, Columbia Business School.

[12] H. Chen and M. Z. Frank. State dependent pricing with a queue. *IIE Transactions*, 32, 2000.

[13] J. G. Dai and T. Tezcan. State space collapse in many server diffusion limits of parallel server systems. 2005. working paper, Georgia Institute of Technology.

[14] G. Edmondson. Customization – BMW. *BusinessWeek*. November 24, 2003.

[15] G. Gallego and G. van Ryzin. Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science*, 40(8):999–1020, 1994.

[16] S. B. Gershwin, B. Tan, and M. H. Veatch. Production control with backlog-dependent demand. *working paper*, 2004.

[17] L. V. Green and P. J. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37(1):84–97, 1991.

[18] J. M. Harrison. *Brownian motion and stochastic flow systems*. John Wiley & Sons, 1985.

[19] R. Hassin and M. Haviv. *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, 2002.

[20] S. Kachani and G. Perakis. A fluid dynamics model of dynamic pricing and inventory control for make-to-stock manufacturing systems. 2002. Working paper, MIT Sloan School of Management.

[21] A. J. Kleywegt. An optimal control problem of dynamic pricing. 2001. Working paper, Georgia Institute of Technology.

[22] D. Lamberton and B. Lapeyre. *Introduction to Stochastic Calculus Applied to Finance*. Chapman & Hall, 1996.

[23] J. D. C. Little. A model of adaptive control of promotional spending. *Oper. Res.*, 14:1075–1097, 1966.

[24] D. W. Low. Optimal dynamic pricing policies for an M/M/s queue. *Oper. Res.*, 22:545–561, 1974.

[25] C. Maglaras. Revenue management for a multi-class single-server queue via a fluid model analysis. *Oper. Res.*, 54(5):914–932, 2006.

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

[27] A. Mandelbaum and G. Pats. State-dependent queues: approximations and applications. In F. Kelly and R. Williams, editors, *Stochastic Networks*, volume 71, pages 239–282. Proceedings of the IMA, 1995.

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

[29] H. Mendelson and S. Whang. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.*, 38(5):870–883, 1990.

[30] P. Naor. The regulation of queue size by levying tolls. *Econometrica*, 37:15–24, 1969.

[31] L. Overbeck and T. Rydén. Estimation in the Cox-Ingersoll-Ross model. *Econometric Theory*, 13:430–461, 1997.

[32] J.-S. Song and P. Zipkin. Inventory control in a fluctuating demand environment. *Oper. Res.*, 41:351–370, 1993.

[33] S. G. Steckley, S. G. Henderson, and V. Mehrotra. Service system planning in the presence of a random arrival rate. 2004. working paper, Cornell University.

A Proofs of the Main Results

Notation: Throughout, we let $C > 0$ and C_p denote Lipschitz constants associated with $\bar{F}$ and a generic policy $p \in \mathcal{P}$, respectively. For any integer $k > 0$, for any $y \in \mathbb{R}^k$, we define $|y| := \max\{|y_i|,\, i = 1, \ldots, k\}$; for any function $f : \mathbb{R}_+ \to \mathbb{R}^k$ and $L > 0$, we define $\|f(\cdot)\|_L := \sup_{0 \le t \le L} |f(t)|$.

Proof of Proposition 1. Let us first consider statement *i.*). We note that one can rewrite the fluid equations (9)-(12) using the reflexion mapping Φ (see, e.g., [27])

$$q = \Phi(x), \quad \text{where } \; x(s) = q(0) + \int_0^s [\bar{\Lambda}\bar{F}\big(p(q(u)/\mu, \bar{\Lambda}/\mu) + (c/\mu)q(u)\big) - \mu]du, \quad \text{(A-1)}$$

and where $\Phi(x)(t) = x(t) + \sup_{0 \le s \le t}(x(s))^-$. $\sup_{0 \le s \le t}(x(s))^-$ represents the correction needed for the process $q(t)$ to stay non-negative. (See [18] for a discussion of one-sided regulators.). Then, the results in [27, Section 3.1] directly imply *i.*), since $\bar{F}(y)$ tends to zero as $y \to \infty$.

We now turn to statement *ii.*). Fix $\delta > 0$. Note that $q(t)$ converges monotonically to $h^{(p)}(\bar{\Lambda})$.

If $h^{(p)}(\bar{\Lambda}) - \delta \le q(0) \le h^{(p)}(\bar{\Lambda}) + \delta$, then $q(t) \in [h^{(p)}(\bar{\Lambda}) - \delta, h^{(p)}(\bar{\Lambda}) - \delta]$ for all $t \ge 0$ and hence one can choose $s(\delta, M, \bar{\Lambda}) = 0$.

Suppose $q(0) > h^{(p)}(\bar{\Lambda}) + \delta$. We know that $q(t) \downarrow h^{(p)}(\bar{\Lambda})$ as $t \to \infty$ and that $q(t)$ is continuous. Let $\tau_\delta = \inf\{s \ge 0 : q(s) = h^{(p)}(\bar{\Lambda}) + \delta\}$, $\bar{p} = p((h^{(p)}(\bar{\Lambda}) + \delta)/\mu, \bar{\Lambda}/\mu)$ and define the following process

$$v = \Phi(y), \qquad \text{where} \qquad y(s) = q(0) + \bar{\Lambda} \int_0^s \bar{F}\big(\bar{p} + (c/\mu)(h^{(p)}(\bar{\Lambda}) + \delta)\big)du - \mu s.$$

Note that we have $y(s) = q(0) + (\bar{\Lambda}\bar{F}\big(\bar{p} + (c/\mu)(h^{(p)}(\bar{\Lambda}) + \delta)\big) - \mu)s$. Let $\tau_1(\bar{\Lambda}, \delta) = \inf\{s \ge 0 : v(s) = h^{(p)}(\bar{\Lambda}) + \delta\}$ and note that its value is readily computed as $\tau_1(\delta, q(0), \bar{\Lambda}) = (q(0) - (h^{(p)}(\bar{\Lambda}) + \delta))/\big(\mu - \bar{\Lambda}\bar{F}\big(\bar{p} + (c/\mu)(h^{(p)}(\bar{\Lambda}) + \delta)\big)\big)$. For all $s \le \tau_\delta$, $v(s) \ge q(s)$ since $y'(\cdot) \le x'(\cdot)$ in that region. We deduce that $\tau_1(\delta, q(0), \bar{\Lambda}) \ge \tau_\delta$ and one can take $s(\delta, M, \bar{\Lambda}) = \tau_1(\delta, M, \bar{\Lambda})$.

Now suppose that $q(0) < h^{(p)}(\bar{\Lambda}) - \delta$. Then necessarily $\bar{\Lambda}\bar{F}(p) > \mu$ (and $h^{(p)}(\bar{\Lambda}) > 0$). We know that $q(t) \uparrow h^{(p)}(\bar{\Lambda})$. Let $\tau_\delta = \inf\{s \ge 0 : q(s) = h^{(p)}(\bar{\Lambda}) - \delta\}$ and $\bar{p} = p((h^{(p)}(\bar{\Lambda}) - \delta)/\mu, \bar{\Lambda}/\mu)$ and define the new process

$$v = \Phi(y), \qquad \text{where} \qquad y(s) = q(0) + \bar{\Lambda} \int_0^s \bar{F}\big(\bar{p} + (c/\mu)(h^{(p)}(\bar{\Lambda}) - \delta)\big)du - \mu s,$$

and note that $y(s) = q(0) + (\bar{\Lambda}\bar{F}\big(\bar{p} + (c/\mu)(h^{(p)}(\bar{\Lambda}) - \delta)\big) - \mu)s$. Let $\tau_2(\delta, q(0), \bar{\Lambda}) = \inf\{s \ge 0 : v(s) = h^{(p)}(\bar{\Lambda}) - \delta\} = (-q(0) + (h^{(p)}(\bar{\Lambda}) - \delta))/(\bar{\Lambda}\bar{F}\big(\bar{p} + (c/\mu)(h^{(p)}(\bar{\Lambda}) - \delta)\big) - \mu)$. For all $s \le \tau_\delta$,

$v(s) \leq q(s)$ since $y'(\cdot) \geq x'(\cdot)$ in that region, implying that $\tau_2(\delta, q(0), \bar{\Lambda}) \geq \tau_\delta$. One can take $s(\delta, M, \bar{\Lambda}) = \tau_2(\delta, 0, \bar{\Lambda})$.

We deduce from all cases that one can take

$$s(\delta, M, \bar{\Lambda}) = \max\left\{\tau_1(\delta, M, \bar{\Lambda}), \tau_2(\delta, 0, \bar{\Lambda}), 0\right\}.$$

This completes the proof. ■

Proof of Theorem 1. Throughout, and in accordance with (6), we denote all quantities in the r^{th} system with a superscript r. For example the arrival process in the r^{th} system is denoted $A^r(\cdot)$. The proof is organized around three main steps. We first define for each r a sequence of processes that are "initialized" at different times in $[0, a_r T]$ and analyze them (Step 1). The underlying idea for introducing these processes is that at any time t, one of them will be "close" to the desired limit $h^{(p)}(\Lambda(t))$. We show that they admit limits that satisfy some fluid equations (Step 2). We conclude in Step 3 by showing that for any $t \in (0, T]$, $\tilde{Q}^r(t)$ is close to one of the processes introduced and the latter is close to the limit $h^{(p)}(\Lambda(t))$ of the solution to some fluid equation using Proposition 1. The proofs of the lemmas required for this result are provided in Appendix B.

We will follow [9] verbatim adjusting his results to our framework whenever needed. We will explicitly reference his corresponding results to facilitate the reading. Our proof is self-contained with the exception of Lemma 3.

$\Lambda(t)$ being continuous and strictly positive on $[0, T]$, it is bounded and there exists a constant $M_1 > 0$ such that $\max_{0 \leq t \leq T} \Lambda(t) \leq M_1$. Note also that since the maximal willingness to pay value $v_{max} < \infty$, $Q^r(u)/r \leq M_2$ a.s. for all $u \geq 0$ where $M_2 = \max\{M, v_{max}\mu/c\}$ (recall that M is such that $Q^r(0)/r \leq M$ and c is the congestion sensitivity parameter). Indeed, whenever $Q^r(u)/r \geq v_{max}\mu/c$, the arrival rate in the r^{th} system is upper bounded by $\Lambda^r(t)\bar{F}((c/\mu)Q^r(u)/r) \leq \Lambda^r(t)\bar{F}(v_{max}) = 0$ and hence is necessarily equal to zero, implying that the queue cannot grow. Since it is assumed that $Q^r(0)/r \leq M$, it follows that $Q^r(u)/r$ will never exceed M_2 a.s..

Step 1. Following Bramson's notations (cf. [9, equation (5.3)]), we define for $m = 1, 2, ..., \lfloor a_r T \rfloor$,

$$
\begin{aligned}
X^{r,m}(t) &= \frac{Q^r(t+m)}{r}, && \text{(A-2)}\\
\mathscr{T}^{r,m}(t) &= \mathscr{T}^r(t+m) - \mathscr{T}^r(m), && \text{(A-3)}\\
A^{r,m}(t) &= \frac{1}{r}\left[A^r(t+m) - A^r(m)\right], && \text{(A-4)}\\
D^{r,m}(t) &= \frac{1}{r}\left[D^r(t+m) - D^r(m)\right], && \text{(A-5)}\\
\Lambda^{r,m}(t) &= \frac{1}{r}\Lambda^r(t+m), && \text{(A-6)}\\
\mathfrak{X}^{r,m}(\cdot) &= \left(X^{r,m}(\cdot), A^{r,m}(\cdot), \mathscr{T}^{r,m}(\cdot), D^{r,m}(\cdot), \Lambda^{r,m}(\cdot)\right). && \text{(A-7)}
\end{aligned}
$$

We now analyze approximations to the processes introduced above. This is the equivalent of [9, Proposition 5.1].

Lemma 1 *Fix $\epsilon > 0$ and $L > 0$. For r large enough,*

$$
\begin{aligned}
&\mathbb{P}_\Lambda\left(\max_{m<a_rT}\left\|A^{r,m}(t) - \int_0^t \Lambda^{r,m}(u) g\left(\frac{1}{\mu}X^{r,m}(u); \frac{1}{\mu}\Lambda^{r,m}(u)\right) du\right\|_L > \epsilon\right) \le \epsilon, && \text{(A-8)}\\
&\mathbb{P}_\Lambda\left(\max_{m<a_rT}\left\|D^{r,m}(t) - \mu\mathscr{T}^{r,m}(t)\right\|_L > \epsilon\right) \le \epsilon, && \text{(A-9)}\\
&\mathbb{P}_\Lambda\left(\max_{m<a_rT}\sup_{t_1,t_2\le L}|D^{r,m}(t_2) - D^{r,m}(t_1)| > \mu|t_2 - t_1| + \epsilon\right) \le \epsilon. && \text{(A-10)}
\end{aligned}
$$

One can choose a sequence $\epsilon(r)$ decreasing to 0 sufficiently slowly such that the inequalities (A-8)-(A-10) still hold. For such a sequence $\epsilon(r)$, let $N = \max(\mu, 1)$ and define

$$
\begin{aligned}
K_0^r &= \left\{w : \sup_{t_1,t_2\le L}|\mathfrak{X}^{r,m}(t_2) - \mathfrak{X}^{r,m}(t_1)| \le N|t_2 - t_1| + \epsilon(r),\ \forall m < a_rT\right\}, && \text{(A-11)}\\
K_1^r &= \left\{w : \max_{m<a_rT}\left\|A^{r,m}(t) - \int_0^t \frac{1}{r}\Lambda^{r,m}(u) g\left(\frac{1}{\mu}X^{r,m}(u); \frac{1}{\mu}\Lambda^{r,m}(u)\right) du\right\|_L \le \epsilon(r)\ ;\right.\\
&\qquad \left.\max_{m<a_rT}\left\|D^{r,m}(t) - \mu\mathscr{T}^{r,m}(t)\right\|_L \le \epsilon(r)\right\}, && \text{(A-12)}\\
K^r &= K_0^r \cap K_1^r. && \text{(A-13)}
\end{aligned}
$$

Paralleling the line of arguments leading to [9, Proposition 5.2, Corollary 5.1], one can establish

Lemma 2

$$\mathbb{P}_\Lambda(K^r) \to 1 \quad as \quad r \to \infty. \tag{A-14}$$

Let E be the space of RCLL functions $x : [0, L] \to \mathbb{R}^5$ and let

$$E' = \{x \in E : |x(0)| \leq \max\{M_1, M_3\} \ , \ |x(t_2) - x(t_1)| \leq N|t_2 - t_1| \ \text{ for all } t_1, t_2 \in [0, L]\},$$
$$E_0^r = \{\mathfrak{X}^{r,m}(\cdot, \omega), \ m < a_r T, \ \omega \in K^r\},$$
$$\mathcal{E}_0 = \{E_0^r, \ r \in \mathbb{R}^+\}.$$

Note that the sample paths of $\mathfrak{X}^{r,m}(\cdot)$ lie in E. We next show that for r sufficiently large, $\omega \in K^r$, the vector of processes $\mathfrak{X}^{r,m}(\cdot, \omega)$ is close to some cluster point of $\mathcal{E}_0$. This essentially parallels [9, Proposition 6.1].

Lemma 3 *Fix $\epsilon > 0$, $L > 0$ and choose r large enough. Then for $\omega \in K^r$ and any $m < a_r T$,*

$$\|\mathfrak{X}^{r,m}(\cdot, \omega) - \tilde{\mathfrak{X}}(\cdot)\|_L \leq \epsilon \tag{A-15}$$

for some cluster point $\tilde{\mathfrak{X}}(\cdot)$ of $\mathcal{E}_0$ with $\tilde{\mathfrak{X}}(\cdot) \in E'$.

Step 2. In this step, we analyze the dynamics of the cluster points. The next lemma highlights a special property of the cluster points $\tilde{\mathfrak{X}}(\cdot)$ that is specific to our analysis and is a consequence of the time scale separation embedded in the parameter scaling of (6). It formalizes the intuition presented in §3.2 that the market size should appear as constant in the limit, a key characteristic of the "Pointwise-Stationary-Approximation"; it essentially states that the fifth element of any cluster point $\tilde{\mathfrak{X}}(\cdot)$, which corresponds to $\tilde{\Lambda}(\cdot)$, cannot have any dependence on time.

Lemma 4 *Any cluster point $\tilde{\mathfrak{X}}(\cdot)$ of $\mathcal{E}_0$ has its component $\tilde{\Lambda}(\cdot)$ constant.*

Having established the above, we can now characterize the equations satisfied by the cluster points by mimicking [9, Proposition 6.2]

Lemma 5 *Fix $L > 0$. Then all cluster points $\tilde{\mathfrak{X}}(\cdot)$ of $\mathcal{E}_0$ are solutions of the fluid equations (9)-(12) on $[0, L]$ (where $(\tilde{X}, \tilde{D}, \tilde{\mathscr{T}}, \tilde{\Lambda})$ replaces $(q, d, \tau, \bar{\Lambda})$).*

Step 3. We are now in a position to conclude the proof. Fix η, ξ, $\epsilon > 0$. By Lemma 2, there exists an $r_0 > 0$ such that for all $r \geq r_0$,

$$\mathbb{P}_\Lambda(K^r) > 1 - \xi. \tag{A-16}$$

Take $L = \sup\{s(\epsilon, M_2, \ell) : \ell \in [0, M_1]\} + 1$ where $s(\cdot, \cdot, \cdot)$ was defined in Proposition 1. Note that the supremum is attained and is finite since $s(\epsilon, M_2, \cdot)$ is continuous. Pick r large enough such

that $L/a_r < \eta$. For all $v \in (L/a_r, T]$, let

$$m_r(v) = \min\{m \in \mathbb{N} : m \leq a_r v,\ \ a_r v \leq m + L\} = \max\left\{\lceil a_r v - L \rceil, 0\right\}. \tag{A-17}$$

Let $\tau' := a_r v - m_r(v)$ and note that $\tau' \geq L - 1 \geq s(\epsilon, M_2, \ell)$ for all $\ell \in [0, M_1]$. $\tilde{Q}^r(v) = Q^r(a_r v)/r = Q^r(\tau' + m_r(v))/r = X^{r,m_r(v)}(\tau')$. We deduce that for all $\omega \in K^r$, and for all $v \in [\eta, T]$,

$$\begin{aligned}
\left|\tilde{Q}^r(v) - h^{(p)}(\Lambda(v))\right| &= \left|X^{r,m_r(v)}(\tau') - h^{(p)}(\Lambda^{r,m_r(v)}(\tau'))\right| \\
&\leq \left|X^{r,m_r(v)}(\tau') - \tilde{X}(\tau')\right| + \left|\tilde{X}(\tau') - h^{(p)}(\tilde{\Lambda}(\tau'))\right| + \left|h^{(p)}(\tilde{\Lambda}(\tau')) - h^{(p)}(\Lambda^{r,m_r(v)}(\tau'))\right| \\
&\leq \epsilon + \epsilon + \epsilon.
\end{aligned}$$

Note that for r sufficiently large, the inequality $|X^{r,m_r(v)}(\tau') - \tilde{X}(\tau')| \leq \epsilon$ is true through (A-15). The latter, combined with the continuity of $h^{(p)}(\cdot)$ also ensures $|h^{(p)}(\tilde{\Lambda}(\tau')) - h^{(p)}(\Lambda^{r,m_r(v)}(\tau'))| \leq \epsilon$. To deduce $\left|\tilde{X}(\tau') - h^{(p)}(\tilde{\Lambda}(\tau'))\right| \leq \epsilon$, we used the fact that the cluster point satisfies the fluid equations (with $\bar{\Lambda} = \tilde{\Lambda}$) and Proposition 1 in combination with the fact that $\tau' \geq L - 1 \geq s(\epsilon, M_3, \ell)$ for all $\ell \in [0, M_1]$. We deduce that for r sufficiently large

$$\mathbb{P}_\Lambda\left(\left\|\tilde{Q}^r(v) - h^{(p)}(\Lambda_v)\right\|_{[\eta,T]} > 3\epsilon\right) < \xi. \tag{A-18}$$

Since η, ϵ, and ξ were arbitrary, the theorem is established. ■

Proof of Proposition 2. Let $\Lambda^* = \mu/\bar{F}(p^*)$ and note that $\Lambda(t) \geq \Lambda^*$ if and only if $\Lambda(t)\bar{F}(p^*) \geq \mu$ which in turn is equivalent to $p^\mu(\Lambda(t)/\mu) \geq p^*$.

i.) Suppose first that $\Lambda(t) \leq \Lambda^*$. Then, setting $p(q/\mu, \Lambda(t)/\mu) = p^*$ yields $h^{(p)}(\Lambda(t)) = 0$ and achieves an instantaneous revenue rate of $\Lambda(t)p^*\bar{F}(p^*)$ which is the maximal possible. Hence, $p(q/\mu, \Lambda(t)/\mu) = p^* = \max\{p^*, p^\mu(\Lambda(t)/\mu)\}$ is optimal in this case.

ii.) Suppose now that $\Lambda(t) > \Lambda^*$. Note first that the policy $\hat{p}(\cdot,\cdot)$ sets the price equal to $p^\mu(\Lambda(t)/\mu)$ at any point time, yields $h^{\hat{p}}(\Lambda(t)) = 0$ and achieves a revenue rate of

$$p^\mu(\Lambda(t)/\mu)\Lambda(t)\bar{F}(p^\mu(\Lambda(t)/\mu)) = p^\mu(\Lambda(t)/\mu)\mu.$$

Consider any policy $p(\cdot,\cdot)$ such that $p(0, \Lambda(t)/\mu) > p^\mu(\Lambda(t)/\mu)$. Such a policy also yields $h^{(p)}(\Lambda(t)) = 0$ but achieves a revenue rate of $p(0, \Lambda(t)/\mu)\Lambda(t)\bar{F}(p(0, \Lambda(t)/\mu))$. The latter is lower than that achieved by $\hat{p}$ by the unimodality assumption ($x \mapsto x\bar{F}(x)$ is non-increasing for $x \geq p^\mu(\Lambda(t)/\mu)$ since $p^\mu(\Lambda(t)/\mu) \geq p^*$). Consider now any policy $p(\cdot,\cdot)$ such that $p(0, \Lambda(t)/\mu) < p^\mu(\Lambda(t)/\mu)$.

For such a policy, the queue length $h^{(p)}(\Lambda(t)) > 0$ and the revenue rate achieved is given by $p(h^{(p)}(\Lambda(t))/\mu, \Lambda(t)/\mu)\mu$. Now note that by the definition of $h^{(p)}(\Lambda(t))$ and $p^{\mu}(\cdot)$, we must have $p(h^{(p)}(\Lambda(t))/\mu, \Lambda(t)/\mu) + (c/\mu)h^{(p)}(\Lambda(t)) = p^{\mu}(\Lambda(t)/\mu)$ and hence $p(h^{(p)}(\Lambda(t))/\mu, \Lambda(t)/\mu) < p^{\mu}(\Lambda(t)/\mu)$. We deduce that the revenue rate achieved by $p(\cdot,\cdot)$ is dominated by that of $\hat{p}(\cdot,\cdot)$. We deduce that an optimal policy must satisfy $p(0, \Lambda(t)/\mu) = p^{\mu}(\Lambda(t)/\mu)$ and that for such a policy $h^{(p)}(\Lambda(t)) = 0$.

From the two cases, we conclude that the pricing policy $\hat{p}(\cdot,\cdot)$ is optimal in the fluid limit. ■

Proof of Proposition 3. Fix $\epsilon > 0$, a policy $p(\cdot,\cdot) \in \mathcal{P}$ and $\Lambda(\cdot)$. Let $\mathcal{B} = \big\{\omega : \sup_{0<t\leq T}\big|\tilde{Q}^r(t) - h^{(p)}(\Lambda(t))\big| \leq \epsilon\big\}$, $M_1 = \max\{\Lambda(u), 0 \leq u \leq T\}$ and $M_2 = v_{\max}$. Note that without loss of generality, we can assume that $p(\cdot,\cdot)$ is bounded above by M_2. For any such policy, the normalized revenues are upper bounded as follows

$$\begin{aligned}\frac{R^r_{\Lambda}(T)}{ra_r} &\leq \mathbb{E}_{\Lambda}\bigg[\int_0^T p^r(u)\Lambda(u)\bar{F}\big(p^r(u) + \frac{c}{\mu}\tilde{Q}^r(u)\big)du \;\bigg|\; \mathcal{B}\bigg] \\ &\quad + \mathbb{E}_{\Lambda}\bigg[\int_0^T p^r(u)\Lambda(u)\bar{F}\big(p^r(u) + \frac{c}{\mu}\tilde{Q}^r(u)\big)du \;\bigg|\; \mathcal{B}^c\bigg]\mathbb{P}_{\Lambda}\big\{\mathcal{B}^c\big\}. \qquad \text{(A-19)}\end{aligned}$$

The first term on the right hand side of (A-19), A_1, can be rewritten and upper bounded as follows

$$\begin{aligned}A_1 &= \mathbb{E}_{\Lambda}\bigg[\int_0^T p^r(u)\Lambda(u)\bar{F}\big(p^r(u) + \frac{c}{\mu}h^{(p)}(\Lambda(u))\big)du\bigg] \\ &\quad + \mathbb{E}_{\Lambda}\bigg[\int_0^T p^r(u)\Lambda(u)\Big(\bar{F}\big(p^r(u) + \frac{c}{\mu}\tilde{Q}^r(u)\big) - \bar{F}\big(p^r(u) + \frac{c}{\mu}h^{(p)}(\Lambda(u))\big)\Big)du \;\bigg|\; \mathcal{B}\bigg] \\ &\leq \mathbb{E}_{\Lambda}\bigg[\int_0^T p^r(u)\Lambda(u)\bar{F}\big(p^r(u) + \frac{c}{\mu}h^{(p)}(\Lambda(u))\big)du\bigg] + \mathbb{E}_{\Lambda}\bigg[\int_0^T p^r(u)\Lambda(u)C\frac{c}{\mu}\big|\tilde{Q}^r(u) - h^{(p)}(\Lambda(u))\big|du \;\bigg|\; \mathcal{B}\bigg] \\ &\leq \mathbb{E}_{\Lambda}\bigg[\int_0^T p^r(u)\Lambda(u)\bar{F}\big(p^r(u) + \frac{c}{\mu}h^{(p)}(\Lambda(u))\big)du\bigg] + (Cc/\mu)M_1M_2T\epsilon.\end{aligned}$$

Using the definition of $h^{(p)}(\cdot)$, we deduce that

$$\begin{aligned}A_1 &\leq \mathbb{E}_{\Lambda}\bigg[\int_0^T p^r(u)\min\{\Lambda(u)\bar{F}(p^r(u)), \mu\}du\bigg] + (Cc/\mu)M_1M_2T\epsilon \\ &\leq \int_0^T \sup_{q\geq 0}\big\{q\ \min\{\Lambda(u)\bar{F}(q), \mu\}\big\}du + (Cc/\mu)M_1M_2T\epsilon. \qquad \text{(A-20)}\end{aligned}$$

The second term on the right hand side of (A-19) can be upper bounded by $M_1M_2T\mathbb{P}_{\Lambda}\{\mathcal{B}^c\}$. Combining this with (A-20), we get

$$\frac{R^r_{\Lambda}(T)}{ra_r} \leq \int_0^T \sup_{q\geq 0}\big\{q\ \min\{\Lambda(u)\bar{F}(q), \mu\}\big\}du + \max\{Cc/\mu, 1\}M_1M_2T(\epsilon + \mathbb{P}_{\Lambda}\{\mathcal{B}^c\}).$$

Letting $r \to \infty$, we get (using Theorem 1)

$$\limsup_{r\to\infty} \frac{R^r_\Lambda(T)}{ra_r} \quad \leq \quad \int_0^T \sup_{q\geq 0}\{q\ \min\{\Lambda(u)\bar{F}(q),\mu\}\}du + \max\{Cc/\mu,1\}M_1M_2T\epsilon.$$

Noting that the result is true for all $\epsilon > 0$, we get the desired inequality. ■

Proof of Proposition 4. Fix $\epsilon > 0$ and $\Lambda(\cdot)$. Suppose that one applies the policy $\hat{p}$. Let $\mathcal{B} = \{\omega : \sup_{0<t\leq T}|\tilde{Q}^r(t)| \leq \epsilon\}$, $M_1 = \max\{\Lambda(u), 0 \leq u \leq T\}$ and $M_2 = v_{max}$.

$$\begin{aligned}
& \mathbb{E}_\Lambda\left[\int_0^T \hat{p}(u)\Lambda(u)\bar{F}\big(\hat{p}(u) + \frac{c}{\mu}\tilde{Q}^r(u)\big)du\right] \\
\geq\ & \mathbb{E}_\Lambda\left[\int_0^T \hat{p}(u)\Lambda(u)\bar{F}\big(\hat{p}(u) + \frac{c}{\mu}\tilde{Q}^r(u)\big)du \ \Big|\ \mathcal{B}\right]\mathbb{P}_\Lambda\{\mathcal{B}\} \\
\geq\ & \left[\int_0^T \hat{p}(u)\Lambda(u)\bar{F}\big(\hat{p}(u) + \frac{c}{\mu}\epsilon\big)du\right]\mathbb{P}_\Lambda\{\mathcal{B}\} \\
=\ & \left[\int_0^T \hat{p}(u)\Lambda(u)\bar{F}(\hat{p}(u))du + \int_0^T \hat{p}(u)\Lambda(u)\big[\bar{F}\big(\hat{p}(u) + \frac{c}{\mu}\epsilon\big) - \bar{F}(\hat{p}(u))\big]du\right]\mathbb{P}_\Lambda\{\mathcal{B}\} \\
\geq\ & \left[\int_0^T \hat{p}(u)\Lambda(u)\bar{F}(\hat{p}(u))du - CM_1M_2\frac{c}{\mu}\epsilon\right]\mathbb{P}_\Lambda\{\mathcal{B}\}.
\end{aligned}$$

Note that under $\hat{p}$, $h^{(p)}(\Lambda(t)) = 0$ for all $0 \leq t \leq T$. Hence, using Theorem 1, we have that $\mathbb{P}_\Lambda\{\mathcal{B}\}$ converges to 1 as $r \to +\infty$ and in turn

$$\liminf_{r\to\infty} \mathbb{E}_\Lambda\left[\int_0^T \hat{p}(u)\Lambda(u)\bar{F}(\hat{p}(u) + \frac{c}{\mu}\tilde{Q}^r(u))du\right] \quad \geq \quad \int_0^T \hat{p}(u)\Lambda(u)\bar{F}(\hat{p}(u))du - CM_1M_2\frac{c}{\mu}\epsilon.$$

This being true for all $\epsilon > 0$, we have

$$\begin{aligned}
\liminf_{r\to\infty} \mathbb{E}_\Lambda\left[\int_0^T \hat{p}(u)\Lambda(u)\bar{F}(\hat{p}(u) + \frac{c}{\mu}\tilde{Q}^r(u))du\right] \ &\geq\ \int_0^T \hat{p}(u)\Lambda(u)\bar{F}(\hat{p}(u))du \\
&=\ \int_0^T \max_q\{q\min\{\Lambda(u)\bar{F}(q),\mu\}\}du.
\end{aligned}$$

This establishes the desired result. ■

Proof of Proposition 5. This proof is a variant of that of Theorem 1. In particular the only step that needs to be adapted is Step 2 where we showed that any cluster point $\tilde{\mathfrak{X}}$ satisfies the fluid equations. Since now $\tilde{p}$ depends on the idleness process, one needs to modify the arguments leading to (B-6) in the proof of Lemma 5. For any function $m(\cdot)$ defined on $[0,T]$, define $\Theta(m)(\cdot) : u \mapsto \beta^- m(u) + \beta^+(u - m(u))$. For $u \in [0,T]$, let $\tilde{f}(u) := p^* + \Phi(\Theta(\tilde{Y}))(u)$ and $\tilde{f}^{r,m}(u) :=$

$p^* + \Phi(\Theta(Y^{r,m}))(u)$. Now we have

$$\begin{aligned}
&\Big\|\tilde{A}(t) - \tilde{\Lambda}\int_0^t \bar{F}\big(\tilde{f}(u) + \frac{c}{\mu}\tilde{X}(u)\big)du\Big\|_L \\
&\quad\leq \|\tilde{A}(t) - A^{r,m}(t)\|_L + \Big\|A^{r,m}(t) - \int_0^t \Lambda^{r,m}(u)\bar{F}\big(\tilde{f}^{r,m}(u) + \frac{c}{\mu}X^{r,m}(u)\big)du\Big\|_L \\
&\qquad + \Big\|\int_0^t \Lambda^{r,m}(u)\big|\bar{F}\big(\tilde{f}^{r,m}(u) + \frac{c}{\mu}X^{r,m}(u)\big) - \bar{F}\big(\tilde{f}(u) + \frac{c}{\mu}\tilde{X}(u)\big)\big|du\Big\|_L \\
&\qquad + \Big\|\int_0^t \|\Lambda^{r,m}(\cdot) - \tilde{\Lambda}\|_L\bar{F}\big(\tilde{f}(u) + \frac{c}{\mu}\tilde{X}(u)\big)du\Big\|_L \\
&\quad\leq \delta + \epsilon(r) + M_1CT\big(\|\tilde{f}^{r,m}(u) - \tilde{f}(u)\|_L + c/\mu\|X^{r,m} - \tilde{X}\|_L\big) + T\delta.
\end{aligned}$$

Recall that the reflection operator $\Phi(\cdot)$ is Lipschitz continuous with respect to the uniform norm in the set of continuous real valued functions, and let C_Φ denote the corresponding constant (see, e.g., [27])

$$\begin{aligned}
\|\tilde{f}^{r,m}(u) - \tilde{f}(u)\|_L &= \|\Phi(\Theta(Y^{r,m}))(u) - \Phi(\Theta(\tilde{Y}))(u)\|_L \\
&\leq C_\Phi\|\Theta(Y^{r,m})(u) - \Theta(\tilde{Y})(u)\|_L \\
&\leq C_\Phi|\beta^- - \beta^+|\|Y^{r,m}(u) - \tilde{Y}(u)\|_L \\
&\leq C_\Phi|\beta^- - \beta^+|\delta.
\end{aligned}$$

We deduce that

$$\Big\|\tilde{A}(t) - \tilde{\Lambda}\int_0^t \bar{F}(\tilde{f}(u) + \frac{c}{\mu}\tilde{X}(u))du\Big\|_L \leq B_1\delta.$$

The rest of the proof establishing that any cluster point $\tilde{\mathfrak{X}}$ satisfies the fluid equations is similar to that of Lemma 5.

The last point that needs to be adapted is the analysis of the fluid model equations. The behavior of the fluid model is summarized in the next result

Lemma 6 *Suppose $\Lambda(\cdot) = \bar{\Lambda}$ for all t and that $q(0) \leq M$. In the fluid system defined through the equations (9)-(12), there exists a finite $\tau(M, \bar{\Lambda}) > 0$ such that*

$$\begin{aligned}
\tilde{p}(t) &= \max\{p^*, p^\mu(\bar{\Lambda}/\mu)\} \qquad \textit{for all } t \geq \tau(M, \bar{\Lambda}), \\
q(t) &= 0 \qquad \textit{for all } t \geq \tau(M, \bar{\Lambda}).
\end{aligned}$$

Hence, we can conclude that for all $\epsilon > 0$, $\mathbb{P}_\Lambda\{\sup_{0<s\leq T}|\tilde{Q}^r(s)| > \epsilon\} \to 0$ and $\mathbb{P}_\Lambda\{\sup_{0<s\leq T}|\tilde{p}^r(s) - \hat{p}(s)| > \epsilon\} \to 0$ as $r \to \infty$.

Finally, establishing that the upper bound in Proposition 3 is achieved follows from the same line of arguments as in the proof of Proposition 4. ■

B Proofs of Auxiliary Lemmas

Proof of Lemma 1. Let the arrival process in the r^{th} system be represented as $A^r(t) = N\left(\int_0^t \Lambda^r(u)\bar{F}\left(p(\frac{Q^r(u)}{\mu r}, \frac{\Lambda^r(u)}{\mu r}) + (c/(\mu r)Q^r(u)\right)dt\right)$, where $N(\cdot)$ is a unit rate Poisson process. Define

$$\begin{aligned}\tilde{N}(y) &= N\left(y + \int_0^m \Lambda^r(u)\bar{F}\left(p(\frac{Q^r(u)}{\mu r}, \frac{\Lambda^r(u)}{\mu r}) + c/(\mu r)Q^r(u)\right)du\right) \\ &\quad -N\left(\int_0^m \Lambda^r(u)\bar{F}\left(p(\frac{Q^r(u)}{\mu r}, \frac{\Lambda^r(u)}{\mu r}) + c/(\mu r)Q^r(u)\right)du\right),\end{aligned}$$

and note that $\tilde{N}(\cdot)$ is a unit rate Poisson process. We can write

$$\begin{aligned}A^{r,m}(t) &= \frac{1}{r}\left[A^r(t+m) - A^r(m)\right] \\ &= \frac{1}{r}\left[N\left(\int_0^{t+m} \Lambda^r(u)\bar{F}\left(p(\frac{Q^r(u)}{\mu r}, \frac{\Lambda^r(u)}{\mu r}) + c/(\mu r)Q^r(u)\right)du\right)\right. \\ &\quad \left. -N\left(\int_0^m \Lambda^r(u)\bar{F}\left(p(\frac{Q^r(u)}{\mu r}, \frac{\Lambda^r(u)}{\mu r}) + (c/(\mu r)Q^r(u)\right)du\right)\right] \\ &= \frac{1}{r}\tilde{N}\left(\int_m^{t+m} \Lambda^r(u)\bar{F}\left(p(\frac{Q^r(u)}{\mu r}, \frac{\Lambda^r(u)}{\mu r}) + c/(\mu r)Q^r(u)\right)du\right) \\ &= \frac{1}{r}\tilde{N}\left(r\int_0^t \Lambda^{r,m}(u)\bar{F}\left(p(\frac{X^{r,m}(u)}{\mu}; \frac{\Lambda^{r,m}(u)}{\mu}) + (c/\mu)X^{r,m}(u)\right)du\right).\end{aligned}$$

The three inequalities are proved in a similar fashion. We start with (A-8). By proposition 4.3 in [9], for large enough r,

$$\mathbb{P}\left(\|\tilde{N}(t) - t\|_{M_1 Lr} > \epsilon M_1 Lr\right) \leq \frac{\epsilon}{M_1 Lr}$$

Note that $0 \leq r\int_0^t \Lambda^{r,m}(u)g\left(X^{r,m}(u)/\mu; \Lambda^{r,m}(u)/\mu\right)du \leq M_1 Lr$ for $t \in [0, L]$ and hence,

$$r\left\|A^{r,m}(t) - \int_0^t \Lambda^{r,m}(u)\bar{F}\left(p(\frac{X^{r,m}(u)}{\mu}; \frac{\Lambda^{r,m}(u)}{\mu}) + (c/\mu)X^{r,m}(u)\right)du\right\|_L \leq \left\|\tilde{N}(t) - t\right\|_{M_1 Lr},$$

implying that

$$\begin{aligned}
&\mathbb{P}\left(\left\|A^{r,m}(t)-\int_0^t \Lambda^{r,m}(u)\bar{F}\left(p(\frac{X^{r,m}(u)}{\mu};\frac{\Lambda^{r,m}(u)}{\mu})+(c/\mu)X^{r,m}(u)\right)du\right\|_L>\epsilon\right)\\
&\quad\leq\ \mathbb{P}\left(\frac{1}{r}\left\|\tilde{N}(t)-t\right\|_{M_1Lr}>\epsilon\right)\\
&\quad\leq\ \frac{\epsilon}{(M_1L)^2r}.
\end{aligned}$$

In turn, we have

$$\begin{aligned}
&\mathbb{P}\left(\max_{m<a_r\tau}\left\|A^{r,m}(t)-\int_0^t \Lambda^{r,m}(u)\bar{F}\left(p(\frac{X^{r,m}(u)}{\mu};\frac{\Lambda^{r,m}(u)}{\mu})+(c/\mu)X^{r,m}(u)\right)du\right\|_L>\epsilon\right)\\
&\leq\ \sum_{m=1}^{\lfloor a_r\tau\rfloor}\mathbb{P}\left(\left\|A^{r,m}(t)-\int_0^t \Lambda^{r,m}(u)\bar{F}\left(p(\frac{X^{r,m}(u)}{\mu};\frac{\Lambda^{r,m}(u)}{\mu})+(c/\mu)X^{r,m}(u)\right)du\right\|_L>\epsilon\right)\\
&\leq\ \frac{\epsilon a_r\tau}{(M_1L)^2r}.
\end{aligned}$$

(A-8) now follows for r large enough since $a_r/r\to 0$ as $r\to\infty$. The inequalities (A-9) and (A-10) follow in a similar fashion. ■

Proof of Lemma 2. We first note that Lemma 1 implies that $\mathbb{P}(K_1^r)\to 1$ as $r\to\infty$. We now show that $\mathbb{P}(K_0^r)\to 1$ as $r\to\infty$ and the result will directly follow. Fix $\epsilon>0$. Let $\gamma(\cdot)$ denote a generic process, $N\geq 0$ be a generic constant and let (P_0) be the following property:

$$\mathbb{P}\left(\max_{m<a_rT}\sup_{t_1,t_2\leq L}|\gamma^{r,m}(t_2)-\gamma^{r,m}(t_1)|>N|t_2-t_1|+\epsilon\right)\leq\epsilon. \tag{B-1}$$

We show that each component of $\mathfrak{X}$ satisfies the property (P_0) for r sufficiently large. We have already shown in Lemma 1 that the component $D^{r,m}(\cdot)$ satisfies (P_0) with $N=\mu$. We also note that $\mathscr{T}^{r,m}(\cdot)$ is Lipschitz with constant 1 and hence trivially satisfies the property with $N=1$. We now focus on $A^{r,m}(\cdot)$. Let $0\leq t_1\leq t_2\leq L$.

$$\begin{aligned}
\left|A^{r,m}(t_2)-A^{r,m}(t_1)\right|\ &\leq\ \left|A^{r,m}(t_2)-\int_0^{t_2}\Lambda^{r,m}(u)\bar{F}\left(p(\frac{X^{r,m}(u)}{\mu};\frac{\Lambda^{r,m}(u)}{\mu})+(c/\mu)X^{r,m}(u)\right)du\right|\\
&\quad+\left|A^{r,m}(t_1)-\int_0^{t_1}\Lambda^{r,m}(u)\bar{F}\left(p(\frac{X^{r,m}(u)}{\mu};\frac{\Lambda^{r,m}(u)}{\mu})+(c/\mu)X^{r,m}(u)\right)du\right|\\
&\quad+\int_{t_1}^{t_2}\Lambda^{r,m}(u)\bar{F}\left(p(\frac{X^{r,m}(u)}{\mu};\frac{\Lambda^{r,m}(u)}{\mu})+(c/\mu)X^{r,m}(u)\right)du\\
&\leq\ 2\left\|A^{r,m}(t)-\int_0^t\Lambda^{r,m}(u)\bar{F}\left(p(\frac{X^{r,m}(u)}{\mu};\frac{\Lambda^{r,m}(u)}{\mu})+(c/\mu)X^{r,m}(u)\right)du\right\|_L\\
&\quad+M_1|t_2-t_1|.
\end{aligned}$$

This implies that

$$\begin{aligned}
&\mathbb{P}\left(\max_{m<a_rT}\sup_{t_1,t_2\le L}|A^{r,m}(t_2)-A^{r,m}(t_1)|>M_1|t_2-t_1|+2\epsilon\right)\\
\le\ &\mathbb{P}\left(2\max_{m<a_rT}\left\|A^{r,m}(t)-\int_0^t\frac{1}{r}\Lambda^{r,m}(u)\bar{F}\left(p(\frac{X^{r,m}(u)}{\mu};\frac{\Lambda^{r,m}(u)}{\mu})+(c/\mu)X^{r,m}(u)\right)du\right\|_L>2\epsilon\right)\\
\le\ &\epsilon,
\end{aligned}$$

where the last inequality follows from Lemma 1. The property is hence satisfied for $N=M_1$. Let us now turn to the component $\Lambda^{r,m}(\cdot)$.

$$\left|\Lambda^{r,m}(t_2)-\Lambda^{r,m}(t_1)\right| \ = \ \left|\Lambda\Big(\frac{t_2+m}{a_r}\Big)-\Lambda\Big(\frac{t_1+m}{a_r}\Big)\right|.$$

Note that $\Lambda(\cdot)$ being continuous on the compact set $[0,\tau]$, it is also uniformly continuous. Let $\delta>0$ be such that $|\Lambda(t_2)-\Lambda(t_1)|<\epsilon$ for all $t_1,t_2\in[0,\tau]$ such that $|t_2-t_1|<\delta$. Let $k=\min\{n: a_n>\tau/\delta\}$. For all $r\ge k$, we have

$$\left|\Lambda^{r,m}(t_2)-\Lambda^{r,m}(t_1)\right| \ \le \ \epsilon,$$

and hence $\Lambda^{r,m}(\cdot)$ satisfies the property (P_0).

The last component of $\mathfrak{X}$ to analyze is $X^{r,m}$. Using the queueing dynamics equations (1)-(4), note that

$$\begin{aligned}
X^{r,m}(t) &= \frac{Q^r(t+m)}{r}\\
&= \frac{1}{r}\left[Q^r(0)+A^r(t+m)-D^r(t+m)\right]\\
&= A^{r,m}(t)-D^{r,m}(t)+\frac{1}{r}\left[Q^r(0)+A^r(m)-D^r(m)\right]\\
&= X^{r,m}(0)+A^{r,m}(t)-D^{r,m}(t). \qquad \text{(B-2)}
\end{aligned}$$

Since both $A^{r,m}$ and $D^{r,m}$ satisfy (P_0) we conclude (using (B-2)) that $X^{r,m}$ also satisfies (P_0) with $N=M_1+\mu$. The lemma is now established. ■

Proof of Lemma 3. Note that for all $m<a_t\tau$, $\mathscr{T}^{r,m}(0)=A^{r,m}(0)=D^{r,m}(0)=0$, $X^{r,m}(0)=\frac{Q^r(m)}{r}\le M_2$, and $\Lambda^{r,m}(0)=\Lambda(m/a_r)\le M_1$. Hence $|\mathfrak{X}^{r,m}(0)|\le\max\{M_1,M_2\}$. The result now directly follows from Proposition 4.1 in [9]. ■

Proof of Lemma 4. Since $\tilde{\mathfrak{X}}(\cdot)$ is a cluster point of $\mathcal{E}_0$, there exists a sequence (r_l,m_l,ω_l) with

$m_l < a_{r_l}\tau$ and $\omega_l \in K^{r_l}$ such that $\left\|\mathfrak{X}^{r_l,m_l}(\cdot,\omega_l) - \tilde{\mathfrak{X}}(\cdot)\right\|_L \to 0$ as $l \to \infty$. In particular, this implies

$$\left\|\Lambda^{r_l,m_l}(\cdot) - \tilde{\Lambda}(\cdot)\right\|_L \to 0 \qquad \text{as} \quad l \to \infty. \tag{B-3}$$

Let $\mathcal{S}$ be the set of limit points of the sequence m_l/a_{r_l}. Note that this set is nonempty since m_l/a_{r_l} lives in the compact set $[0,\tau]$ and hence has at least one convergent subsequence. Let $\mathcal{L}$ be the image of $\mathcal{S}$ by $\Lambda(\cdot)$.

Suppose $\mathcal{L}$ contains two distinct elements $y_1 \neq y_2$. Take any $u \in [0,L]$. Denote by l' and l'' the indexes of two subsequences such that $m_{l'}/a_{r_{l'}} \to x_1$ and $\Lambda(x_1) = y_1$ and $m_{l''}/a_{r_{l''}} \to x_2$ and $\Lambda(x_2) = y_2$.

$$\begin{aligned}
\Lambda^{r_{l'},m_{l'}}(u) &= \Lambda(u/a_{r_{l'}} + m_{l'}/a_{r_{l'}}) \to \Lambda(x_1) = y_1 \qquad \text{as} \quad l \to \infty, \\
\Lambda^{r_{l''},m_{l''}}(u) &= \Lambda(u/a_{r_{l''}} + m_{l'}/a_{r_{l''}}) \to \Lambda(x_2) = y_2 \qquad \text{as} \quad l \to \infty.
\end{aligned}$$

Having $y_1 \neq y_2$ contradicts (B-3). We deduce that $\mathcal{L}$ is necessarily a singleton. Denote by y the unique point in $\mathcal{L}$. Consider any converging subsequence of m_l/a_{r_l} and denote by l' its indexes and x its limit ($x \in \mathcal{S}$).

$$\Lambda^{r_{l'},m_{l'}}(u) = \Lambda(u/a_{r_{l'}} + m_{l'}/a_{r_{l'}}) \to \Lambda(x) = y \qquad \text{as} \quad l \to \infty. \tag{B-4}$$

Hence, we must have $\tilde{\Lambda}(u) = y$ for all $u \in [0,L]$. In other words, for any cluster point, $\tilde{\Lambda}(\cdot)$ is constant. ■

Proof of Lemma 5. Select any cluster point $\tilde{\mathfrak{X}}(\cdot)$ of $\mathcal{E}_0$. For a given $\delta > 0$, choose (r,m,ω) so that $\epsilon(r) \leq \delta$,

$$\left\|\tilde{\mathfrak{X}}(\cdot) - \tilde{\mathfrak{X}}^{r,m}(\cdot,\omega)\right\|_L \leq \delta. \tag{B-5}$$

Let $\omega \in K^r$. We have

$$\begin{aligned}
& \left\| \tilde{A}(t) - \tilde{\Lambda} \int_0^t \bar{F}(p(\tilde{X}(u)/\mu, \tilde{\Lambda}(u)/\mu) + \frac{c}{\mu}\tilde{X}(u))du \right\|_L \\
\leq \; & \left\| \tilde{A}(t) - A^{r,m}(t) \right\|_L + \left\| A^{r,m}(t) - \int_0^t \Lambda^{r,m}(u)\bar{F}(p(X^{r,m}(u)/\mu; \Lambda^{r,m}(u)/\mu) + \frac{c}{\mu}X^{r,m}(u))du \right\|_L \\
& + \left\| \int_0^t \Lambda^{r,m}(u) \left| \bar{F}(p(X^{r,m}(u)/\mu, \Lambda^{r,m}(u)/\mu) + \frac{c}{\mu}X^{r,m}(u)) - \bar{F}(p(\tilde{X}(u)/\mu, \tilde{\Lambda}(u)/\mu) + \frac{c}{\mu}\tilde{X}(u)) \right| du \right\|_L \\
& + \left\| \int_0^t \|\Lambda^{r,m}(\cdot) - \tilde{\Lambda}\|_L \bar{F}(p(\tilde{X}(u)/\mu; \tilde{\Lambda}(u)/\mu) + \frac{c}{\mu}\tilde{X}(u))du \right\|_L \\
\overset{(a)}{\leq} \; & \delta + \epsilon(r) + LM_1C(C_p\delta + (c/\mu)\delta) + L\delta \\
\leq \; & B_1\delta, \qquad \text{(B-6)}
\end{aligned}$$

where B_1 is a suitably large positive constant and (a) follows from (B-5), the fact that $\omega \in K^r$ and the Lipschitz continuity of $\bar{F}$ and p. Now

$$\begin{aligned}
& \left\| \tilde{X}(t) - (\tilde{X}(0) + \tilde{A}(t) - \mu\tilde{\mathscr{T}}(t)) \right\|_L \\
\leq \; & \left\| \tilde{X}(t) - X^{r,m}(t) \right\|_L + \left\| X^{r,m}(0) - \tilde{X}(0) \right\|_L + \left\| A^{r,m}(t) - \tilde{A}(t) \right\|_L + \mu \left\| \tilde{\mathscr{T}}(t) - \mathscr{T}^{r,m}(t) \right\|_L \\
& + \left\| X^{r,m}(t) - (X^{r,m}(0) + A^{r,m}(t) - D^{r,m}(t)) \right\|_L + \left\| \mu\mathscr{T}^{r,m}(t) - D^{r,m}(t) \right\|_L \\
\leq \; & B_2\delta, \qquad \text{(B-7)}
\end{aligned}$$

where we made use of (B-5), the fact that $\omega \in K^r$ and $B_2 > 0$ is chosen suitably large. Using (B-6) and (B-7), we have

$$\begin{aligned}
& \left\| \tilde{X}(t) - (\tilde{X}(0) + \tilde{\Lambda} \int_0^t \bar{F}(p(\frac{\tilde{X}(u)}{\mu}; \frac{\tilde{\Lambda}(u)}{\mu}) + \frac{c}{\mu}\tilde{X}(u))du - \mu\tilde{\mathscr{T}}(t)) \right\|_L \\
\leq \; & \left\| \tilde{X}(t) - (\tilde{X}(0) + \tilde{A}(t) - \mu\tilde{\mathscr{T}}(t)) \right\|_L + \left\| \tilde{A}(t) - \tilde{\Lambda} \int_0^t \bar{F}(p(\frac{\tilde{X}(u)}{\mu}; \frac{\tilde{\Lambda}(u)}{\mu}) + \frac{c}{\mu}\tilde{X}(u))du \right\|_L \\
\leq \; & (B_1 + B_2)\delta.
\end{aligned}$$

Let $\tilde{Y}(t) = t - \tilde{\mathscr{T}}(t)$. We next show that $\tilde{X}(s) > \delta$ for all $s \in [t_1, t_2] \subset [0, L]$ implies that $\tilde{Y}(t_2) - \tilde{Y}(t_1) \leq 2\delta$. Suppose that $\tilde{X}(s) > \delta$ for all $s \in [t_1, t_2] \subset [0, L]$. This implies that $X^{r,m}(s) > 0$ for all $s \in [t_1, t_2]$ since $\|\tilde{X}(s) - X^{r,m}(s)\|_L \leq \delta$. We deduce that $Y^{r,m}(t_2) - Y^{r,m}(t_1) = 0$. Now

$$\begin{aligned}
\tilde{Y}(t_2) - \tilde{Y}(t_1) &= \tilde{Y}(t_2) - Y^{r,m}(t_2) + Y^{r,m}(t_2) - Y^{r,m}(t_1) + Y^{r,m}(t_1) - \tilde{Y}(t_1) \\
&\leq 2\delta.
\end{aligned}$$

Since δ was arbitrary, we conclude that $(\tilde{X}, \tilde{\mathscr{T}}, \tilde{D}, \tilde{\Lambda})$ satisfy the fluid equations. (9)-(12).

Proof of Lemma 6. Note that in the fluid system, q is bounded above by $\max\{q(0), q_{max}\}$ where $q_{max} = (\mu/c)v_{max}$. Note also that when $\Lambda(\cdot)$ is constant over time, so is $\hat{p}(t) = \max\{p^*, p^\mu\}$. Hence, we will drop the dependence in time for $\hat{p}(t)$. Rewriting the fluid equations as a reflection, we have

$$q = \Phi(x), \qquad \text{where} \qquad x(s) = q(0) + \int_0^s [\bar{\Lambda}\bar{F}(\tilde{p}(s) + (c/\mu)q(s)) - \mu]du.$$

Let $\tau^{(1)} = \inf\{s \geq 0 : x(s) = 0\}$. For any $\gamma > 0$, we have

$$\begin{aligned}
\tau^{(1)} &= \inf\Big\{s \geq 0 : q(0) + \int_0^s \big(\bar{\Lambda}\bar{F}\big(p^* + \beta u + (c/\mu)q(u)\big) - \mu\big)du = 0\Big\} \\
&\leq \inf\Big\{s \geq \gamma : q(\gamma) + \int_\gamma^s \big(\bar{\Lambda}\bar{F}(p^* + \beta\gamma) - \mu\big)du = 0\Big\} \\
&= \gamma + \frac{q(\gamma)}{\mu - \bar{\Lambda}\bar{F}(p^* + \beta\gamma)} \\
&\leq \gamma + \frac{\max\{M, q_{max}\}}{\mu - \bar{\Lambda}\bar{F}(p^* + \beta\gamma)}.
\end{aligned}$$

We deduce that $\tau^{(1)}$ is finite.

Suppose first that $\bar{\Lambda}\bar{F}(p^*) \leq \mu$. Then $\hat{p} = p^*$. If $q(0) = 0$, then $q(t) = 0$ and $\tilde{p}(t) = p^* = \hat{p}$ for all $t \geq 0$. If $q(0) > 0$, Following $\tau^{(1)}$, $q(t)$ stays equal to zero since $x'(t) \leq 0$. We deduce that following $\tau^{(1)}$, $\tilde{p}(\cdot)$ decreases at rate β^- until it reaches p^* and hence one can take $\tau = \tau^{(1)} + (\beta^+/\beta^-)\tau^{(1)} \leq (1 + \beta^+/\beta^-)\big(\gamma + \max\{M, q_{max}\}/(\mu - \bar{\Lambda}\bar{F}(p^* + \beta\gamma)\big)$.

Suppose now that $\bar{\Lambda}\bar{F}(p^*) > \mu$. This implies that $\hat{p} = p^\mu = \bar{F}^{-1}(\mu/\bar{\Lambda})$. We can assume without loss of generality that $q(0) > 0$ (since if $q(0) = 0$, then $q(\delta) > 0$ for some $\delta > 0$). Let $\tau^{(1)} = \inf\{s \geq 0 : q(t) = 0\}$. Choose γ such that $\gamma > (p^\mu - p^*)/\beta$. Since $q(u) > 0$ for all $u < \tau^{(1)}$, we have that in addition, $q(\tau^{(1)}) = 0$ and $\tilde{p}(\tau^{(1)}) \geq p^\mu$. Let $\tau^{(2)} = \tau^{(1)} + (\tilde{p}(\tau^{(1)}) - p^\mu)/\beta^-$. By definition, $\tilde{p}(\cdot)$ is decreasing at rate β^- on $[\tau^{(1)}, \tau^{(2)}]$ and the queue length stays at zero on that interval. At $\tau^{(2)}$, we have $q(\tau^{(2)}) = 0$ and $\tilde{p}(\tau^{(2)}) = p^\mu$.

We next show that for all $t \geq \tau^{(2)}$, $q(t) = 0$ and $\tilde{p}(t) = p^\mu$. Let $\tau^{(3)} = \inf\{s \geq \tau^{(2)} : \tilde{p}(t) \neq p^\mu\}$. Suppose $\tau^{(3)} < \infty$ and $\tilde{p}(t)$ decreases for $\epsilon > 0$ units of time following $\tau^{(3)}$. $\tilde{p}(t)$ would have had to decrease at rate β^-. By definition of $\tilde{p}$, this would imply that the system was idling on $(\tau^{(3)}, \tau^{(3)} + \epsilon)$ and hence $q(t) = 0$ on that interval. In turn, this would imply that $\bar{\Lambda}\bar{F}(\tilde{p}(t) + (c/\mu)q(t)) - \mu > 0$ and that $x(t)$ was increasing on $(\tau^{(3)}, \tau^{(3)} + \epsilon)$, contradicting that $q(t)$ stays equal to zero. Similarly, if one supposes $\tau^{(3)} < \infty$ and $\tilde{p}(t)$ increases for $\epsilon > 0$ units of

time following $\tau^{(3)}$. Then this would imply that $x(t)$ is decreasing on $(\tau^{(3)}, \tau^{(3)} + \epsilon)$, and hence that the system is idling, contradicting that $\tilde{p}(t)$ increases on that interval. We conclude that necessarily $\tau^{(3)} = \infty$ and one can take $\tau = \tau^{(2)}$ in the case where $\bar{\Lambda}\bar{F}(p^*) > \mu$. This concludes the proof. ■