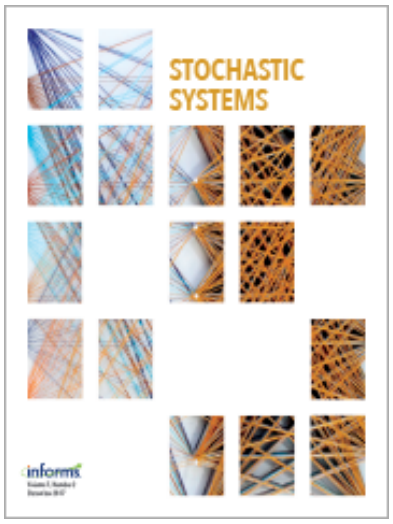



Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Optimal Control of a Time-Varying Double-Ended Production Queueing Model

Chihoon Lee, Xin Liu, Yunan Liu, Ling Zhang

To cite this article:
Chihoon Lee, Xin Liu, Yunan Liu, Ling Zhang (2021) Optimal Control of a Time-Varying Double-Ended Production Queueing Model. Stochastic Systems

Published online in Articles in Advance 09 Feb 2021

. https://doi.org/10.1287/stsy.2019.0066

http://pubsonline.informs.org/journal/stsy

STOCHASTIC SYSTEMS
*Articles in Advance*, pp. 1–34
ISSN 1946-5238 (online)

# Optimal Control of a Time-Varying Double-Ended Production Queueing Model

**Chihoon Lee,**[a,b] **Xin Liu,**[c] **Yunan Liu,**[d] **Ling Zhang**[d]

[a] School of Business, Stevens Institute of Technology, Hoboken, New Jersey 07030; [b] School of Data Science, The Chinese University of Hong Kong, 518172 Shenzhen, China; [c] School of Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina 29634; [d] Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina 27695-7906

**Contact:** clee4@stevens.edu (CL); xliu9@clemson.edu, https://orcid.org/0000-0002-2326-9657 (XL); yliu48@ncsu.edu, http://orcid.org/0000-0001-9961-2610 (YL); lzhang42@ncsu.edu (LZ)

**Received:** April 5, 2019
**Revised:** November 25, 2019; June 7, 2020
**Accepted:** September 15, 2020
**Published Online in Articles in Advance:** February 9, 2021

**https://doi.org/10.1287/stsy.2019.0066**



**Abstract.** Motivated by production systems with nonstationary stochastic demand, we study a double-ended queueing model having back orders and customer abandonment. One side of our model stores back orders, and the other side represents inventory. We assume first-come-first-served instantaneous fulfillment discipline. Our goal is to determine the optimal (nonstationary) production rate over a finite time horizon to minimize the costs incurred by the system. In addition to the inventory-related (holding and perishment) and demand-related (waiting and abandonment) costs, we consider a cost that penalizes rapid fluctuations of production rates. We develop a deterministic fluid-control problem (FCP) that serves as a performance lower bound for the original queueing-control problem (QCP). We further consider a high-volume system for which an upper bound of the gap between the optimal values of the QCP and FCP is characterized and construct an asymptotically optimal production rate for the QCP, under which the FCP lower bound is achieved asymptotically. Demonstrated by numerical examples, the proposed asymptotically optimal production rate successfully captures the time variability of the nonstationary demand.



**Keywords:** time-varying demand • asymptotic optimality • double-ended queue • abandonment • optimal control

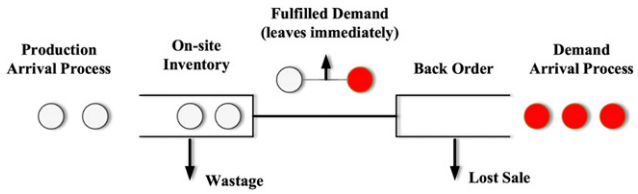
## 1. Introduction

One of the fundamental challenges that precludes desired dynamic behavior of stochastic systems is that of nonstationarity. For instance, a number of service systems, such as production lines, call centers, hospitals, and online trading, are subject to nonstationarity of customer demand, service times, and staffing levels. In a pharmaceutical-industry context, factors such as consumer perceptions (e.g., from advertising or time since entry), over-the-counter offerings, or ongoing market competition may lead to nonstationary demand behavior over time; for instance, figure 1 in Arcidiacono et al. (2013) shows monthly prescription fluctuations for brand and generic drugs.

Our work is motivated by the recent attention of industry, government, and academia to continuous manufacturing. In contrast to the traditional batch-production process, continuous manufacturing can produce more reliable products, while achieving higher efficiency in cost. Recognizing its value, the Food and Drug Administration approved in April 2016, for the first time, Janssen Products, LP's change in their production method from batch to continuous manufacturing; see Yu (2016). Moreover, continuous manufacturing allows manufacturers to adjust for nonstationary (i.e., time-varying) demand much more quickly, hence, preventing a potential shortage or large back-order quantities. The emerging question is: When and how much should one change the production rate so that the nonstationary demand and resource capacities can be timely and carefully balanced?

The goal of this paper is to address the above question for a class of production systems that exhibit the following three features: (i) Product demand is stochastic and *nonstationary*; (ii) produced goods are *perishable* and customers are *impatient*—that is, back orders are subject to cancellations; and (iii) rapid changes in production rate incur operational cost, hence, the term *inflexible* production system. Mathematically, we model the production system by a double-ended queue (see Figure 1), in which goods are produced according to a Poisson process with time-varying production rate, and demands arrive according to another Poisson process

**Figure 1.** A Production/Inventory System Modeled by a Double-Ended Queue




with time-dependent and state-dependent arrival rate. Upon arrival of a demand, if there are available products in the inventory, it will be fulfilled immediately, and if no product is available, it will be backlogged and wait for the upcoming products. The system state can be described by a one-dimensional queue-length process.

Our goal is to obtain an optimal (nonstationary) production rate over a *finite-time* production-planning horizon to minimize the costs incurred by the system. These costs include the production cost; inventory-related costs, consisting of holding and perishment costs; demand-related costs, consisting of waiting and abandonment costs; and a cost that penalizes the rapid fluctuations of production rates, measured by the total variation of the production-rate function.

Daunting challenges as they seem to be, we now describe our *solution approaches and contributions* to the existing body of literature.

- We develop a (deterministic) fluid control problem (FCP) that serves as a lower bound for the original queueing control problem (QCP) under any admissible control policy. Our result (Theorem 3.2) is established without assuming any asymptotics. We construct an auxiliary control problem to tackle the difficulty arising from the nonlinear drift of the expected queue-length process (Lemma 3.2). We believe our approach can be potentially applied for other control problems for double-ended queues and matching queues with similar nonlinear drift structure.
- We develop an asymptotic framework by considering high-volume systems, under which we characterize the gap between the optimal values of the QCP and the corresponding FCP. Further, we construct an asymptotically optimal production rate for the QCP based on the optimal solution of the FCP and show that the FCP lower bound is achieved asymptotically under the proposed production rate. To consider an asymptotic framework, we assume that the product demand is high and the system capacity can scale up large in response to the demand. This enables us to pursue approximate solutions based on asymptotic analysis. We show in Theorem 4.2 that the optimal value of the FCP serves asymptotically as a performance lower bound to the fluid-scaled QCP as a scaling factor $n$ (which measures the "size" of the system—for example, average product demand rate) grows large. We then propose a production rate by using the optimal solution of the FCP and show that under the proposed policy, the FCP lower bound is achieved asymptotically.
- We develop simple effective algorithms based on a linear programming (LP) formulation to numerically solve the FCP and conduct various simulations to demonstrate the effectiveness of the proposed asymptotically optimal production rate for different system sizes. One salient feature of our control problem is the consideration of production inflexibility cost—that is, the cost for rapidly increasing or decreasing production levels. We use the total variation of a production-rate function as a quantitative measure for the production-flexibility cost. As a result, the analytical solution of the FCP becomes intractable. Instead, we develop effective, yet simple, algorithms based on an LP formulation to numerically obtain the optimal production rates. Our simulations demonstrate that the proposed policy based on the FCP solutions is optimal with error $o(n)$.
- We solve the QCP numerically by formulating a dynamic programming (DP) problem and validate the asymptotic optimality of the FCP solution. The simulated average optimal control path based on the optimal DP policy is compared with the numerical solution of the FCP. Although our numerical studies are based on academic-sized examples, we point out that the computational burden associated with the DP is tremendous. To solve the DP problem with the system scale size $n = 1{,}000$, it requires 20 gigabytes of memory and six hours of computing. Under the same hardware setting, the FCP obtains its optimal solution in less than a minute, and the memory required is negligible compared with that of the DP case. For the small to moderate-size $n$, the DP problem still requires significantly more computing resources. We acknowledge that when the system scale is small, the FCP results do deviate significantly from the DP solutions. However, the difference decreases quickly as the system scale increases. An example with varying system scale is included in Section 6.

There are two streams of literature that are related to our work.

### 1.1. Double-Ended Queueing Models

Double-ended queues have been studied for many applications, including taxi-service systems, perishable-inventory systems, organ-transplant systems, and finance (cf. Kaspi and Perry 1983, Zenios 1999, Prabhakar et al. 2000, Boxma et al. 2011, Afeche et al. 2014, and He et al. 2018). In particular, Kaspi and Perry (1983) studied a perishable-inventory system with Poisson arrivals for both supplies and demands. Supplies are assumed to have constant lifetimes, and demands leave the system if not matched immediately. Extensions of such perishable-inventory system were considered in Kaspi and Perry (1984) and Perry and Stadje (1999). When renewal arrivals and/or generally distributed patience times are considered, exact analysis becomes intractable. In Liu et al. (2015), rigorous fluid and diffusion models were developed for double-ended queues with renewal arrivals and exponential patience times. Later on, Liu (2019) established the heavy traffic asymptotics for the system with renewal arrivals and generally distributed patience times. Double-ended queues are the simplest matching queues. Multiclass matching queues have also drawn attention recently—for example, Gurvich and Ward (2014), Ozkan and Ward (2020), and Khademi and Liu (2019) study matching systems with applications in assemble-to-order systems, ride-sharing systems, and organ-transplant systems and focus on developing asymptotically optimal matching policies.

### 1.2. Staffing and Control of Time-Varying Queues

Heavy-traffic fluid and diffusion limits were developed by Mandelbaum et al. (1998) for time-varying Markovian queueing networks with Poisson arrivals and exponential service times. Gaussian approximation methods for Markovian queues have been developed by Niyirora and Pender (2016) and Pender (2016). Adopting a two-parameter queue-length descriptor, the pioneering work by Whitt (2006) studied the $G/GI/s+GI$ fluid model having nonexponential service and abandonment times. Extending the work by Whitt (2006), Liu and Whitt (2012a) developed a fluid approximation for the $G_t/GI/s_t+GI$ queue with time-varying arrivals and nonexponential distributions; they later extended it to the framework of fluid networks (Liu and Whitt 2011, Liu and Whitt 2014]. A functional weak law of large numbers (Liu and Whitt 2012b) was established to substantiate the fluid approximation in Liu and Whitt (2012a), and a functional central limit theorem (Liu and Whitt 2012c) was developed for the $G_t/M/s_t+GI$ model with exponential service times.

Asymptotic optimal control for time-homogeneous queueing systems is well studied (cf. Harrison 2000, Ata and Kumar 2005, and Budhiraja and Ghosh 2006). For time-inhomogeneous queueing systems, the asymptotic control is usually considered under fluid scaling—cf. Bassamboo et al. (2005), Cudina and Ramanan (2011), Ozkan and Ward (2020), and Khademi and Liu (2019), in all of which high-volume systems were considered, and the FCP was shown to be the best performance bound for the original QCP asymptotically. In Armoney et al. (2019), the authors considered the problem of scheduling appointments for a service facility with customer no-shows. Therein, a similar scaling is considered, and the optimal fluid-limit path exhibits piecewise linear drift in time variable. In Atar et al. (2019), the authors studied a load balancing for time-varying systems and established subdiffusive balance, extending far beyond the heavy-traffic setting. A recent paper (Liu et al. 2021) investigated a scheduling problem for a time-varying multiclass queueing model under a dynamic prioritization rule that is both state-dependent and time-dependent, with an objective of achieving stable and differentiated service levels measured by the so-called *tail probability of delay* (TPoD); also see Liu (2018) for discussions of TPoD and its applications.

Without assuming any asymptotics, in many situations, deterministic models can also be shown to be the best performance bound for the expected stochastic performance. Indeed, our Theorem 3.2 provides such nonasymptotic lower bound on the finite time horizon cumulative costs. To the best of our knowledge, the majority of the existing nonasymptotic results are established for systems with linear drift—for example, in Gallego and Van Ryzin (1994), the deterministic revenue was shown to be the upper bound for its stochastic counterparts. In our work, a piecewise linear drift appears in the fluid process, and our approach for such nonlinearity can be potentially applied for other control problems.

It is worth mentioning that Bassamboo and Randhawa (2010) studied the accuracy of the $M/M/s+GI$ fluid model for capacity sizing, where they quantified the approximation errors and observed that fluid model does not always serve as a lower bound.

**1.2.1. Organization of the Paper.** In Section 2, we introduce our double-ended queueing model with abandonment at both sides and formulate a finite time horizon QCP. In Section 3, we develop a deterministic FCP that provides a lower bound for the QCP. In Section 4, we consider high-volume systems, for which we characterize the asymptotic gap between the optimal values of the QCP and the corresponding FCP and construct an asymptotically optimal production rate for the QCP based on the optimal solution of a suitable FCP.

Scaling-limit theorems are provided to show the asymptotic optimality. Numerical examples to evaluate the effectiveness of the FCP are presented in Section 5. In Section 6, we develop a numerical solution to the QCP using a DP problem to validate the asymptotic optimality of the FCP numerical solution. Additional proofs are given in Section 7, and a short conclusion is given in Section 8. And, last, the appendix collects numerical methods to solve the FCP and the extensions considering some practical constraints.

## 2. Model Formulation

We are motivated by a production/inventory system, in which single commodities are produced according to a Poisson process with a time-varying production rate, and demands arrive following another Poisson process, whose rate also fluctuates over time and further depends on the inventory level. Upon the arrival of demand, if there are available products in the inventory, it will be fulfilled immediately, and if no product is available, it will be backlogged and wait for the upcoming products. Demand fulfillment follows the first-come-first-served principle. We further assume that the products are perishable, and their lifetimes are identically and independently distributed (i.i.d.) exponential random variables, demands are impatient, and their patience times are also i.i.d. exponentially distributed. Such a system can be modeled as a double-ended queueing system, which is schematically depicted in Figure 1.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. All the random variables and stochastic processes in this section are assumed to be defined on this space. The expectation under $\mathbb{P}$ will be denoted by $\mathbb{E}$.

### 2.1. System Processes

A precise mathematical description of the system is given as follows. For $t \geq 0$, let $A_p(t)$ denote the number of goods produced by time $t$, and $A_d(t)$ denote the number of demands arrived by time $t$. Next, let $G_p(t)$ count the total number of perished products by time $t$, and $G_d(t)$ the total number of abandoned demands by time $t$. Let $\{Q(t); t \geq 0\}$ be the queue-length process, and at time $t$, there are $Q^+(t)$ number of products in the inventory, and $Q^-(t)$ number of backlogged demands waiting in the system, where for $x \in \mathbb{R}$, $x^+ \equiv \max(x, 0)$ and $x^- \equiv \max(-x, 0)$, and, thus, $Q^+(t)$ and $Q^-(t)$ denote the positive and negative parts of $Q(t)$, respectively.

Flow conservation yields

$$Q(t) = Q(0) + A_p(t) - A_d(t) - G_p(t) + G_d(t), \quad t \geq 0. \tag{1}$$

Let $N_i, i = 1, 2, 3, 4$ be four independent unit rate Poisson processes, and then we can formulate the system processes as follows.

$$\begin{aligned} A_p(t) &\equiv N_1\left(\int_0^t \lambda_p(s)ds\right), \quad A_d(t) \equiv N_2\left(\int_0^t \lambda_d(s, Q(s))ds\right), \\ G_p(t) &\equiv N_3\left(\theta_p \int_0^t Q^+(s)ds\right), \quad G_d(t) \equiv N_4\left(\theta_d \int_0^t Q^-(s)ds\right), \end{aligned} \tag{2}$$

where $\lambda_p : [0, \infty) \to [0, \infty)$, $\lambda_d : [0, \infty) \times \mathbb{R} \to [0, \infty)$ are measurable functions that represent the production rate and demand arrival rate, respectively. Parameters $\theta_p \geq 0$ and $\theta_d \geq 0$ are the goods-perishment and back-order-abandonment rates (the reciprocals of the means of product shelf life and demand patience time). For the production system with nonperishable goods or infinitely patient customers, we can let $\theta_p = 0$ or $\theta_d = 0$, which is a special case of our model here. We allow the demand rate to depend on the inventory level; such dependence may be attributed to the "selective effect" and "advertising effect" known in the marketing literature; see Khmelnitsky and Gerchak (2002) and Corstjens and Doyle (1981).

### 2.2. A Queueing Control Problem

The goal of the production manager is to minimize the expected cost functional over a finite time horizon $[0, T]$ by controlling the production rate $\{\lambda_p(t); t \in [0, T]\}$. More precisely, denote by $h_p$ and $h_d$ the holding costs for each product in inventory and each backlogged demand, $c_p$ and $c_d$ the penalty costs for each perished product and lost demand, and $C : [0, \infty) \to [0, \infty)$ the convex cost function for production. Furthermore, to quantify the degree to which the production rate varies, we denote the total variation of the production-rate function $\{\lambda_p(t); t \in [0, T]\}$ by

$$V_T(\lambda_p) \equiv \sup \sum_{i=0}^{n-1} \left|\lambda_p(t_{i+1}) - \lambda_p(t_i)\right|,$$

where the supremum runs over the set of all partitions $\{0 = t_0 \leq \cdots \leq t_n = T\}$.

We associate $V_T(\lambda_p)$ with a penalty cost $c_f$ to penalize the rapid fluctuations of the production rate.

Let $\mathcal{M} = (Q(0), \lambda_d, \Lambda_p, \theta_p, \theta_d, C, h_p, h_d, c_p, c_d, c_f)$ denote the input data of the system. Our QCP is to choose $\{\lambda_p(t); t \in [0,T]\} \in \mathcal{U}$ to minimize

$$\mathcal{R}(\lambda_p; \mathcal{M}) \equiv \mathbb{E}\left(\int_0^T \left[h_p Q^+(t) + h_d Q^-(t) + C(\lambda_p(t))\right]dt\right) + \mathbb{E}\left(c_p G_p(T) + c_d G_d(T) + c_f V_T(\lambda_p)\right), \tag{3}$$

where $\mathcal{U}$ is the space of all admissible production-rate functions. A production-rate function $\{\lambda_p(t); t \in [0,T]\}$ is called *admissible* if it satisfies the following conditions:

i. (**Nonanticipativity**) For $t \in [0,T]$,

$$\lambda_p(t) \in \mathcal{F}_t \equiv \sigma\{(Q(s), A_p(s), A_d(s), G_p(s), G_d(s)); 0 \le s \le t\}.$$

The $\sigma$-field $\mathcal{F}_t$ collects all information available to the production manager at time $t$.

ii. (**Boundedness**) For $t \in [0,T]$, $0 \le \lambda_p(t) \le \Lambda_p$, where $\Lambda_p > 0$ is the maximum production rate.

iii. (**Bounded Variation**) $\mathbb{E}[V_T(\lambda_p)] < \infty$.

Unfortunately, the QCP is too complex to be analyzed directly. We, thus, develop an FCP in the next section, which provides a lower bound for the QCP. We further develop an asymptotic framework, and derive an asymptotically optimal production rate for (3) (see Section 4).

## 3. Fluid Control Problem

We formulate a deterministic FCP which is tied to the QCP by serving as a lower bound for (3) (see Theorem 3.2). Structural properties of the FCP, including existence of optimal solutions (Theorem 3.1), linear scalability (Proposition 3.1), and monotonicity of total variation (Proposition 3.2), are presented. The optimal solution of the FCP will be used in Section 4, in which we consider an asymptotic framework and construct an asymptotically optimal control for the QCP (3).

A natural way to develop a fluid model is to consider the expectations of the stochastic processes introduced in Section 2 (see Liu and Whitt 2012b). We note that from (2), for $t \ge 0$,

$$\mathbb{E}(A_p(t)) = \int_0^t \mathbb{E}[\lambda_p(s)]ds,\ \mathbb{E}(A_d(t)) = \int_0^t \mathbb{E}[\lambda_d(s, Q(s))]ds,$$

and

$$\mathbb{E}(G_p(t)) = \int_0^t \theta_p \mathbb{E}[Q^+(s)]ds,\ \mathbb{E}(G_d(t)) = \int_0^t \theta_d \mathbb{E}[Q^-(s)]ds.$$

The objective function (3) can then be written as

$$\begin{aligned}\mathcal{R}(\lambda_p; \mathcal{M}) = &\int_0^T \left[(h_p + c_p\theta_p)\mathbb{E}[Q^+(t)] + (h_d + c_d\theta_d)\mathbb{E}[Q^-(t)]\right]dt \\ &+ \int_0^T \mathbb{E}[C(\lambda_p(t))]dt + c_f \mathbb{E}[V_T(\lambda_p)].\end{aligned}$$

Noting that the cost function $C$, the positive and negative part functionals, and the total variation functional are all convex, from Jensen's inequality, we have

$$\mathbb{E}[Q(t)^+] \ge \mathbb{E}[Q(t)]^+,\ \mathbb{E}[Q(t)^-] \ge \mathbb{E}[Q(t)]^-, \tag{4}$$

$$\mathbb{E}[C(\lambda_p(t))] \ge C(\mathbb{E}[\lambda_p(t)]),\ \mathbb{E}[V_T(\lambda_p)] \ge V_T(\mathbb{E}[\lambda_p]). \tag{5}$$

This yields that $\mathcal{R}(\lambda_p; \mathcal{M}) \ge \tilde{\mathcal{R}}(\mathbb{E}[\lambda_p]; \mathcal{M})$, where

$$\begin{aligned}\tilde{\mathcal{R}}(\mathbb{E}[\lambda_p]; \mathcal{M}) = &\int_0^T \left[(h_p + c_p\theta_p)(\mathbb{E}[Q(t)])^+ + (h_d + c_d\theta_d)(\mathbb{E}[Q(t)])^-\right]dt \\ &+ \int_0^T C(\mathbb{E}[\lambda_p(t)])dt + c_f V_T(\mathbb{E}[\lambda_p]).\end{aligned} \tag{6}$$

We next observe that the expected queue-length process $\mathbb{E}[Q(t)]$ satisfies the following equation. For $t \in [0,T]$,

$$\mathbb{E}[Q(t)] = \mathbb{E}[Q(0)] + \int_0^t \mathbb{E}\big[\lambda_p(s)\big]ds - \int_0^t \mathbb{E}[\lambda_d(s,Q(s))]ds - \int_0^t \theta_p \mathbb{E}[Q^+(s)]ds + \int_0^t \theta_d \mathbb{E}[Q^-(s)]ds. \tag{7}$$

Note that, in general, $\mathbb{E}[Q^+(s)] \neq (\mathbb{E}[Q(t)])^+$ and $\mathbb{E}[Q^-(s)] \neq (\mathbb{E}[Q(t)])^-$, and, hence, $\{\mathbb{E}[Q(t)]; t \in [0,T]\}$ cannot be determined by (7). However, to develop a fluid model, we will replace $\mathbb{E}[Q(t)]$ by a deterministic $q(t)$, $\mathbb{E}[Q^+(s)]$ by $q^+(t)$, and $\mathbb{E}[Q^-(s)]$ by $q^-(t)$ in (6) and (7), and derive the following deterministic control problem, which will be referred to as the fluid control problem associated with $\mathcal{M}$.

**Definition 3.1** (FCP). *The FCP associated with $\mathcal{M}$ is to choose a deterministic function $\{\bar{\lambda}_p(t); t \in [0,T]\}$ to minimize*

$$\bar{\mathcal{R}}(\bar{\lambda}_p; \mathcal{M}) \equiv \int_0^T \big[(h_p + c_p\theta_p)q^+(t) + (h_d + c_d\theta_d)q^-(t) + C(\bar{\lambda}_p(t))\big]dt + c_f V_T(\bar{\lambda}_p), \tag{8}$$

*subject to*

i. *For $t \in [0,T]$,*

$$q(t) = \mathbb{E}[Q(0)] + \int_0^t \Big[\bar{\lambda}_p(s) - \lambda_d\big(s, q(s)\big) - \theta_p q^+(s) + \theta_d q^-(s)\Big]ds; \tag{9}$$

ii. *For $t \in [0,T]$, $0 \le \bar{\lambda}_p(t) \le \Lambda_p$;*

iii. *$V_T(\bar{\lambda}_p) < \infty$.*

Throughout this section, we make the following assumptions, which contain some natural regularity conditions on the demand arrival-rate function $\lambda_d$ and the production cost function $C$. In particular, the local Lipschitz continuity assumption (11) below will guarantee the existence of the solution to the (fluid) state process $\{q(t); t \in [0,T]\}$ defined by (9), and the linear growth assumption (10) below will be used to show the uniform boundedness of $\{q(t); t \in [0,T]\}$. Such assumptions are standard for state-dependent rate functions, for example, in Mandelbaum and Pats (1998). Lastly, Assumption 1(ii) below is required for the existence of an optimal solution to the FCP (see Theorem 3.1).

**Assumption 1.**

i. *There exists a positive constant $L_1$ such that for $t \ge 0$ and $x \in \mathbb{R}$,*

$$\lambda_d(t,x) \le L_1(1 + |x|), \tag{10}$$

ii. *and for any compact set $K_1 \times K_2 \subset [0,\infty) \times \mathbb{R}$, there exists a positive constant $L_2$ such that*

$$\sup_{t \in K_1, x, y \in K_2} |\lambda_d(t,x) - \lambda_d(t,y)| \le L_2|x - y|. \tag{11}$$

iii. *The function $C : [0,\infty) \to [0,\infty)$ is continuous.*

The following theorem establishes the existence of an optimal solution to the FCP, which essentially follows from theorem 1.1 of Matula (1987). (The cited theorem guarantees the existence, but not uniqueness, of an optimal solution. In general, uniqueness requires further conditions on the model parameters; cf. Kabe and Gouranga Rao 1986).

**Theorem 3.1** (FCP Existence). *Under Assumption 1, there exists an optimal solution to the FCP.*

**Proof of Theorem 3.1.** Our proof follows from theorem 1.1 in Matula (1987). It suffices to verify the sufficient conditions for that theorem. More precisely, defining for $t \in [0,T]$ and $q \in \mathbb{R}$ and $u \in \mathbb{R}_+$,

$$f(q,u) = (h_p + c_p\theta_p)q^+ + (h_d + c_d\theta_d)q^- + C(u),$$
$$g(t,q,u) = u - \lambda_d(t,q) - \theta_p q^+ + \theta_d q^-,$$

we need to verify the following conditions: (i) $f(\cdot,\cdot)$ is lower semicontinuous on $\mathbb{R}\times\mathbb{R}_+$; (ii) there exists an integrable function $\mu(\cdot)$ such that $\mu(t)\le f(q(t),\lambda_p(t))$ for any admissible pair $(q,\lambda_p)$ and for almost all $t\in[0,T]$; (iii) $g(t,q,u)$ is continuous with respect to $(q,u)$ and measurable with respect to $t$; and (iv) there exists an integrable function $m(\cdot)$ such that $|g(t,q(t),\lambda_p(t))|\le m(t)$ for each admissible pair $(q,\lambda_p)$ and for almost all $t\in[0,T]$.

Clearly $f$ is continuous, and has a lower bound which can be taken to be zero, and, furthermore, $g$ is continuous, and from (10) in Assumption 1, $|g(t,q(t),u(t))|\le\Lambda_p+L_1(1+|q(t)|)+(\theta_d+\theta_p)|q(t)|$. To show the integrability of $q$, we observe that

$$\begin{aligned}|q(t)| &\le |q_0|+\Lambda_p t+\int_0^t\big(L_1\big(1+|q(s)|\big)ds+\theta_p|q(s)|ds+\theta_d|q(s)|\big)ds\\ &= |q_0|+\big(\Lambda_p+L_1\big)t+\big(L_1+\theta_p+\theta_d\big)\int_0^t|q(s)|ds,\end{aligned}$$

and from Gronwall's inequality (see section 1 of the online appendix of Aras et al. 2018 for a reference), we have

$$|q(t)|\le\big(|q_0|+\big(\Lambda_p+L_1\big)t\big)e^{\left(L_1+\theta_p+\theta_d\right)t}. \tag{12}$$

This verifies the sufficient conditions. □

The following lemma confirms the admissibility of an optimal FCP solution for the corresponding QCP.

**Lemma 3.1.** *Let $\bar\lambda_p^*\equiv\{\bar\lambda_p^*(t);t\in[0,T]\}$ be an optimal solution to the FCP associated with $\mathcal{M}$; then, $\bar\lambda_p^*$ is an admissible solution to the QCP associated with $\mathcal{M}$.*

**Proof. Theorem 3.1.** guarantees the existence of $\bar\lambda_p^*$. It suffices to verify that $\bar\lambda_p^*$ satisfies all three admissibility conditions, which are listed below (3)—that is, nonanticipativity, boundedness, and bounded variation. Noting that $\bar\lambda_p^*$ is deterministic, $\bar\lambda_p^*(\cdot)\in\mathcal{F}_0\subset\mathcal{F}_t$ for $t\in[0,T]$. So the nonanticipativity condition is trivially satisfied. In Definition 3.1, the second constraint requires that any feasible solution to the FCP is bounded from above by $\Lambda_p$, which verifies the boundedness condition. Lastly, the existence of an optimal solution to the FCP ensures that the optimal value of the objective function is finite—that is, $0\le\bar{\mathcal{R}}(\bar\lambda_p^*;\mathcal{M})<\infty$—which implies that the total variation of the optimal solution is also bounded—that is, $V_T(\bar\lambda_p^*)<\infty$. It follows now that $\bar\lambda_p^*$ is an admissible solution to the corresponding QCP. □

In the following, we impose stronger assumptions on the demand arrival rate function and the production-cost functional. These assumptions will be required in some of the following results.

**Assumption 2.**

i. *The demand arrival rate is bounded and state-independent—that is, for $t\ge0$ and $x\in\mathbb{R}$, $\lambda_d(t,x)\equiv\lambda_d(t)$ and* $\sup_{t\ge0}\lambda_d(t)<\infty$.

ii. *The production cost $C(\cdot)$ is linear—that is, for some $C_0>0$, $C(x)=C_0x$, $x\in\mathbb{R}_+$.*

We now study the linear-scalability property of the FCP. The scalable FCP is proportional to the increasing demand input. In the high-volume system, the associated FCP can be scaled properly for computational convenience (see Section 5). Let $\mathcal{V}(\mathcal{M})$ denote the optimal value of the FCP with given data $\mathcal{M}$—that is,

$$\mathcal{V}(\mathcal{M})\equiv\min_{\bar\lambda_p}\bar{\mathcal{R}}\left(\bar\lambda_p;\mathcal{M}\right).$$

For a constant $\kappa>0$, define the linearly scaled data with respect to $Q(0)$, $\lambda_d$, and $\Lambda_p$ as follows:

$$\mathcal{M}^\kappa\equiv\left(\kappa Q(0),\kappa\lambda_d,\kappa\Lambda_p,\theta_p,\theta_d,C,h_p,h_d,c_p,c_d,c_f\right).$$

**Proposition 3.1** (Linear-Scalability Property)**.** *Under Assumption 2, for any constant $\kappa>0$,*

$$\kappa\mathcal{V}(\mathcal{M})=\mathcal{V}(\mathcal{M}^\kappa). \tag{13}$$

*Furthermore, if $\bar\lambda_p^*$ is an optimal solution to the FCP with $\mathcal{M}$, then $\kappa\bar\lambda_p^*$ is an optimal solution to the FCP with $\mathcal{M}^\kappa$.*

**Proof.** Following Theorem 3.1, there exists $\bar{\lambda}_p^*$ such that $\mathcal{V}(\mathcal{M}) = \bar{\mathcal{R}}(\bar{\lambda}_p^*; \mathcal{M})$. The corresponding optimal state process $q^*$ satisfies,

$$q^*(t) = \mathbb{E}[Q(0)] + \int_0^t \left[\bar{\lambda}_p^*(s) - \lambda_d(s) - \theta_p q^{*,+}(s) + \theta_d q^{*,-}(s)\right] ds.$$

It is straightforward to see that $\kappa\bar{\lambda}_p^*$ and $\kappa q^*$ together satisfy the state process of the FCP with $\mathcal{M}^\kappa$. That is,

$$\kappa q^*(t) = \kappa\mathbb{E}[Q(0)] + \int_0^t \left[\kappa\bar{\lambda}_p^*(s) - \kappa\lambda_d(s) - \theta_p \kappa q^{*,+}(s) + \theta_d \kappa q^{*,-}(s)\right] ds. \tag{14}$$

Hence, $\kappa\bar{\lambda}_p^*$ is an admissible control to the problem $\min_{\bar{\lambda}_p} \bar{\mathcal{R}}(\bar{\lambda}_p, \mathcal{M}^\kappa)$. By the optimality of $\mathcal{V}(\mathcal{M}^\kappa)$,

$$\kappa\mathcal{V}(\mathcal{M}) = \bar{\mathcal{R}}\left(\kappa\bar{\lambda}_p^*, \mathcal{M}^\kappa\right) \geq \mathcal{V}(\mathcal{M}^\kappa), \tag{15}$$

where the first equality is due to the linearity of $\bar{\mathcal{R}}$ with respect to $q^+$, $q^-$ and $\bar{\lambda}_p$. More specifically,

$$\begin{aligned}\bar{\mathcal{R}}\left(\kappa\bar{\lambda}_p^*, \mathcal{M}^\kappa\right) &= \int_0^T \left[c_1 \kappa q^{*,+}(t) + c_2 \kappa q^{*,-}(t) + C\left(\kappa\bar{\lambda}_p^*(t)\right)\right] dt + c_f V_T\left(\kappa\bar{\lambda}_p^*\right) \\ &= \kappa\left(\int_0^T \left[c_1 q^{*,+}(t) + c_2 q^{*,-}(t) + C\left(\bar{\lambda}_p^*(t)\right)\right] dt + c_f V_T\left(\bar{\lambda}_p^*\right)\right) = \kappa\mathcal{V}(\mathcal{M}),\end{aligned}$$

where $c_1 = h_p + c_p\theta_p$ and $c_2 = h_d + c_d\theta_d$.

Similarly, there also exists a control $\bar{\lambda}_{p,\kappa}^*$ for the FCP with $\mathcal{M}^\kappa$ such that $\mathcal{V}(\mathcal{M}^\kappa) = \bar{\mathcal{R}}(\bar{\lambda}_{p,\kappa}^*; \mathcal{M}^\kappa)$ and the resulting state process $q_\kappa^*$ satisfies

$$q_\kappa^*(t) = \kappa\mathbb{E}[Q(0)] + \int_0^t \left[\bar{\lambda}_{p,\kappa}^*(s) - \kappa\lambda_d(s) - \theta_p q_\kappa^{*,+}(s) + \theta_d q_\kappa^{*,-}(s)\right] ds.$$

Next, $\bar{\lambda}_{p,\kappa}^*/\kappa$ can be shown to be an admissible control to the FCP with input $\mathcal{M}$—that is,

$$\frac{q_\kappa^*(t)}{\kappa} = \mathbb{E}[Q(0)] + \int_0^t \left[\frac{\bar{\lambda}_{p,\kappa}^*(s)}{\kappa} - \lambda_d(s) - \theta_p \frac{q_\kappa^{*,+}(s)}{\kappa} + \theta_d \frac{q_\kappa^{*,-}(s)}{\kappa}\right] ds,$$

where the state process is $q_\kappa^*/\kappa$. Noting that $\mathcal{V}(\mathcal{M})$ is the optimal value for the FCP associated with $\mathcal{M}$,

$$\frac{\mathcal{V}(\mathcal{M}^\kappa)}{\kappa} = \bar{\mathcal{R}}\left(\frac{\bar{\lambda}_{p,\kappa}^*}{\kappa}, \mathcal{M}\right) \geq \mathcal{V}(\mathcal{M}). \tag{16}$$

Together with (15), we have the equality (13). Furthermore, given the optimal rate $\bar{\lambda}_p^*$ for FCP with $\mathcal{M}$, we have confirmed in (14) that $\kappa\bar{\lambda}_p^*$ is also an admissible control to FCP with $\mathcal{M}^\kappa$, where the state process is given by $\kappa q^*$. Additionally, because of (13), we know that $\bar{\mathcal{R}}(\kappa\bar{\lambda}_p^*; \mathcal{M}^\kappa) = \kappa\mathcal{V}(\mathcal{M}) = \mathcal{V}(\mathcal{M}^\kappa)$. Therefore, we conclude that $\kappa\bar{\lambda}_p^*$ is an optimal solution to FCP with $\mathcal{M}^\kappa$. □

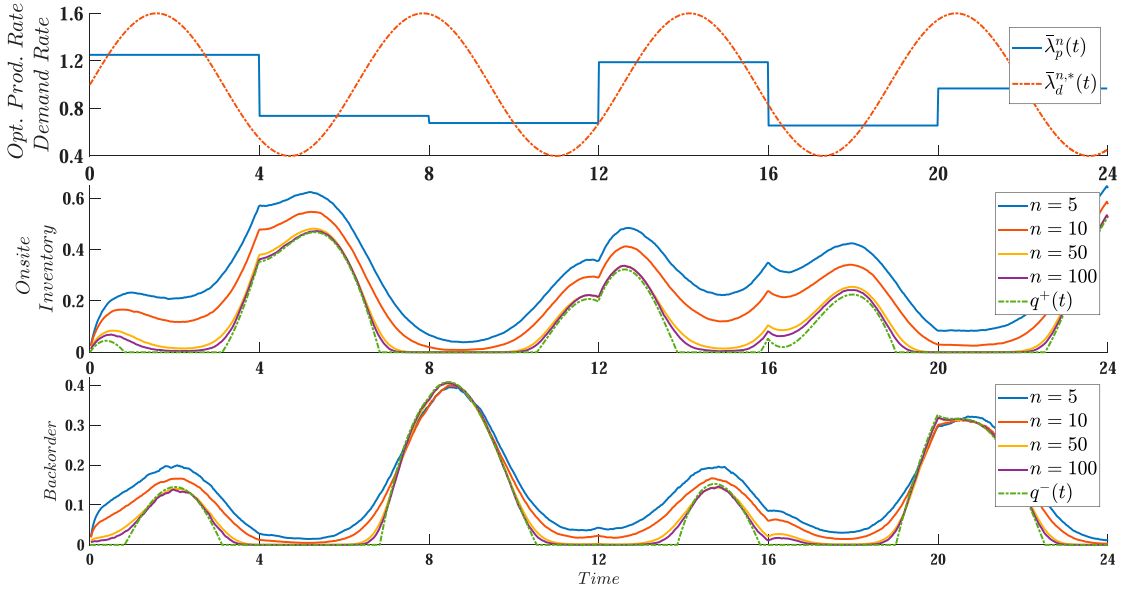
One salient feature of our control problem is the inclusion of the cost associated with total variation of the production-rate function. The following result studies the monotonicity of the total variation of the optimal production rate with respect to the associated cost $c_f$. A numerical example is provided in Figure 3, which shows that when $c_f$ is within a moderate range, the total variation term affects the optimal fluid production rate in a nontrivial manner.

**Proposition 3.2** (Monotonicity of Total Variation)**.** *Under Assumption 1, fix all the input data in $\mathcal{M}$ except the flexibility cost $c_f$, and let $\bar{\lambda}_p^*(c_f) \equiv \{\bar{\lambda}_p^*(t; c_f); t \in [0, T]\}$ be an optimal solution of the FCP for the given $c_f$. The total variation $V_T(\bar{\lambda}_p^*(c_f))$ is monotonically decreasing in $c_f$.*

**Proof.** Let

$$f(\bar{\lambda}_p) = \int_0^T \left[(h_p + c_p\theta_p)q^+(t) + (h_d + c_d\theta_d)q^-(t) + C(\bar{\lambda}_p(t))\right] dt.$$

**Figure 2.** $n = 5, 10, 50, 100$




*Notes.* Let $T = 24$, demand rate $\lambda_d^n(t) = n\bar{\lambda}_d(t)$, $\bar{\lambda}_d(t) = 1 + 0.6\sin(t)$, and $\theta_p^n = 0.5$, $\theta_d^n = 2$, and unit costs are $c_d = 5$, $c_p = 3$, $h_d = 5$, $h_d = 3$, $c_f = 1$, $C_0 = 2$. Allow production-rate change to occur every $L = 4$. Opt. Prod. Rate, optimal production rate.

Define an optimal value function parameterized by $c_f$ as

$$V(c_f) = \bar{\mathcal{R}}\left(\bar{\lambda}_p^*(c_f), c_f\right),$$

which can be also written as

$$V(c_f) = \min_{\bar{\lambda}_p} \bar{\mathcal{R}}(\bar{\lambda}_p; c_f) = \min_{\lambda_p}\left[f(\bar{\lambda}_p) + c_f V_T(\bar{\lambda}_p)\right],$$

where the $\mathcal{R}(\bar{\lambda}_p; c_f)$ is linear in $c_f$. Recall that the minimum of a family of linear functions forms a concave function. Therefore, $V(c_f)$ is concave in its only parameter $c_f$.

Following the envelop theorem in Milgrom and Segal (2002), which concerns the differentiability properties of the objective function of a parameterized optimization problem and, in this case, determines the value of the derivative in (17), we have

$$\frac{dV(c_f)}{dc_f} = V_T\left(\bar{\lambda}_p^*(c_f)\right) \geq 0. \tag{17}$$

The positive part of the first-order derivative of a concave function must be decreasing in its parameter. Hence, the total variation of the optimal production rate $V_T(\bar{\lambda}_p^*(c_f))$ as a function is decreasing in the flexibility cost $c_f$. □

The next theorem establishes that the FCP (8) provides a lower bound for the QCP (3).

**Theorem 3.2** (Lower-Boundedness of the FCP)**.** *Under Assumption 2, for any admissible control $\lambda_p \in \mathcal{U}$ of the queueing control problem (3), we have*

$$\mathcal{R}(\lambda_p; \mathcal{M}) \geq \bar{\mathcal{R}}(\mathbb{E}[\lambda_p]; \mathcal{M}). \tag{18}$$

**Proof of Theorem 3.2.** We note that both the cost functionals $\mathcal{R}$ and $\bar{\mathcal{R}}$ consist of queue-length holding cost, production cost, and flexibility cost. As discussed in (5), under any admissible control $\lambda_p$ for the QCP, $\mathbb{E}[C(\lambda_p(t))] \geq C(\mathbb{E}[\lambda_p(t)])$ and $\mathbb{E}[V_T(\lambda_p)] \geq V_T(\mathbb{E}[\lambda_p])$. Thus, it suffices to establish that

$$\int_0^T \left[(h_p + c_p\theta_p)\mathbb{E}[Q^+(t)] + (h_d + c_d\theta_d)\mathbb{E}[Q^-(t)]\right] dt \geq \int_0^T \left[(h_p + c_p\theta_p)q^+(t) + (h_d + c_d\theta_d)q^-(t)\right] dt, \tag{19}$$

where for $t \in [0,T]$,

$$q(t) = \mathbb{E}[Q(0)] + \int_0^t \big[\mathbb{E}\big[\lambda_p(s)\big] - \lambda_d(s) - \theta_p q^+(s) + \theta_d q^-(s)\big]ds. \tag{20}$$

We first treat the special case $\theta_p = \theta_d$. Assume that $\theta_p = \theta_d = \theta$. Then,

$$\theta_p \mathbb{E}[Q(t)^+] - \theta_d \mathbb{E}[Q(t)^-] = \theta \mathbb{E}[Q(t)].$$

Hence, $\{(\mathbb{E}[\lambda_p(t)], \mathbb{E}[Q(t)]); t \in [0,T]\}$ satisfies (20), and the inequality (19) follows from Jensen's inequality.

We now consider the case $\theta_p \neq \theta_d$. We consider an auxiliary problem defined as the following:

$$\min_{\{u(t);t\in[0,T]\}} \check{\mathcal{R}}(u) \equiv \int_0^T \big[(h_p + c_p\theta_p)q^+(t) + (h_d + c_d\theta_d)q^-(t) + \big(h_p + c_p\theta_p + h_d + c_d\theta_d\big)u(t)\big]dt, \tag{21}$$

subject to, for $t \in [0,T]$,

$$q(t) = \mathbb{E}[Q(0)] + \int_0^t \big[d(s) - \theta_p q^+(s) + \theta_d q^-(s) - \big(\theta_p - \theta_d\big)u(s)\big]ds, \tag{22}$$

$$0 \le u(t) \le \Delta_0, \tag{23}$$

where $\Delta_0$ is a given constant and $d(t) = \mathbb{E}[\lambda_p(t)] - \lambda_d(t)$.

**Lemma 3.2.** *Assume that $\theta_p \neq \theta_d$ and fix an admissible control $\lambda_p \in \mathcal{U}$. Then, under Assumption 2, $u^*(t) = 0, t \in [0,T]$, is an optimal control for the problem (21)–(23).*

The proof of Lemma 3.2 is given in Section 7. In the following, we finish the proof of Theorem 3.2. Let

$$u_1(t) \equiv \mathbb{E}[Q^+(t)] - \mathbb{E}[Q(t)]^+ = \mathbb{E}[Q^-(t)] - \mathbb{E}[Q(t)]^- \ge 0.$$

Now, the constraint (7) becomes

$$\mathbb{E}[Q(t)] = \mathbb{E}[Q(0)] + \int_0^t \big[\mathbb{E}\big[\lambda_p(t)\big] - \lambda_d(t) - \theta_p \mathbb{E}[Q(s)]^+ + \theta_d \mathbb{E}[Q(s)]^- - \big(\theta_p - \theta_d\big)u_1(s)\big]ds.$$

Furthermore, Lemma 4.1 shows that $\sup_{0\le t\le T} \mathbb{E}[|Q(t)|] < \infty$, which yields that there exists $\Delta_0 > 0$ such that $u_1(t) \le \Delta_0$ for all $t \in [0,T]$. This shows that $u_1(t)$ is an admissible control to the auxiliary control problem in Lemma 3.2, and the corresponding state process is $\{\mathbb{E}[Q(t)]; t \in [0,T]\}$.

From Lemma 3.2, $u^*(t) = 0, t \in [0,T]$ is an optimal solution, and we have

$$\check{\mathcal{R}}(u_1) = \int_0^T \big[(h_p + c_p\theta_p)\mathbb{E}[Q^+(t)] + (h_d + c_d\theta_d)\mathbb{E}[Q^-(t)]\big]dt \ge \int_0^T \big[(h_p + c_p\theta_p)q^+(t) + (h_d + c_d\theta_d)q^-(t)\big]dt = \check{\mathcal{R}}\big(u^*\big),$$

where

$$q(t) = \mathbb{E}[Q(0)] + \int_0^t \big[\mathbb{E}\big[\lambda_p(s)\big] - \lambda_d(s) - \theta_p q^+(s) + \theta_d q^-(s)\big]ds.$$

The theorem now follows. □

**Remark 3.1.** The exact analysis of the QCP is of challenge mainly due to the intractability that stems from: (i) the nonlinearity of holding costs, (ii) the nonstationarity of demand rate, and (iii) the total-variation term in the cost functional. The FCP, however, is a continuous-time, continuous-space optimal control problem. The standard solution technique is to apply Pontryagin's Maximum Principle; see Pontryagin (2018) and Seierstad and Sydsaeter (1986). The facts that the drift of the FCP state process as in (9) is piece-wise and there is involvement

of the total variation term render it difficult to obtain a closed-form solution. Therefore, we resort to a discrete-time LP reformulation that can be efficiently solved by commercial solvers. See Section A.1 in the appendix for details of the LP reformulation. A numerical example on the sensitivity analysis of the discrete-time interval is also included in Section 5, where we show the convergence of the LP optimal solution as the discrete-time interval shrinks. The proof of the convergence of the LP solution to that of the FCP is outside the scope of this paper.

The FCP provides a convenient performance lower bound. Theorem 3.2 states the existence of such lower bound for any admissible control. In Section 4, we introduce the notion of system scale and consider a sequence of QCPs indexed by the increasing system scale. We establish a heavy traffic-limit theorem (Theorem 4.2), which extends Theorem 3.2 and shows that the FCP lower bound is asymptotically tight. Section 5 provides numerical examples to evaluate the tightness of the FCP lower bound when the system scale varies. Also, see Section 6 for exact solutions to QCP via a dynamic-programming approach.

## 4. Asymptotic Optimality of FCP

In this section, we develop an asymptotic framework, in which the performance of the suitably scaled QCP attains the FCP lower bound asymptotically, which extends Theorem 3.2. We first introduce a scaling parameter $n$ that can be considered as the maximum possible quantity of demand at any given time point on a finite interval $[0,T]$ when offering the lowest possible price. For example, $n$ stands for the potential market size, and our key assumption is that such a market size is large. Without loss of generality, we assume $n$ takes an integer value. Our asymptotic analysis makes it possible to embed the underlying production system onto a sequence of systems indexed by $n$.

More precisely, we consider a sequence of double-ended queues considered in Section 2, and for the $n^{\text{th}}$ system, we append a superscript $n$ to the quantities introduced in Section 2. In particular, we have $\lambda_p^n$, $\lambda_d^n, \theta_p^n, \theta_d^n$ to denote the production rate, demand-arrival rate, product-perishment rate, and demand-abandonment rate, respectively. We will assume that $\lambda_d^n$ is $\mathcal{O}(n)$, and $\theta_p^n$ and $\theta_d^n$ are $\mathcal{O}(1)$ (see (25) and Assumption 3, (i) and (ii)). The unit-rate Poisson processes, arrival processes, abandonment processes, and queue-length process are denoted by $N_i^n, i=1,2,3,4$, $A_p^n, A_d^n, G_p^n, G_d^n$, and $Q^n$, respectively, which are defined on a complete probability space $(\Omega^n, \mathbb{P}^n, \mathcal{F}^n)$. The expectation under $\mathbb{P}^n$ is denoted by $\mathbb{E}^n$, and, for convenience, we omit the parameter $n$ and simply use $\mathbb{P}$ and $\mathbb{E}$. We assume that all the costs are independent of $n$.

The control problem in the $n^{\text{th}}$ system is to choose $\{\lambda_p^n(t); t\in[0,T]\} \in \mathcal{U}^n$ with an objective of minimizing

$$\mathcal{R}^n\left(\lambda_p^n; \mathcal{M}^n\right) \equiv \mathbb{E}\left(\int_0^T \left[h_p Q^{n,+}(t) + h_d Q^{n,-}(t) + C\left(\lambda_p^n(t)\right)\right]dt + c_p G_p^n(T) + c_d G_d^n(T) + c_f V_T\left(\lambda_p^n\right)\right), \tag{24}$$

where $\mathcal{M}^n = (Q^n(0), \lambda_d^n, \Lambda_p^n, \theta_p^n, \theta_d^n, C, h_p, h_d, c_p, c_d, c_f)$, and $\mathcal{U}^n$ is the space of all admissible production-rate functions in the $n^{\text{th}}$ system. A production-rate function $\{\lambda_p^n(t); t\in[0,T]\}$ is *admissible* in the $n^{\text{th}}$ system if it satisfies the following conditions:

i. (**Nonanticipativity**) for $t\in[0,T]$,

$$\lambda_p^n(t) \in \mathcal{F}_t^n \equiv \sigma\left(\left(Q^n(s), A_p^n(s), A_d^n(s), G_p^n(s), G_d^n(s)\right); 0\le s\le t\right);$$

ii. ($O(n)$**-boundedness**) for $t\in[0,T]$, $0\le\lambda_p^n(t)\le\Lambda_p^n = n\bar{\Lambda}_p$, where $\bar{\Lambda}_p$ is a positive constant;

iii. (**Bounded Variation**) $\mathbb{E}[V_T(\lambda_p^n)] < \infty$.

### 4.1. Gap Between the QCP and the Corresponding FCP

For each $n\in\mathbb{N}$, suppose Assumption 2 holds for the $n^{\text{th}}$ system—that is, for $x\in\mathbb{R}^+$, $C^n(x) = C_0^n x$ for some $C_0^n > 0$, and for $t\ge 0$ and $x\in\mathbb{R}$, $\lambda_d^n(t,x) \equiv \lambda_d^n(t)$ and $\sup_{t\ge0}\lambda_d^n(t) < \infty$. From Theorem 3.1, the FCP associated with $\mathcal{M}^n$ admits an optimal solution and denote it by $\tilde{\lambda}_p^{n,*} \equiv \{\tilde{\lambda}_p^{n,*}(t); t\in[0,T]\}$. Lemma 3.1 ensures that $\tilde{\lambda}_p^{n,*}$ is an admissible solution to the QCP associated with the same $\mathcal{M}^n$. Let $\mathcal{V}^n(\mathcal{M}^n)$ and $\bar{\mathcal{V}}^n(\mathcal{M}^n)$ denote the optimal values of the QCP and the associated FCP for the $n^{\text{th}}$ system with $\mathcal{M}^n$.

**Theorem 4.1.** *For each $n \in \mathbb{N}$, suppose Assumption 2 holds for the $n^{\text{th}}$ system—that is, for $x \in \mathbb{R}^+$, $C^n(x) = C_0^n x$ for some $C_0^n > 0$, and for $t \geq 0$ and $x \in \mathbb{R}$, $\lambda_d^n(t,x) \equiv \lambda_d^n(t)$ and $\sup_{t\geq 0} \lambda_d^n(t) < \infty$. We further assume that*

$$\sup_{n\in\mathbb{N}} \mathbb{E}\left(\left(\bar{Q}^n(0)\right)^2\right) < \infty, \sup_{n\in\mathbb{N}} \mathrm{Var}(Q^n(0))/n < \infty,$$
$$\sup_{n\in\mathbb{N}} \sup_{t\geq 0} \lambda_d^n(t)/n < \infty, \sup_{n\in\mathbb{N}} \theta_p^n < \infty, \sup_{n\in\mathbb{N}} \theta_d^n < \infty. \tag{25}$$

*Then, for some $\kappa_0 > 0$,*

$$0 \leq \mathcal{V}^n(\mathcal{M}^n) - \bar{\mathcal{V}}^n(\mathcal{M}^n) \leq \mathcal{R}^n\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) - \bar{\mathcal{R}}\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) \leq \kappa_0\sqrt{n}. \tag{26}$$

**Remark 4.1.** Theorem 4.1 says for a large-scale system, under the stronger Assumption 2, applying the optimal production rate $\{\tilde{\lambda}_p^{n,*}(t); t \in [0,T]\}$ of the corresponding FCP yields an error $O(\sqrt{n})$.

**Proof of Theorem 4.1.** From Theorem 3.2 on the lower-boundedness of FCP, we have

$$\mathcal{R}^n\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) \geq \bar{\mathcal{R}}\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) = \bar{\mathcal{V}}^n(\mathcal{M}^n), \text{ and } \mathcal{V}^n(\mathcal{M}^n) \geq \bar{\mathcal{V}}^n(\mathcal{M}^n).$$

Noting that $\mathcal{V}^n(\mathcal{M}^n) \leq \mathcal{R}^n(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n)$, we have

$$\mathcal{V}^n(\mathcal{M}^n) - \bar{\mathcal{V}}^n(\mathcal{M}^n) \leq \mathcal{R}^n\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) - \bar{\mathcal{R}}\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right).$$

In view of (26), it suffices to show $\mathcal{R}^n(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n) - \bar{\mathcal{R}}(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n) = O(\sqrt{n})$. For the rest of the proof, the systems we consider are under the control $\{\tilde{\lambda}_p^{n,*}\}$. For each $n \in \mathbb{N}$, define

$$\check{q}^n(t) = \mathbb{E}[Q^n(0)] + \int_0^t \left[\tilde{\lambda}_p^{n,*}(s) - \lambda_d^n(s) - \theta_p^n \check{q}^{n,+}(s) + \theta_d^n \check{q}^{n,-}(s)\right] ds.$$

We observe that

$$\begin{aligned}
&\mathcal{R}^n\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) - \bar{\mathcal{R}}\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) \\
&= \int_0^T h_p \mathbb{E}[Q^{n,+}(t)] + h_d \mathbb{E}[Q^{n,-}(t)] + C^n\left(\tilde{\lambda}^{n,*}(t)\right) dt + c_p \mathbb{E}\left[G_p^n(T)\right] + c_d \mathbb{E}[G^n(T)] + c_f V_T\left(\tilde{\lambda}_p^{n,*}\right) \\
&- \int_0^T \left(h_p + c_p\theta_p^n\right)\check{q}^{n,+}(t) + \left(h_d + c_d\theta_d^n\right)\check{q}^{n,-}(t) + C^n\left(\tilde{\lambda}_p^{n,*}(t)\right) dt - c_f V_T\left(\tilde{\lambda}_p^{n,*}\right) \\
&= \int_0^T h_p\left(\mathbb{E}\left[Q^{n,+}(t) - \check{q}^{n,+}(t)\right]\right) + h_d\left(\mathbb{E}\left[Q^{n,-}(t) - \check{q}^{n,-}(t)\right]\right) dt \\
&\quad + c_p\left(\mathbb{E}\left[G_p^n(T)\right] - \theta_p^n \int_0^T \check{q}^{n,+}(t) dt\right) + c_d\left(\mathbb{E}\left[G_d^n(T)\right] - \theta_d^n \int_0^T \check{q}^{n,-}(t) dt\right)
\end{aligned}$$

We note that for $t \geq 0$,

$$\mathbb{E}\left[G_p^n(T)\right] = \theta_p^n \int_0^T \mathbb{E}[Q^{n,+}(t)] dt, \quad \mathbb{E}\left[G_d^n(T)\right] = \theta_d^n \int_0^T \mathbb{E}[Q^{n,-}(t)] dt.$$

Hence,

$$\mathcal{R}^n\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) - \bar{\mathcal{R}}\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) = \int_0^T \left(h_p + c_p\theta_p^n\right)\left(\mathbb{E}[Q^{n,+}(t)] - \check{q}^{n,+}(t)\right) + \left(h_d + c_d\theta_d^n\right)\left(\mathbb{E}[Q^{n,-}(t)] - \check{q}^{n,-}(t)\right) dt.$$

Using the inequalities $|x^+ - y^+| \le |x-y|$ and $|x^- - y^-| \le |x-y|$ for $x, y \in \mathbb{R}$, we have that

$$\mathcal{R}^n\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) - \bar{\mathcal{R}}\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) \le \left(h_p + h_d + c_p\theta_p^n + c_d\theta_d^n\right) \int_0^T \mathbb{E}|Q^n(t) - \check{q}^n(t)|dt. \tag{27}$$

We claim that for some $c_0 > 0$, which is independent of $n$,

$$\frac{1}{\sqrt{n}} \sup_{t\in[0,T]} \mathbb{E}\left[\left|Q^n(t) - \check{q}^n(t)\right|\right] \le c_0. \tag{28}$$

Using the claim (28), we have

$$\frac{1}{\sqrt{n}}\left(\mathcal{R}^n\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right) - \bar{\mathcal{R}}\left(\tilde{\lambda}_p^{n,*}; \mathcal{M}^n\right)\right) \le \left(h_p + h_d + c_p\theta_p^n + c_d\theta_d^n\right) T c_0,$$

and the theorem follows. Finally, we prove the claim (28). For $t \ge 0$,

$$\begin{aligned}\frac{1}{\sqrt{n}}\left(Q^n(t) - \check{q}^n(t)\right) &= \frac{1}{\sqrt{n}}\left(Q^n(0) - \mathbb{E}[Q^n(0)]\right) + \hat{N}_1^n\left(\int_0^t \frac{\tilde{\lambda}_p^{n,*}(s)}{n} ds\right) - \hat{N}_2^n\left(\int_0^t \frac{\lambda_d^n(s)}{n} ds\right) \\ &\quad - \hat{N}_3^n\left(\theta_p^n \int_0^t \bar{Q}^{n,+}(s) ds\right) + \hat{N}_4^n\left(\theta_d^n \int_0^t \bar{Q}^{n,-}(s) ds\right),\end{aligned} \tag{29}$$

where $\hat{N}_i^n(t) = (N_i^n(nt) - nt)/\sqrt{n}$, $t \ge 0$, and $i = 1, \ldots, 4$ being the diffusion-scaled Poisson processes, and $\bar{Q}^n(t) = Q^n(t)/n$ for $t \ge 0$ being the fluid-scaled queue-length process that will be formally introduced in the next section. From functional central-limit theorem for Poisson processes, $\hat{N}_i^n$, $i = 1, \ldots, 4$, is a martingale with respect to the $\sigma$-fields generated by itself. Using (25), and Doob's inequality, for $t \ge 0$,

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{N}_1^n\left(\int_0^t \frac{\tilde{\lambda}_p^{n,*}(s)}{n} ds\right)\right)^2 + \left(\hat{N}_2^n\left(\int_0^t \bar{\lambda}_d^n(s) ds\right)\right)^2\right] \\ &\quad \le \mathbb{E}\left(\sup_{0\le t\le \Lambda_p T}\left(\hat{N}_1^n(t)\right)^2\right) + \mathbb{E}\left(\sup_{0\le t\le C_1 T}\left(\hat{N}_2^n(t)\right)^2\right) \\ &\quad \le 4\mathbb{E}\left(\hat{N}_1^n(\Lambda_p T)\right)^2 + 4\mathbb{E}\left(\hat{N}_2^n(C_1 T)\right)^2 \\ &\quad = 4T\left(\Lambda_p + C_1\right) < \infty,\end{aligned} \tag{30}$$

where $C_1 = \sup_n \sup_t \lambda_d(t)/n$. Next, we observe that $\hat{N}_3^n(\theta_p^n \int_0^\cdot \bar{Q}^{n,+}(s) ds)$ and $\hat{N}_4^n(\theta_d^n \int_0^\cdot \bar{Q}^{n,-}(s) ds)$ are both $\{\mathcal{F}_t\}$ martingales with the following quadratic variations

$$\left[\hat{N}_3^n\left(\theta_p^n \int_0^\cdot \bar{Q}^{n,+}(s) ds\right), \hat{N}_3^n\left(\theta_p^n \int_0^\cdot \bar{Q}^{n,+}(s) ds\right)\right]_t = \bar{N}_3^n\left(\theta_p^n \int_0^t \bar{Q}^{n,+}(s) ds\right),$$

and

$$\left[\hat{N}_4^n\left(\theta_p^n \int_0^\cdot \bar{Q}^{n,-}(s) ds\right), \hat{N}_4^n\left(\theta_p^n \int_0^\cdot \bar{Q}^{n,-}(s) ds\right)\right]_t = \bar{N}_4^n\left(\theta_p^n \int_0^t \bar{Q}^{n,-}(s) ds\right),$$

where $\bar{N}_i^n(t) = N_i^n(nt)/n$ for $t \ge 0$ being the fluid-scaled Poisson processes. It follows that

$$\left(\hat{N}_3^n\left(\theta_p^n \int_0^t \bar{Q}^{n,+}(s) ds\right)\right)^2 - \theta_p^n \int_0^t \bar{Q}^{n,+}(s) ds,$$
$$\left(\hat{N}_4^n\left(\theta_d^n \int_0^t \bar{Q}^{n,-}(s) ds\right)\right)^2 - \theta_d^n \int_0^t \bar{Q}^{n,-}(s) ds,$$

are also $\{\mathcal{F}_t\}$ martingales, and

$$\mathbb{E}\left(\hat{N}_3^n\left(\theta_p^n\int_0^t \bar{Q}^{n,+}(s)ds\right)\right)^2 = \theta_p^n\int_0^t \mathbb{E}\big[\bar{Q}^{n,+}(s)\big]ds, \tag{31}$$

$$\mathbb{E}\left(\hat{N}_4^n\left(\theta_d^n\int_0^t \bar{Q}^{n,-}(s)ds\right)\right)^2 = \theta_d^n\int_0^t \mathbb{E}\big[\bar{Q}^{n,-}(s)\big]ds. \tag{32}$$

See Pang et al. (2007) and the references therein for more details of the martingale representations of the random time change of Poisson processes. To proceed, we introduce the following lemma, which shows that $\{\bar{Q}^n(t); t\in[0,T]\}$ is uniformly integrable.

**Lemma 4.1.** *Assume that* $\sup_{n\in\mathbb{N}}\mathbb{E}[(|\bar{Q}^n(0)|)^2]<\infty$, *and for each* $n$, $\lambda_d^n$ *and* $C^n$ *satisfy Assumption 1. Then, under (25), for any admissible production-rate process* $\{\lambda_p^n(t); t\in[0,T]\}$, *there exists a constant* $L\equiv L(T)$ *such that*

$$\sup_{n\in\mathbb{N}}\mathbb{E}\left(\sup_{0\le t\le T}\left|\bar{Q}^n(t)\right|^2\right)\le L.$$

The proof of Lemma 4.1 can be found in Section 7. Using Lemma 4.1, Hölder's inequality and Jensen's inequality, we see that (31) and (32) have the following estimates:

$$\begin{aligned}\mathbb{E}\left(\hat{N}_3^n\left(\theta_p^n\int_0^t \bar{Q}^{n,+}(s)ds\right)\right)^2 &= \theta_p^n\int_0^t \mathbb{E}\big[\bar{Q}^{n,+}(s)\big]ds\\ &\le \sqrt{\left(\theta_p^n\right)^2\int_0^t\left(\mathbb{E}\big[\bar{Q}^n(s)\big]\right)^2 ds}\\ &\le \sqrt{\left(\theta_p^n\right)^2\int_0^t \mathbb{E}\Big[\left(\bar{Q}^n(s)\right)^2\Big]ds}\\ &\le \theta_p^n\sqrt{LT},\end{aligned} \tag{33}$$

And, similarly,

$$\mathbb{E}\left(\hat{N}_4^n\left(\theta_d^n\int_0^t \bar{Q}^{n,-}(s)ds\right)\right)^2 = \theta_d^n\int_0^t \mathbb{E}\big[\bar{Q}^{n,-}(s)\big]ds \le \theta_d^n\sqrt{LT}. \tag{34}$$

Now back to (29), using Jensen's inequality and the inequality that $(\sum_{i=1}^K c_i)^2\le K\sum_{i=1}^K c_i^2$ for any $c_i\in\mathbb{R}$ and $K\in\mathbb{N}$, we have for each $t\in[0,T]$,

$$\begin{aligned}\left(\frac{1}{\sqrt{n}}\mathbb{E}\big[|Q^n(t)-\breve{q}^n(t)|\big]\right)^2 &\le \mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\left(Q^n(t)-\breve{q}^n(t)\right)\right)^2\right]\\ &\le \frac{5}{n}\mathbb{E}(Q^n(0)-\mathbb{E}[Q^n(0)])^2+5\mathbb{E}\left(\hat{N}_1^n\left(\int_0^t\frac{\tilde{\lambda}_p^{n,*}(s)}{n}ds\right)\right)^2+5\mathbb{E}\left(\hat{N}_2^n\left(\int_0^t\bar{\lambda}_d^n(s)ds\right)\right)^2\\ &\quad+5\mathbb{E}\left(\hat{N}_3^n\left(\theta_p^n\int_0^t\bar{Q}^{n,+}(s)ds\right)\right)^2+5\mathbb{E}\left(\hat{N}_4^n\left(\theta_d^n\int_0^t\bar{Q}^{n,-}(s)ds\right)\right)^2.\end{aligned}$$

Now applying (30), (33), and (34) to the above estimate yields

$$\frac{1}{\sqrt{n}}\mathbb{E}\big[|Q^n(t)-\breve{q}^n(t)|\big]\le\sqrt{5\mathrm{Var}(Q^n(0))/n+20T\left(\Lambda_p+C_1\right)+5\left(\theta_p^n+\theta_d^n\right)\sqrt{LT}},$$

which concludes the claim (28) in view of the assumption (25). □

### 4.2. Asymptotic Optimal Production Rate

In this section, we study the asymptotic properties of the system under Assumption 1 and some convergent properties of the parameter functions (see Assumption 3). We first introduce the fluid-scaled forms of system processes, cost functional, and rate functions. Roughly speaking, in fluid scaling, we scale down the quantity size by $n$. Define

$$\bar{Q}^n(t) \equiv \frac{Q^n(t)}{n},\ \bar{\mathcal{R}}^n\left(\bar{\lambda}_p^n;\mathcal{M}^n\right) \equiv \frac{\mathcal{R}^n\left(\lambda_p^n;\mathcal{M}^n\right)}{n},$$
$$\bar{\lambda}_p^n(t) \equiv \frac{\lambda_p^n(t)}{n},\ \bar{\lambda}_d^n(t,x) \equiv \frac{\lambda_d^n(t,nx)}{n},\ \bar{C}^n(x) \equiv \frac{C(nx)}{n},$$
$$\bar{A}_p^n(t) \equiv \frac{A_p^n(t)}{n},\ \bar{A}_d^n(t) \equiv \frac{A_d^n(t)}{n},\ \bar{G}_p^n(t) \equiv \frac{G_p^n(t)}{n},\ \bar{G}_d^n(t) \equiv \frac{G_d^n(t)}{n}.$$

Consequently, we have the *fluid-scaled control problem*, which minimizes the following fluid-scaled cost functional by controlling the fluid-scaled production rate $\{\bar{\lambda}_p^n(t); t \in [0,T]\}$:

$$\bar{\mathcal{R}}^n\left(\bar{\lambda}_p^n;\mathcal{M}^n\right) \equiv \mathbb{E}\left(\int_0^T \left[h_p\bar{Q}^{n,+}(t) + h_d\bar{Q}^{n,-}(t) + \bar{C}^n\left(\bar{\lambda}_p^n(t)\right)\right]dt + c_p\bar{G}_p^n(T) + c_d\bar{G}_d^n(T) + c_f V_T\left(\bar{\lambda}_p^n\right)\right). \tag{35}$$

Our goal is to find a sequence of admissible production-rate functions $\{\lambda_p^{n,*}\}_{n\geq 1}$, which is *asymptotically optimal*—that is, it satisfies

$$\lim_{n\to\infty} \bar{\mathcal{R}}^n\left(\bar{\lambda}_p^{n,*};\mathcal{M}^n\right) = \inf \liminf_{n\to\infty} \bar{\mathcal{R}}^n\left(\bar{\lambda}_p^n;\mathcal{M}^n\right), \tag{36}$$

where the infimum is taken over all admissible production-rate functions $\{\lambda_p^n \in \mathcal{U}^n\}_{n\geq 1}$. We next introduce the main assumptions on parameters and functions.

**Assumption 3.**

i. *There exist* $\bar{\theta}_p, \bar{\theta}_d \geq 0$ *such that* $\theta_p^n \to \bar{\theta}_p$, $\theta_d^n \to \bar{\theta}_d$, *as* $n \to \infty$.

ii. *There exists a nonnegative measurable function* $\bar{\lambda}_d : [0,\infty)\times\mathbb{R} \to [0,\infty)$ *such that for any* $t \geq 0$ *and any* $L_0 > 0$,

$$\sup_{|x|\leq L_0} \left|\bar{\lambda}_d^n(t,x) - \bar{\lambda}_d(t,x)\right| \to 0, \quad \text{as } n \to \infty. \tag{37}$$

iii. *Furthermore, the function* $\bar{\lambda}_d^n : [0,\infty)\times\mathbb{R} \to [0,\infty)$ *is continuous on* $x \in \mathbb{R}$, *and there exists a* $L_1 > 0$ *such that for* $t \geq 0$ *and* $x \in \mathbb{R}$,

$$\bar{\lambda}_d^n(t,x) \leq L_1(1+|x|), \tag{38}$$

iv. *and for any compact set* $K_1 \times K_2 \subset [0,\infty)\times\mathbb{R}$, *there exists a positive constant* $L_2$ *such that*

$$\sup_{t\in K_1, x,y\in K_2} \left|\bar{\lambda}_d(t,x) - \bar{\lambda}_d(t,y)\right| \leq L_2|x-y|. \tag{39}$$

v. *There exists a continuous function* $\bar{C} : [0,\infty) \to [0,\infty)$ *such that for any* $L_3 > 0$,

$$\sup_{|x|\leq L_3} \left|\bar{C}^n(x) - \bar{C}(x)\right| \to 0, \text{as } n \to \infty. \tag{40}$$

**Remark 4.2.**

i. From Assumption 3, (i) and (ii), we see that the goods-perishment and back-order-abandonment rates $\theta_p^n$ and $\theta_d^n$ are $\mathcal{O}(1)$, while the arrival rate of demands is $\mathcal{O}(n)$.

ii. The assumptions (38) and (39) guarantee the limit arrival rate function $\bar{\lambda}_d(t,x)$ has linear growth and locally Lipschitz continuity in $x$—that is, $\bar{\lambda}_d(t,x)$ satisfies Assumption 1(i). Both assumptions are also required in proving Lemmas 4.1 and 4.2.

iii. From (40) and the definition of $\bar{C}^n(x)$, it is straightforward to see that $\bar{C}(\cdot)$ is linear, and, hence, $\bar{C}^n(\cdot)$ is asymptotically linear. In particular, $\bar{C}(\cdot)$ satisfies Assumption 2(ii).

iv. It is clear that Assumption 3 implies the assumption in (25).

Assume that $\mathbb{E}[|\bar{Q}^n(0) - q_0|^2] \to 0$ for some deterministic point $q_0 \in \mathbb{R}$. Note that if $\bar{\lambda}_p^n(\cdot)$ converges to some nonnegative function $\bar{\lambda}_p(\cdot)$, we would expect that $\bar{Q}^n$ converges to $q$ in probability and uniformly on $[0, T]$ (see Lemma 4.2), where, for $t \in [0, T]$,

$$q(t) = q_0 + \int_0^t \left[\bar{\lambda}_p(s) - \bar{\lambda}_d\big(s, q(s)\big) - \bar{\theta}_p q^+(s) + \bar{\theta}_d q^-(s)\right] ds, \tag{41}$$

and the associated fluid-scaled cost $\bar{\mathcal{R}}^n(\lambda_p^n; \mathcal{M}^n)$ is expected to converge to $\bar{\mathcal{R}}(\bar{\lambda}_p; \bar{\mathcal{M}})$, where $\bar{\mathcal{R}}(\bar{\lambda}_p; \bar{\mathcal{M}})$ is the cost of the FCP associated with $\bar{\mathcal{M}} = (q_0, \bar{\lambda}_d, \bar{\Lambda}_p, \bar{\theta}_p, \bar{\theta}_d, \bar{C}, h_p, h_d, c_p, c_d, c_f)$ (see (8)).

From Theorem 3.1, under Assumption 3, the FCP associated with $\bar{\mathcal{M}}$ admits an optimal solution. Denote by $\bar{\lambda}_p^*$ this optimal solution and $\mathcal{V}(\bar{\mathcal{M}})$ the optimal value. The theorem below establishes the asymptotic optimality of the FCP solution $\bar{\lambda}_p^*$.

**Theorem 4.2** (Asymptotic Optimality of FCP)**.** *Under Assumption 3, $\{n\bar{\lambda}_p^*\}_{n\geq 1}$ is asymptotically optimal for the QCP (24) of the $n^{\text{th}}$ system under the fluid scaling—namely,*

$$\lim_{n\to\infty} \bar{\mathcal{R}}^n\left(\bar{\lambda}_p^*; \mathcal{M}^n\right) = \mathcal{V}(\bar{\mathcal{M}}), \tag{42}$$

*and for any admissible sequence $\{\lambda_p^n\}_{n\geq 1}$ of production-rate functions,*

$$\liminf_{n\to\infty} \bar{\mathcal{R}}^n\left(\bar{\lambda}_p^n; \mathcal{M}^n\right) \geq \mathcal{V}(\bar{\mathcal{M}}). \tag{43}$$

An immediate consequence of Theorem 4.2 gives the following corollary, and we omit its proof.

**Corollary 4.1.** *Under Assumption 3, considering a production rate $\lambda_p^n$ such that*

$$\mathbb{E}\left[\sup_{0\leq t\leq T} \left|\lambda_p^n(t) - n\bar{\lambda}_p^*(t)\right|\right] = o(n),$$

*then $\lambda_p^n$ is asymptotically optimal under the fluid scaling.*

In view of Theorems 4.1 and 4.2, under Assumptions 2 and 3, one may expect that $\tilde{\lambda}_p^{n,*}$, which is an optimal solution of the FCP associated with $\mathcal{M}^n$, satisfies $\tilde{\lambda}_p^{n,*}/n \to \bar{\lambda}_p^*$, where $\bar{\lambda}_p^*$ is an optimal solution of the FCP associated with $\bar{\mathcal{M}}$. Unfortunately, because the optimal solution of an FCP may not be unique, the above convergence may not be true in general. However, it is true for the following special case. Consider

$$\mathcal{M}^n = \left(nq_0, \{n\bar{\lambda}_d(t); t \in [0, T]\}, n\bar{\Lambda}_p, \bar{\theta}_p, \bar{\theta}_d, \{C_0 x; x \in \mathbb{R}_+\}, h_p, h_d, c_p, c_d, c_f\right), \tag{44}$$

where $q_0, \bar{\lambda}_d, \bar{\Lambda}_p, \bar{\theta}_p$, and $\bar{\theta}_d$ are as introduced in Assumption 3 of this section. It is easily seen that Assumptions 2 and 3 hold for $\mathcal{M}^n$. The corresponding limit input is given by

$$\bar{\mathcal{M}} = \left(q_0, \{\bar{\lambda}_d(t); t \in [0, T]\}, \bar{\Lambda}_p, \bar{\theta}_p, \bar{\theta}_d, \{C_0 x; x \in \mathbb{R}_+\}, h_p, h_d, c_p, c_d, c_f\right).$$

Combining Proposition 3.1 and Theorems 4.1 and 4.2 yields the following corollary, whose proof will be omitted. Recall that $\mathcal{V}^n(\mathcal{M}^n)$ and $\bar{\mathcal{V}}^n(\mathcal{M}^n)$ denote the optimal values of the QCP and the associated FCP for the $n^{\text{th}}$ system with $\mathcal{M}^n$, and $\mathcal{V}(\bar{\mathcal{M}})$ denotes the optimal value of the FCP associated with the limit input $\bar{\mathcal{M}}$.

**Corollary 4.2.** *Suppose the $n^{\text{th}}$ system has input $\mathcal{M}^n$ given in (44). The following hold.*

i. *Let $\bar{\lambda}_p^*$ denote an optimal solution of the FCP associated with $\bar{\mathcal{M}}$. Then, $n\bar{\lambda}_p^*$ is an optimal solution of the FCP associated with $\mathcal{M}^n$. Thus, for each $n \in \mathbb{N}$, $\bar{\mathcal{V}}^n(\mathcal{M}^n) = n\mathcal{V}(\bar{\mathcal{M}})$.*

ii. *As $n \to \infty$,*

$$0 \leq \mathcal{V}^n(\mathcal{M}^n) - n\mathcal{V}(\bar{\mathcal{M}}) \leq \mathcal{R}^n\left(n\bar{\lambda}_p^*; \mathcal{M}^n\right) - n\bar{\mathcal{R}}\left(\bar{\lambda}_p^*; \bar{\mathcal{M}}\right) = O(\sqrt{n}).$$

**Remark 4.3.** Theorems 4.1 and 4.2 provide a useful framework to construct asymptotically optimal production-rate functions for systems with large demand rates. Specifically, Theorem 4.1 says that the optimal solution of the FCP is optimal for the QCP with error $O(\sqrt{n})$ under the stronger Assumption 2 on $\lambda_d^n$ and $C^n$ and a weaker asymptotic assumption (25). On the other hand, Theorem 4.2 implies the production rate constructed from the optimal solution of the limit FCP is optimal for the QCP with error $o(n)$, under the weaker assumption for $\lambda_d^n$ and $C^n$ and the stronger asymptotic assumptions (as in Assumption 3).

**Remark 4.4.** For a large-scale nonstationary stochastic model (e.g., our double-ended queueing model), the FCP successfully captures the system's temporal variability (performance trend in time) and ignores the stochastic variability. When the scale is large or medium, the FCP tends to be effective because time variability often dominates the stochastic variability (see Liu and Whitt 2012a for a similar observation).

In Section 5, we illustrate the effectiveness of FCP through numerical examples. In Section 6, we develop numerical solutions to QCP by formulating a DP problem and demonstrate the asymptotic optimality of the FCP solutions. The detailed solution procedure for the FCP can be found in the appendix.

**Proof of Theorem 4.2.** The following lemma and Lemma 4.1 will be used in the proof of Theorem 4.2. In particular, Lemma 4.1 will be used to establish the uniform integrability of $\bar{Q}^n$, and Lemma 4.2 is essentially the fluid approximation under an arbitrary admissible production rate $\lambda_p^n$, in which the process $\tilde{q}^n$ can be interpreted as the fluid limit under $\lambda_p^n$. For proofs of these two lemmas, see Section 7. Given a deterministic $q_0 \in \mathbb{R}$, under an admissible production rate $\lambda_p^n$, define the following stochastic process.

$$\tilde{q}^n(t) = q_0 + \int_0^t \left[\bar{\lambda}_p^n(s) - \bar{\lambda}_d\left(s, \tilde{q}^n(s)\right) - \bar{\theta}_p \tilde{q}^{n,+}(s) + \bar{\theta}_d \tilde{q}^{n,-}(s)\right] ds. \tag{45}$$

**Lemma 4.2.** *Assume that* $\mathbb{E}[|\bar{Q}^n(0) - q_0|^2] \to 0$ *for some deterministic* $q_0 \in \mathbb{R}$. *Then, under Assumption 3, we have*

$$\mathbb{E}\left(\sup_{0 \le s \le T} \left|\bar{Q}^n(s) - \tilde{q}^n(s)\right|\right) \to 0, \text{ as } n \to \infty.$$

Let $n\bar{\lambda}_p^*$ be the production rate for the $n^{\text{th}}$ system. The admissibility of $n\bar{\lambda}_p^*$ to the QCP associated with $\mathcal{M}^n$ can be verified analogously to Lemma 3.1. From Lemma 4.2, under the production rate $n\bar{\lambda}_p^*$, we have

$$\mathbb{E}\left(\sup_{0 \le t \le T} \left|\bar{Q}^n(t) - q^*(t)\right|\right) \to 0, \tag{46}$$

Where, for $t \ge 0$,

$$q^*(t) = q_0 + \int_0^t \left[\bar{\lambda}_p^*(s) - \bar{\lambda}_d\left(s, q^*(s)\right) - \bar{\theta}_p q^{*,+}(s) + \bar{\theta}_d q^{*,-}(s)\right] ds. \tag{47}$$

We proceed to prove Theorem 4.2 using Lemmas 4.1 and 4.2. We first show (42) in Theorem 4.2. Note that from (35), we have

$$\begin{aligned}\bar{\mathcal{R}}^n\left(\bar{\lambda}_p^*; \mathcal{M}^n\right) = \mathbb{E}\left(\int_0^T \left[h_p \bar{Q}^{n,+}(t) + h_d \bar{Q}^{n,-}(t) + \bar{C}^n\left(\bar{\lambda}_p^*(t)\right)\right] dt\right) + c_f V_T\left(\bar{\lambda}_p^*\right) \\ + c_p \mathbb{E}\left[\bar{N}_3^n\left(\theta_p^n \int_0^T \bar{Q}^{n,+}(s)\, ds\right)\right] + c_d \mathbb{E}\left[\bar{N}_4^n\left(\theta_d^n \int_0^T \bar{Q}^{n,-}(s)\, ds\right)\right].\end{aligned}$$

From (46), we have that

$$\begin{aligned}&\mathbb{E}\left(\sup_{0\le t\le T}\left|\bar{Q}^{n,+}(t)-q^{*,+}(t)\right|\right)+\mathbb{E}\left(\sup_{0\le t\le T}\left|\bar{Q}^{n,-}(t)-q^{*,-}(t)\right|\right)\\&\le 2\mathbb{E}\left(\sup_{0\le t\le T}\left|\bar{Q}^{n}(t)-q^{*}(t)\right|\right)\to 0.\end{aligned}\tag{48}$$

We next note that

$$\begin{aligned}&\mathbb{E}\left[\sup_{0\le t\le T}\left|\bar{N}_3^n\left(\theta_p^n\int_0^t\bar{Q}^{n,+}(s)ds\right)-\bar{\theta}_p\int_0^t q^{*,+}(s)ds\right|\right]\\&\quad\le\mathbb{E}\left[\sup_{0\le t\le T}\left|\bar{N}_3^n\left(\theta_p^n\int_0^t\bar{Q}^{n,+}(s)ds\right)-\theta_p^n\int_0^t\bar{Q}^{n,+}(s)ds\right|\right],\qquad(49)\\&\qquad+\left|\theta_p^n-\bar{\theta}_p\right|\int_0^T q^{*,+}(s)ds+\theta_p^n\int_0^T\mathbb{E}\left[\left|\bar{Q}^{n,+}(s)-q^{*,+}(s)\right|\right]ds.\qquad(50)\end{aligned}$$

From (48) and Assumption 3(i), the summation in (51) converges to zero. The expectation in (49) also converges to zero, which follows from the proof of Lemma 4.1 (see (82)). Thus,

$$\mathbb{E}\left[\sup_{0\le t\le T}\left|\bar{N}_3^n\left(\theta_p^n\int_0^t\bar{Q}^{n,+}(s)ds\right)-\bar{\theta}_p\int_0^t q^{*,+}(s)ds\right|\right]\to 0.\tag{51}$$

Using a similar argument, we have

$$\mathbb{E}\left[\sup_{0\le t\le T}\left|\bar{N}_4^n\left(\theta_d^n\int_0^t\bar{Q}^{n,-}(s)ds\right)-\bar{\theta}_d\int_0^t q^{*,-}(s)ds\right|\right]\to 0.\tag{52}$$

Using the convergence in (48), (51), and (52), and Assumption 3(iii),

$$\begin{aligned}\bar{\mathcal{R}}^n\left(\bar{\lambda}_p^*;\mathcal{M}^n\right)&\to\int_0^T\left[h_pq^{*,+}(t)+h_dq^{*,-}(t)+\bar{C}\left(\bar{\lambda}_p^*(t)\right)\right]dt+c_fV_T\left(\bar{\lambda}_p^*\right)\\&\quad+c_p\bar{\theta}_p\int_0^T q^{*,+}(t)dt+c_d\bar{\theta}_d\int_0^T q^{*,-}(t)dt\\&=\int_0^T\left[\left(h_p+c_p\bar{\theta}_p\right)q^{*,+}(t)+\left(h_d+c_d\bar{\theta}_d\right)q^{*,-}(t)+\bar{C}\left(\bar{\lambda}_p^*(t)\right)\right]dt+c_fV_T\left(\bar{\lambda}_p^*\right)\\&=\mathcal{V}(\bar{\mathcal{M}}).\end{aligned}$$

This shows (42). To show (43), let $\{\lambda_p^n\}_{n\ge1}$ be an arbitrary admissible sequence of production rates, and define $\tilde{q}^n$, as in (45). We first note that

$$\liminf_{n\to\infty}\bar{\mathcal{R}}^n\left(\bar{\lambda}_p^n;\mathcal{M}^n\right)\le\limsup_{n\to\infty}\bar{\mathcal{R}}^n\left(\bar{\lambda}_p^n;\mathcal{M}^n\right)<\infty,$$

which follows from the uniform integrability of $\{\bar{Q}^n(t);t\in[0,T]\}$ in Lemma 4.1. Next, from Lemma 4.2, we have

$$\mathbb{E}\left(\sup_{0\le t\le T}\left|\bar{Q}^n(t)-\tilde{q}^n(t)\right|\right)\to 0,\tag{53}$$

and using similar arguments in the proof of (51) and (52), it can be shown that

$$\begin{aligned}\mathbb{E}\left[\left|\bar{N}_2^n\left(\theta_p^n\int_0^T\bar{Q}^{n,+}(t)dt\right)-\bar{\theta}_p\int_0^T\tilde{q}^{n,+}(t)dt\right|\right]\to 0,\\ \mathbb{E}\left[\left|\bar{N}_4^n\left(\theta_d^n\int_0^T\bar{Q}^{n,-}(t)dt\right)-\bar{\theta}_d\int_0^T\tilde{q}^{n,-}(t)dt\right|\right]\to 0.\end{aligned}\tag{54}$$

It follows from (53) and (54) that

$$\left|\bar{\mathcal{R}}^n\left(\bar{\lambda}_p^n;\mathcal{M}^n\right)-\mathbb{E}\left(\bar{\mathcal{R}}\left(\bar{\lambda}_p^n;\bar{\mathcal{M}}\right)\right)\right|\to 0,\tag{55}$$

where

$$\bar{\mathcal{R}}\left(\bar{\lambda}_p^n;\bar{\mathcal{M}}\right)=\int_0^T\left[\left(h_p+c_p\bar{\theta}_p\right)\tilde{q}^{n,+}(t)+\left(h_d+c_d\bar{\theta}_d\right)\tilde{q}^{n,-}(t)+\bar{C}\left(\bar{\lambda}_p^n(t)\right)\right]dt+V_T\left(\bar{\lambda}_p^n\right).$$

Note that $\bar{\mathcal{R}}(\bar{\lambda}_p^n;\bar{\mathcal{M}})$ is a random variable because $\tilde{q}^n$ and $\bar{\lambda}_p^n$ are stochastic. Because $\mathcal{V}(\bar{\mathcal{M}})$ is the optimal value of the FCP, we must have $\bar{\mathcal{R}}(\bar{\lambda}_p^n;\bar{\mathcal{M}})\geq\mathcal{V}(\bar{\mathcal{M}})$ for each $n$ almost surely. Thus, from (55), we have

$$\liminf_{n\to\infty}\bar{\mathcal{R}}^n\left(\bar{\lambda}_p^n;\mathcal{M}^n\right)=\liminf_{n\to\infty}\mathbb{E}\left(\bar{\mathcal{R}}\left(\bar{\lambda}_p^n;\bar{\mathcal{M}}\right)\right)\geq\mathcal{V}(\bar{\mathcal{M}}).$$

## 5. Numerical Examples

In this section, we provide numerical examples to evaluate the effectiveness of the fluid approximation as a performance lower bound and illustrate asymptotic optimality of the FCP solutions as the system scale grows. We also demonstrate the importance of including the flexibility cost in the proposed control problem. As is discussed in Remark 3.1, we solve a linear reformulation of the FCP in discrete time, and the detailed reformulation steps are reported in the appendix.

### 5.1. Asymptotic Effectiveness of FCP

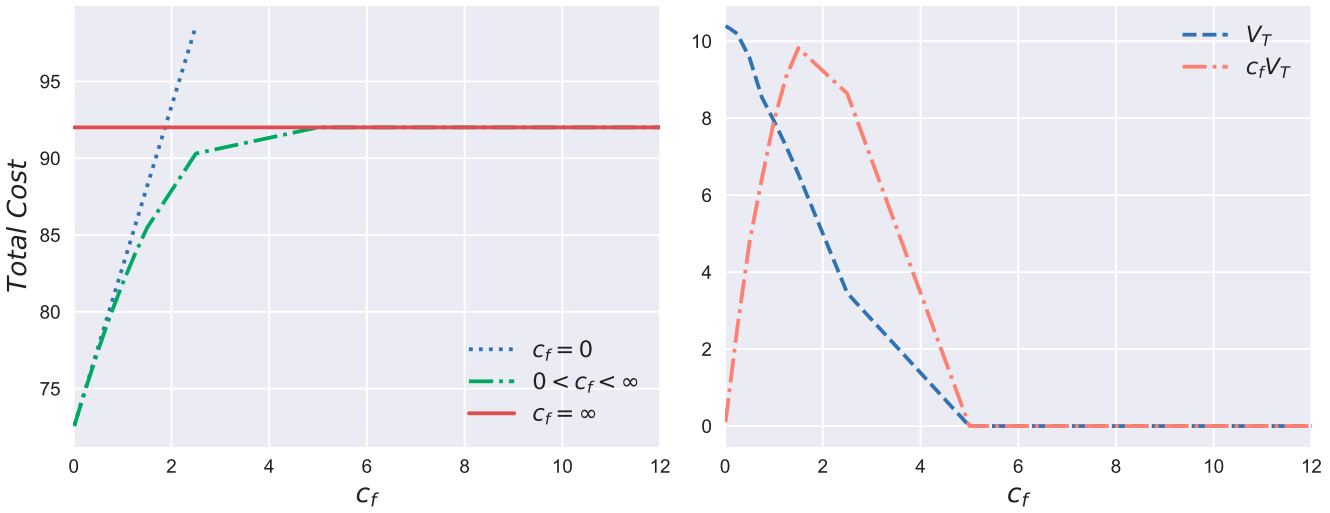
In the first example, let $T=24$, $\bar{\lambda}_d(t)=1+0.6\sin(t)$, $\bar{\theta}_p=0.5$, and $\bar{\theta}_d=2$. Without loss of generality, initial conditions are all set to zero. We consider $n=5,10,50,100$, and Monte Carlo simulations are computed by 20,000 independent iterations when $n=5,10$, and 2,000 independent iterations when $n=50,100$. In the scale $n$ simulation, we have $\lambda_d^n(t)=n\bar{\lambda}_d(t)$, $\theta_p^n=\bar{\theta}_p$, and $\theta_d^n=\bar{\theta}_d$. The production cost is linear with unit cost $C_0=2$, and other cost coefficients are $c_d=5$, $c_p=3$, $h_d=5$, $h_d=3$, and $c_f=1$. We note that the example input satisfies (44) and Corollary 4.2 holds.

As an example to extend the LP reformulation, we also consider a requirement that changes made to production rate can only occur every $L$ amount of time. Here, we let $L=4$, which models a work shift in manufacturing settings. At the beginning of each work shift, the production rate is updated and kept unchanged throughout the entire shift. We note that $L$ and $\Delta t$ should be distinguished from each other. The time-discretization step $\Delta t$ is only introduced in formulating the discrete-time LP. It is void of any physical interpretation and only exists to control the accuracy of the LP reformulation to approximate continuous-time FCP. Detailed formulation and another example with the extension of realistic constraints are presented in Section A.2 in the appendix.

The optimal production rate $\bar{\lambda}_p^*(t)$ of the FCP associated with $\bar{\mathcal{M}}$ is the dashed line in the top panel of Figure 2, where the solid line is the demand rate $\bar{\lambda}_d(t)$. We denote the optimal inventory and back-order queue length associated with $\bar{\lambda}_p^*(t)$ and $\bar{\mathcal{M}}$ by $q^{*,+}(t)$ and $q^{*,-}(t)$, where $q^*(t)$ is as defined in (47). For the scale $n$ system, we implement production rate $n\bar{\lambda}_p^*$ and set parameters to be of $\mathcal{M}^n$. In Figure 2, we plot $\mathbb{E}[\bar{Q}^{n,+}(t)]$ and $\mathbb{E}[\bar{Q}^{n,-}(t)]$ (solid lines in the middle and bottom panels), and $q^{*,+}(t)$ and $q^{*,-}(t)$ (dotted lines in the last two panels). In Table 1, we report the percentage difference, which is defined as the difference between the simulation and FCP results divided by the FCP result for itemized costs and the total cost.

As system scale $n$ increases, inventory, back order, and total cost from simulations can be better approximated by their fluid counterparts. See the middle and bottom panels of Figure 2 and Table 1 for the significant improvement as the system scale increases. Note that the on-site inventory level (positive part of the queue length) and the back-order size (negative part of the queue length) are proportional to the total

**Figure 3.** Total Cost of Three Cases with Different Production Flexibility




inventory holding cost and the total back-order cost, respectively. Therefore, we omit the percentage difference for the queue length. Also, the flexibility cost $c_f V_T(n\bar{\lambda}_p^*)$ is fixed and, hence, is also omitted here.

### 5.2. Impact of the Flexibility Cost

Now, we discuss the impact of production flexibility cost. In Figure 3, the total costs of three cases with different levels of production flexibility are plotted. Without loss of generality, we consider zero initial condition. The first one considers the case of full flexibility (i.e., flexibility cost $c_f = 0$), under which the optimal fluid production rate should track the demand rate—that is, $\bar{\lambda}_p^*(t) = \bar{\lambda}_d(t)$. Secondly, we consider the FCP, which balances the cost associated with queue length and limited flexibility (i.e., flexibility cost $c_f \in (0, \infty)$). In the third case, the production rate is constrained to be constant (i.e., no flexibility, $c_f = \infty$). The total costs for the first and second cases are directly calculated from the fluid model without optimization.

Notice for the free- and constant-production cases, the total variation of their chosen production rate is independent of the flexibility cost $c_f$. For the free-production case, the total variation is the fixed value of $V_T(\lambda_d^n)$, which results in a linear total cost as $c_f$ increases. For the constant-production case, the total variation is zero. For the FCP, the total cost varies in a nonlinear fashion as $c_f$ increases. In the left panel of Figure 3, when the flexibility is extremely small or large, the controlled case degenerates to the free or constant case. However, when $c_f$ takes a value that allows flexibility cost to be comparable with other cost items, FCP achieves the lowest total cost. In the right panel of Figure 3, we illustrate the monotonicity of total variation as a function of $c_f$, which is proved in Proposition 3.2.

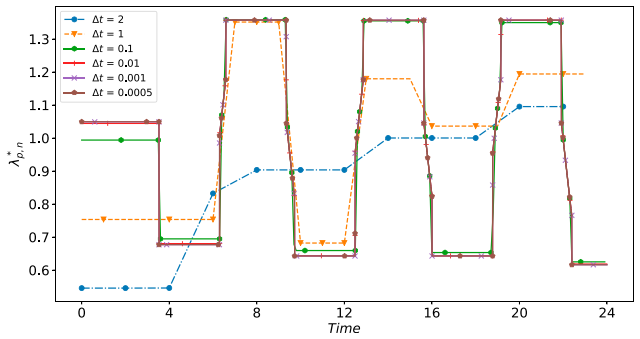
### 5.3. Convergence of Discrete-Time LP Solutions

To obtain numerical solution of the FCP, we resort to the discrete-time LP reformulation. Detailed steps in constructing the LP can be found in the appendix. The key factor that greatly affects the shape of the output solution is the discrete-time step size $\Delta t$. Although this paper does not pursue a theoretical proof that the solution of the discrete-time LP reformulation converges to the true FCP optimal solution, we do provide an

**Table 1.** Relative Difference Between FCP and Simulation Results under $n\bar{\lambda}_p^*$

| | System scale, % | | | | | |
|---|---|---|---|---|---|---|
| Cost breakdown | $n = 5$ | $n = 10$ | $n = 25$ | $n = 50$ | $n = 100$ | $n = 1{,}000$ |
| Inventory holding | 62.8 | 49.2 | 30.9 | 18.5 | 11.7 | 1.7 |
| Inventory expiration | 62.7 | 49.1 | 30.7 | 18.4 | 11.8 | 1.6 |
| Back-order holding | 35.1 | 23.6 | 12.0 | 6.6 | 3.3 | 0.4 |
| Lost sale | 35.1 | 23.6 | 12.0 | 6.6 | 3.3 | 0.4 |
| Production | 0.05 | 0.06 | 0.02 | 0.01 | 0.09 | 0.07 |
| Total cost without production | 80.9 | 46.45 | 20.1 | 10.7 | 5.8 | 0.65 |
| Total cost | 29.2 | 19.1 | 9.6 | 5.2 | 2.9 | 0.4 |

**Figure 4.** Optimal LP Solution $\lambda^*_{p,n}$ with $\Delta t$ Decreases from 2 to 0.0005




example that presents optimal LP solutions under a sequence of decreasing $\Delta t$. We keep refining the discrete time grid until the change in the LP optimal solution is negligible. In Figure 4, we consider a series of numerical examples with the roughest case of $\Delta t = 2$ and the finest of $\Delta t = 0.0005$. Optimal LP solution converges quickly after the case $\Delta t = 0.01$. Note the settings of these examples follow Section 5.1, except $L = 3$. This provides convincing foundations for the LP reformulation of the FCP.

## 6. Exact QCP Solution Using Dynamic Programming

To supplement the discussion on the *asymptotic optimality* of FCP, we now solve QCP via dynamic programming. We write a discrete-time and discrete-space DP problem based on the dynamics and cost structure of the (continuous-time) QCP model. The numerical solution to our DP problem should be close to that of QCP when the discrete time step is small. In Section 3 and Section 4, we show that the FCP serves as a lower bound to QCP regardless of system scale $n$ and produces an asymptotically optimal solution. This indicates that as $n$ grows large, the optimal numerical solution independently obtained from the DP and FCP should eventually coincide. Also, for any given $n$, the optimal cost of FCP should be a lower bound for the DP optimal cost. Although solutions to the DP problem are successfully obtained for academic-sized scale (i.e., small to moderate size) problems, we remark that the computational burden of DP significantly exceeds that of FCP, especially when $n$ is large. In this section, we properly develop the DP problem and present several numerical examples to compare various outputs from the DP and FCP numerical solutions.

### 6.1. Dynamic Programming Formulation

We define a DP problem with a two-dimensional system state, queue length $\hat{Q}$, and production rate $\lambda_p$. Note that the production rate is a decision variable as well as a system-state variable. Specifically, the immediate past decision of production rate (at the previous time step) becomes a system state, which is used to calculate the discrete version of the total variation.

Let $J_k(\hat{Q}, \lambda_p)$ be the optimal value function, given that the system has a queue length of $\hat{Q}$ and production rate is set to $\lambda_p$, with $k$ steps to go. The state variable $\hat{Q}$ takes integer values, and variable $\lambda_p$ takes values in a discrete set $\mathcal{D} := \{\lambda_{p,n} : \lambda_{p,n} = n\hat{\lambda}, n = 0, \ldots, N\}$, $N = \lfloor \Lambda_p/\hat{\lambda} \rfloor$ and $\hat{\lambda}$ is taken to be a fairly small value such that $\mathcal{D}$ can reasonably approximate $[0, \Lambda_p]$. One can regard $\hat{\lambda}$ as an accuracy level of the controller of a production machine. The discrete time space is $\mathcal{K} = \{k\delta : k = 0, \ldots, K\}$, where $K = T/\delta$. The value of $\delta$ should be very small. The initial condition is given by $J_0(\hat{Q}, \lambda_p) = 0$ for all $\hat{Q}$ and $\lambda_p$, which corresponds to zero terminal costs in the original QCP. For a given state $(\hat{Q}, \lambda_p)$ with $k+1$ stages to go, possible state transitions in the next time epoch include $(\hat{Q}+1, \lambda'_p)$, $(\hat{Q}-1, \lambda'_p)$ and $(\hat{Q}, \lambda'_p)$, with probabilities $p_1 = (\lambda_p + \theta_d \hat{Q}^-)\delta$, $p_2 = (\lambda_{d,\cdot} + \theta_p \hat{Q}^+)\delta$, and $1 - p_1 - p_2$ respectively. The discrete-time DP problem is formulated as the following:

$$J_{k+1}(\hat{Q}, \lambda_p) = \min_{\lambda'_p \in \mathcal{D}} \Big\{ \big|\lambda_p - \lambda'_p\big| c_f + \Big( (h_p + \theta_p c_p)\hat{Q}^+ + (h_d + \theta_d c_d)\hat{Q}^- + C(\lambda_p) \Big)\delta \,, \tag{56}$$

$$+ (\lambda_p + \theta_d \hat{Q}^-)\delta J_k\Big(\hat{Q}+1, \lambda'_p\Big) + \Big(\lambda_{d,k+1} + \theta_p \hat{Q}^+\Big)\delta J_k\Big(\hat{Q}-1, \lambda'_p\Big), \tag{57}$$

$$+ \Big(1 - \Big(\lambda_p + \theta_d \hat{Q}^- + \lambda_{d,k+1} + \theta_p \hat{Q}^+\Big)\delta\Big) J_k\Big(\hat{Q}, \lambda'_p\Big) \Big\}, \; k > 0, \tag{58}$$

$$J_0(\hat{Q}, \lambda_p) = 0, \; \hat{Q} \in \mathbb{Z}, \; \lambda_p \in \mathcal{D}, \; k \in \mathcal{K}.$$

in (56) characterize costs incurred in the $(k+1)^{\text{th}}$ stages. The first term is the flexibility cost. Notice that $\lambda'_p$ is the optimal production to be determined with $k$ periods to go and $\lambda_p$ is given as the state of production rate with $k+1$ periods to go. Holding costs in a small time interval $\delta$ on the production and back-order side are $h_p Q^+ \delta$ and $h_d Q^- \delta$, respectively. Costs resulting from expired products and back orders are $c_p \theta_p Q^+ \delta$ and $c_d \theta_d Q^- \delta$. Finally, the production cost is $C(\lambda_p)\delta$. For any given state $(Q, \lambda_p)$ in the $(k+1)^{\text{th}}$ period and policy of $\lambda'_p$ in the $k^{\text{th}}$ period, the queue size in the $k^{\text{th}}$ period can only be $\hat{Q}+1$, $\hat{Q}-1$ or $\hat{Q}$, as indicated by (57) and (58).

**6.1.1. Implementation of DP Recursion (56)–(58).** Based on this recursive equation and the initial condition, we can obtain a numerical solution to the QCP. The optimal policy from DP is stored as a four-dimensional matrix, which corresponds to queue length, production rate, discrete time, and the optimal production rate, given a set of previous three variables—that is, $(\hat{Q}, \lambda_p, k+1, \lambda'_p)$. To implement this DP optimal policy, one first observes the current time and system state. The corresponding optimal control can then be looked up. The value of optimal DP value function $J_k(\hat{Q}, \lambda)$ is also stored in a matrix with coordinates of $(\hat{Q}, \lambda_p, k, J_k(\hat{Q}, \lambda))$. However, the vector of optimal value function that matters is when $\hat{Q} = Q_0$, $k = K$. The average value of this vector of $J_k(\hat{Q}, \lambda)$ is the mean total cost $\mathcal{R}$ in (3), which is compared with the FCP numerical solutions.

**Remark 6.1.** (Implementation of DP Recursion)**.** The main obstacle in the implementation of the DP program above is the successful storing and indexing high-dimensional matrix. Moreover, in QCP and DP, there is no upper bound imposed on the queue-length state variable. However, in any computer program, there are only finite possible values that any variable can take. It requires that we prespecify an upper bound. Therefore, the challenge at hand is to reduce the size of state space while keeping all feasible paths intact. If the bound is too low, then the state space is overly truncated, and the result is only suboptimal due to reduced state space. If it is too high, then the search space is unnecessarily enlarged, and the computational burden is increased as a consequence. To overcome this obstacle, we utilize FCP results to provide some guidance, because from a moderate to large scale, FCP solutions should provide a good approximation to QCP. We first solve the FCP and find the largest unscaled queue length, according to which we set the upper bound. We then proceed to solve the DP problem and conduct simulations of the DP problem. We are most confident in the results when the simulated DP queue length does not hit the boundary. When the boundaries are hit, we increase the bound and resolve the DP.

Similarly, the choices of $\hat{\lambda}$ and $\delta$ face the same problem; when they are too big, it reduces the accuracy of the DP numerical solution, and when too small, it increases the computational cost. Because we develop the DP problem as a benchmark for the FCP, in order to ensure the accuracy of DP results, we have to sacrifice computational efficiency and choose small values of $\hat{\lambda}$ and $\delta$ in our numerical studies.

**6.1.2. Simulation Verification of DP Solutions.** The optimal policy obtained from DP, unlike the *deterministic* numerical optimal solution of the FCP, is a *state-dependent* roadmap for decision makers. This fact also affects the data structure for storing the computed optimal value. Given the same problem inputs, it takes a high-dimensional matrix to store the DP output. In contrast, the FCP results are stored as vectors, which exhibit only time dependency, but no state dependency (i.e., queue length and previous production rate). The DP optimal policy is an extensive plan that includes the optimal path for every possible scenario to the extent of the entire state space. To validate the effectiveness of the DP numerical solution, we conduct simulation experiments; we also compare the outputs of simulated DP to that of the FCP to illustrate the asymptotic optimality results.

Our simulation experiment adopts a uniformization approach. That is, all possible events, such as the arrival of a production unit or a demand unit, expiration of a piece of on-site inventory, or back orders, occur at some exponential rates. Given the system state $(\hat{Q}, \lambda_p)$ with $k+1$ periods to go, the aggregated rate of the exponential clock of the "next event" (of any kind) is

$$\tilde{\lambda} = \lambda_p + \theta_d \hat{Q}^- + \lambda_{d,k+1} + \theta_p \hat{Q}^+.$$

In a small time step $\delta$, the probability that one event occurs is approximately $\tilde{\lambda}\delta$. In addition, if one event happens, the probabilities of demand arrival, production arrival, back-order expiration, and on-site inventory wastage are

$$\lambda_{d,k+1}/\tilde{\lambda}, \quad \lambda_p/\tilde{\lambda}, \quad \theta_d \hat{Q}^-/\tilde{\lambda} \quad \text{and} \quad \theta_p \hat{Q}^+/\tilde{\lambda},$$

respectively. Starting from the same initial condition, we simulate and record a large number of independent sample paths under the optimal DP policy. Finally, to compare with FCP results, we calculate the mean value of all sample paths. We report results in the next section.

### 6.2. Comparing DP and FCP Solutions

We compare the optimal numerical solutions of QCP and FCP to demonstrate the *asymptotic optimality* of the FCP solution. As we stated previously, the optimal policy of DP is a state-dependent roadmap that records optimal paths for all possible scenarios to the extent of the state space, and the numerical solution to FCP is state-independent. One cannot directly compare these two types of results. We bridge the gap by conducting a simulation of DP under the optimal DP policy and obtain the average of realized optimal control and queue-length paths in DP.

To control the problem size, we set DP parameters as the following: $T = 4$, $\delta = 0.0001$, $\lambda^n_{d,k} = n(1 + 0.6\sin(k\delta))$ for $k = 0, \ldots, T/\delta$, $\hat{\lambda} = 0.001\Lambda$, $C = 2$, $c_f = 6$, $h_p = 3$, $c_p = 3$, $\theta_p = 1$, $h_d = 5$, $c_d = 5$, $\theta_d = 1$, $\Lambda^n_p = 2n$, where $n$ is the system scale taking integer values in $[1, 1000]$. For the given scale tuple $(1, 3, 5, 25, 200, 1000)$, we repeated the simulation for $(10^4, 10^4, 10^4, 10^3, 10^2, 50)$ times. The setup for LP is outlined in Section 5. For simplicity, we also ignore the constraint to fix $\lambda^n_p$ within each interval of length $L$.

In Figure 5, we compare the optimal total costs of DP numerical solutions, simulations of DP and FCP for a series of system scales ranging from $n = 1$ to $n = 100$. In the left panel of Figure 5, numerical solutions to DP overlap with the simulations of DP, regardless of the system scale, which verifies the accuracy of DP results. Unlike DP, the solution of FCP is independent of the system scale, which is shown as the constant blue line in the left panel. As expected, the FCP solution approaches DP results as the system scale grows large and eventually matches with the DP. In the right panel of Figure 5, we show the decrease in the percentage error of the numerical solution of FCP results, with respect to simulation results of DP. This example illustrates the asymptotic optimality and lower-boundedness of the FCP optimal solution, as stated in Theorem 4.2 and Theorem 3.2.

In addition to Figure 5, we present one more plot on queue length and optimal production rate in Figures 6. In the top left panel of Figure 6, we plot the mean optimal queue length from the simulated DP performance. It is evident that as the system scale increases from $n = 1$ to $n = 100$, the mean queue length converges quickly to the optimal FCP queue length $q$. Similarly, in the top right panel of Figure 6, we plot the optimal production rate. Again, we observe the convergence of the mean path of the optimal production rate to the optimal FCP production rate. We also plot the scaled on-site inventory and back-order queue length and their corresponding FCP counterparts in the bottom panels of Figure 6.

## 7. Additional Proofs

### 7.1. Proof of Lemma 3.2

To solve the control problem in Lemma 3.2, we remark that it is nonsmooth, as it involves the positive and negative parts of system state, which calls on an extension of the general maximum principle. See Feichtinger and Hartl (1985). Furthermore, the structure of the mean queue-length process has certain patterns, which serves as the backbone of this proof. Starting from the initial condition, the $q(t)$ either goes to zero before terminal time $T$ or stays in the same quadrant for $[0, T]$. If it is the first case, $q(t)$ may or may not remain at zero after first reaching zero. Then, $q(t)$ may stay at zero until terminal time $T$ or at some point increase (decrease) to be positive (negative). The system may terminate before $q(t)$ returns to zero. If not, then system will again

**Figure 5.** An Illustration of the Asymptotic Optimality of FCP Solution

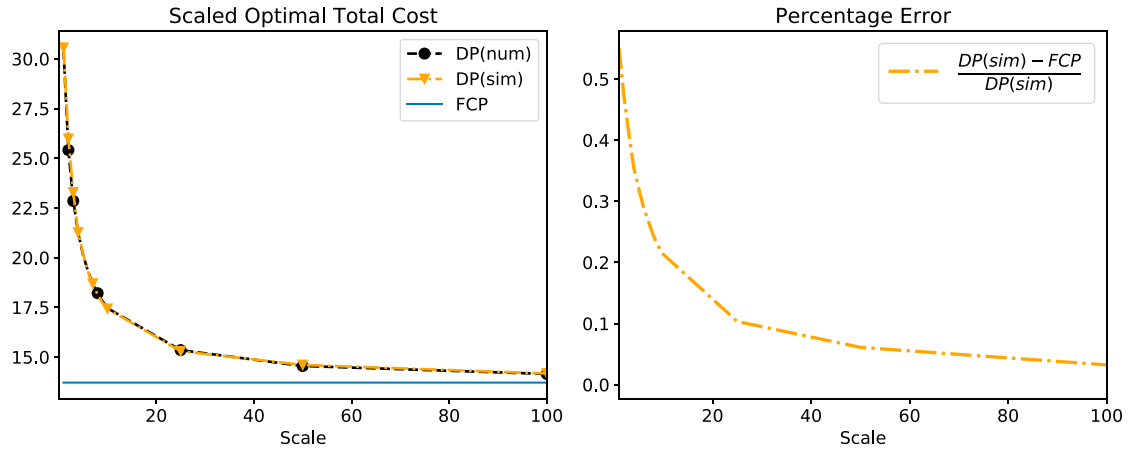
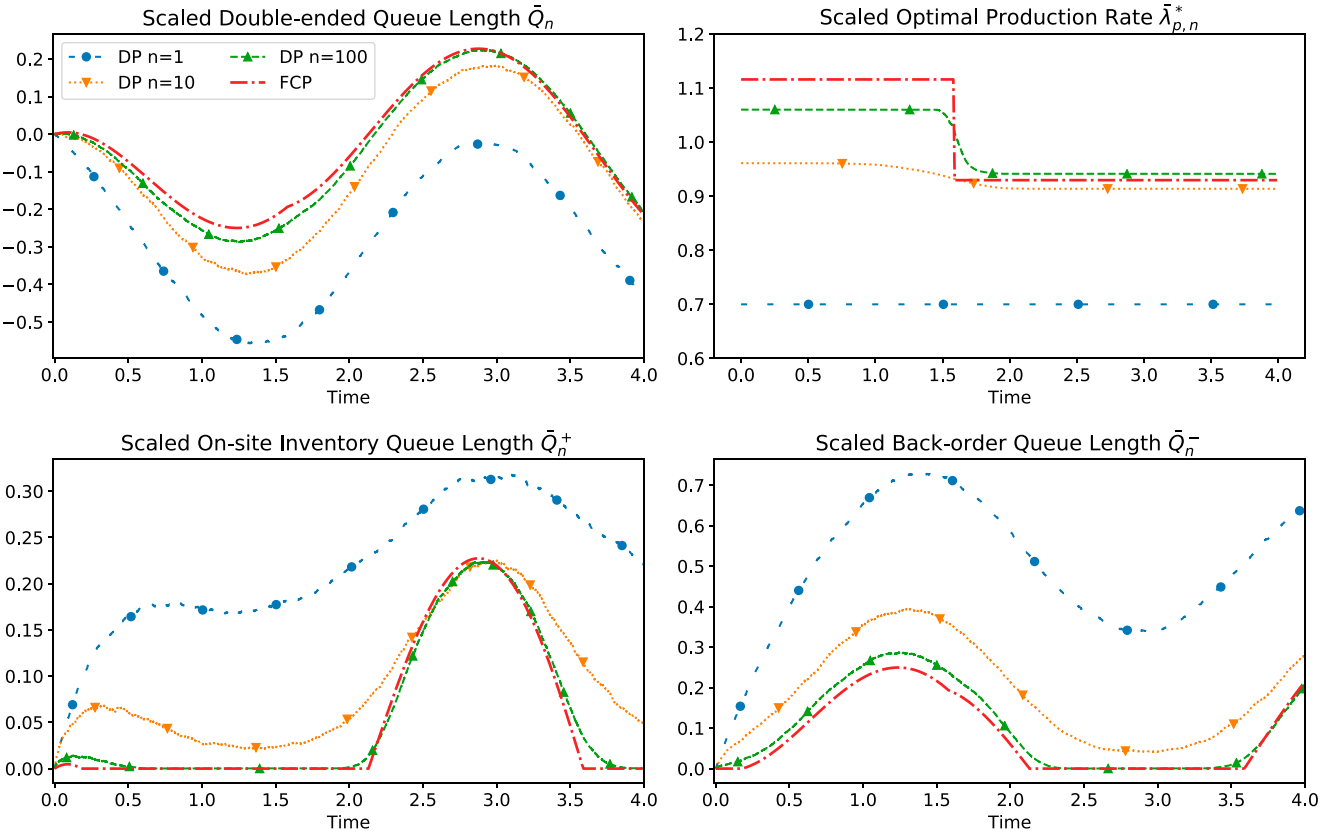


**Figure 6.** Scaled Queue Length $\bar{Q}_n$, $\bar{Q}_n^+$, $\bar{Q}_n^-$, and Optimal Production Rate $\bar{\lambda}_p^*$, with $\Delta t = 0.001$




reach back to zero and repeatedly move away from and then back to zero. The main ideas of this proof are to show the optimality of $u^*(t) = 0$ for all possible paths.

Following Pontryagin's Maximum Principle (see Seierstad and Sydsaeter 1986), we first rewrite (21)–(23) as a maximization problem. Let $c_1 = h_p + c_p\theta_p$, $c_2 = h_d + c_d\theta_d$, and $d(t) = \mathbb{E}[\lambda_p(t)] - \lambda_d(t)$. The Hamiltonian is

$$\begin{aligned}\mathcal{H}\big(q(t), u(t), t\big) = &-c_1 q^+(t) - c_2 q^-(t) - (c_1 + c_2)u(t) \\ &+ \lambda(t)\big(d(t) - \theta_p q^+(t) + \theta_d q^-(t) - \big(\theta_p - \theta_d\big)u(t)\big),\end{aligned} \tag{59}$$

where $\lambda(t)$ is the costate variable. Notice that this Hamiltonian is nonsmooth only at discrete time points. In order to define the costate, we divide the time interval $[0, T]$ into four disjoint subsets $A_T$, $B_T$, $C_T$, and $D_T$—that is, $A_T \subset [0, T]$, $B_T \subset [0, T]$, $C_T \subset [0, T]$, $D_T \subset [0, T]$, $A_T \cup B_T \cup C_T \cup D_T = [0, T]$, and $A_T \cap B_T \cap C_T \cap D_T = \emptyset$—such that the optimal queue length in problem (21)–(23) satisfies $q^*(t) > 0$ for $t \in A_T$, $q^*(t) < 0$ for $t \in B_T$, and $q^*(t) = 0$ for $t \in C_T \cup D_T$. Noting that the set $C_T$ consists of finite number of disjoint intervals, we write $C_T = \cup_{k=1}^m (t^{(k)}, s^{(k)})$ such that $0 \le t^{(1)} \le s^{(1)} < t^{(2)} \le s^{(2)} < \cdots < t^{(m)} \le s^{(m)} \le T$, where $m$ is the number of disjoint intervals with zero optimal queue length. In the special case when $q^*(t)$ just passes through zero, $t^{(i)} = s^{(i)}$. The set $D_T = \{t^{(1)}, s^{(1)}, \cdots, t^{(m)}, s^{(m)}\}$ consists only the end points of these open intervals, which are all the nondifferentiable point of $\mathcal{H}$ in $q$.

If $(q^*(t), u^*(t))$ is the optimal control for problem (21)–(23), following theorem 2.1 of Feichtinger and Hartl (1985),

$$\frac{d\lambda(t)}{dt} = -\frac{d\mathcal{H}\big(q(t), u(t), t\big)}{dq(t)} = \begin{cases} c_1 + \theta_p\lambda(t), & \text{if } t \in A_T, \\ -c_2 + \theta_d\lambda(t), & \text{if } t \in B_T, \\ 0, & \text{if } t \in C_T, \\ \partial_q \mathcal{H}\big(q, u, t\big), & \text{if } t \in D_T, \end{cases} \tag{60}$$

$$\lambda(T) = 0,$$

where $\partial_q$ denotes the generalized gradient with respect to $q$, and

$$\mathcal{H}(q^*(t), u^*(t), t) = \max_u \mathcal{H}(q(t), u(t), t). \tag{61}$$

Notice that $\mathcal{H}$ is linear in $u$; therefore, the optimizer of (61) is

$$u^*(t) = \begin{cases} 0, & \text{if } -(c_1 + c_2) - (\theta_p - \theta_d)\lambda(t) < 0, \\ \Delta_0, & \text{if } -(c_1 + c_2) - (\theta_p - \theta_d)\lambda(t) > 0, \\ \delta \in [0, \Delta_0], & \text{if } -(c_1 + c_2) - (\theta_p - \theta_d)\lambda(t) = 0. \end{cases} \tag{62}$$

To show that the optimality of $u^*(t) = 0$ for $t \in [0, T]$, it is essential to show that for any given problem parameter $\mathcal{M}$ and admissible control $\lambda_p$, the auxiliary function $\beta(t) = -(c_1 + c_2) - (\theta_p - \theta_d)\lambda(t) < 0$, which, in essence, requires to solve for costate variable $\lambda(t)$.

We start from the terminal time $T$. Divide the time horizon by time points where the optimal queue length $q^*(t)$ enters/leaves zero. Roughly speaking, there are four parts. The first period, which starts from the terminal time $T$, is the last period where $q^*(t)$ is nonzero. The second period connects with the first period, and it is the last period where $q^*(t)$ remains at zero. The third period follows the second period, and it is the second-to-last period where $q^*(t)$ is nonzero. The last period is the rest of the total horizon, and the analysis here repeats the second and third periods.

**7.1.1. The First Period.** Let $s^{(m)} = \sup\{t \in [0, T]; q^*(t) = 0\}$. If $s^{(m)} < T$, then it is either $(s^{(m)}, T] \in A_T$ or $(s^{(m)}, T] \in B_T$. Hence, from (60), solving $\lambda(t)$ in $t \in (s^{(m)}, T]$, we obtain

$$\lambda(t) = \frac{c_1}{\theta_p}\left(e^{\theta_p(t-T)} - 1\right), \qquad \text{if } \left(s^{(m)}, T\right] \subset A_T, \tag{63}$$

$$\lambda(t) = \frac{c_2}{\theta_d}\left(1 - e^{\theta_d(t-T)}\right), \qquad \text{if } \left(s^{(m)}, T\right] \subset B_T. \tag{64}$$

Checking the sign of the auxiliary function $\beta(t)$ in $t \in (s^{(m)}, T]$ under (63) and (64), respectively,

$$\beta(t) = -c_2 - c_1\left(1 - \left(1 - e^{\theta_p(t-T)}\right)\left(1 - \frac{\theta_d}{\theta_p}\right)\right) < 0, \ \left(s^{(m)}, T\right] \subset A_T, \tag{65}$$

$$\beta(t) = -c_1 - c_2\left(1 - \left(1 - e^{\theta_d(t-T)}\right)\left(1 - \frac{\theta_p}{\theta_d}\right)\right) < 0, \ \left(s^{(m)}, T\right] \subset B_T. \tag{66}$$

If $s^{(m)}$ does not exist, we conclude that $q(t)$ does not change sign throughout the entire decision horizon. Therefore, (65) and (66) hold on the entire $[0, T]$. Hence, the optimal control $u^*(t) = 0$ for $t \in [0, T]$. This trivial case occurs when the initial condition $q_0$ is extremely large or small.

**7.1.2. The Second Period.** The case where $s^{(m)} = T$ indicates a trivial first period. Define $t^{(m)} = \sup\{t \in [0, T]; q^*(t) \neq 0\}$. The interval $\left(t^{(m)}, T\right] \subset C_T$ and from (60), $\lambda(t) = 0$ and $\beta(t) = -(c_1 + c_2)$ for $t \in (t^{(m)}, T]$. For the interval $\left[0, t^{(m)}\right)$, the analysis returns to cases in (63) and (64). Such $t^{(m)}$ exists either when the initial queue length is nonzero, or there exists a time interval where $\lambda_p(t) \neq \lambda_d(t)$. If such $t^{(m)}$ does not exist, then we must have zero initial queue length and $\lambda_p(t) = \lambda_d(t)$ for $t \in [0, T]$, which indicates that the expected queue length $\mathbb{E}(Q(t))$ in (7) and the fluid queue length $q(t)$ in (9) are both zero in $[0, T]$. Theorem 3.2 is trivial under this case.

Assuming the existence of nontrivial $s^{(m)} < T$, let

$$t^{(m)} = \sup\left\{t \in \left[0, s^{(m)}\right];\ q^*(t) \neq 0\right\}. \tag{67}$$

The second period is $(t^{(m)}, s^{(m)})$ and $(t^{(m)}, s^{(m)}) \subset C_T$. To determine the value of $\beta(t)$ in this time interval, we first need to treat the nonsmooth point at $t = s^{(m)}$, where the state process $q(t)$ becomes zero from nonzero value. Because of the continuity of the costate process, at $t = s^{(m)}$, depending on the sign of $q^*(t)$ in the first period, we have

$$\lambda\left(s^{(m)}\right) = \frac{c_1}{\theta_p}\left(e^{\theta_p\left(s^{(m)}-T\right)} - 1\right), \text{ or } \lambda\left(s^{(m)}\right) = \frac{c_2}{\theta_d}\left(1 - e^{\theta_d\left(s^{(m)}-T\right)}\right),$$

and we can check that $\beta(s^{(m)}) < 0$ for both cases from (65) and (66). Following (60), the costate is constant on the interval $[t^{(m)}, s^{(m)}]$—that is,

$$\lambda(t) = \lambda\left(s^{(m)}\right),\ t \in \left[t^{(m)}, s^{(m)}\right]. \tag{68}$$

It is straightforward that $\beta(t) < 0$ for $t \in [t^{(m)}, s^{(m)}]$. If $t^{(m)} = s^{(m)}$, the second period is trivial, and the above analysis still holds. A trivial second period means that at $t = t^{(m)}$, the process $q^*(t)$ either touches zero and goes back to the same sign it had, or it crosses zero and changes sign. When $t^{(m)}$ does not exist, the time horizon $[0, T]$ can be divided into one first period and one second period. More specifically, we must have $q^*(t) = 0$ for $t \in [0, s^{(m)}]$ and $q^*(t) \neq 0$ for $t \in (s^{(m)}, T]$. The auxiliary function $\beta(t) < 0$ for $t \in [0, T]$ follows the analysis in (65), (66), and (68), and, hence, $u^*(t) = 0$ for $t \in [0, T]$.

**7.1.3. The Third Period.** Assuming the existence of nontrivial first and second periods—that is, $0 < t^{(m)} < s^{(m)} < T$—we define

$$s^{(m-1)} = \sup\left\{t \in \left[0, t^{(m)}\right]; q^*(t) = 0\right\}. \tag{69}$$

The third period is $(s^{(m-1)}, t^{(m)})$, and it is either $(s^{(m-1)}, t^{(m)}) \subset A_T$ or $(s^{(m-1)}, t^{(m)}) \subset B_T$. Similarly, to solve the costate process $\lambda(\cdot)$, it follows (60), and the terminal value is now $\lambda(t^{(m)})$ as in (68) due to the continuity of $\lambda(t)$. Depending on the sign of $q^*(t)$ in $(s^{(m)}, T]$ and $(s^{(m-1)}, t^{(m)})$, the solution of $\lambda(t)$ can be written as

$$\lambda(t) = \frac{a}{b}\left(e^{b\left(t - t^{(m)}\right)} - 1\right) + \lambda\left(s^{(m)}\right)e^{b\left(t - t^{(m)}\right)}, \tag{70}$$

where $a = c_1$, $b = \theta_p$ or $a = -c_2$, $b = \theta_d$. There are four possible cases, each corresponding to different representation of $\beta(t)$, and we can show, regardless of different cases, that $\beta(t) < 0$—for example, if $q^*(t) > 0$ on both intervals,

$$\beta(t) = -c_2 - c_1\left(1 - \left(1 - \frac{\theta_d}{\theta_p}\right)\left(1 - e^{\theta_p\left(t - t^{(m)}\right)}\right) - \left(1 - \frac{\theta_d}{\theta_p}\right)\left(1 - e^{\theta_p\left(s^m - T\right)}\right)e^{\theta_p\left(t - t^{(m)}\right)}\right).$$

If $q^*(t) > 0$ on $(s^{(m)}, T]$ and $q^*(t) < 0$ on $(s^{(m-1)}, t^{(m)})$,

$$\beta(t) = -c_2\left(1 - \left(1 - \frac{\theta_p}{\theta_d}\right)\left(1 - e^{\theta_d\left(t - t^{(m)}\right)}\right)\right) - c_1\left(1 - \left(1 - \frac{\theta_d}{\theta_p}\right)\left(1 - e^{\theta_p\left(s^m - T\right)}\right)e^{\theta_p\left(t - t^{(m)}\right)}\right).$$

If $s^{(m-1)}$ does not exist, then we must have $q^*(t) \neq 0$ for $t \in \left[0, t^{(m))}\right)$, $q^*(t) = 0$ for $t \in [t^{(m)}, s^{(m)}]$ and $q^*(t) \neq 0$ for $t \in (s^{(m)}, T]$. For each period, we have shown that $\beta(t) < 0$ and, hence, $u^*(t) = 0$. If $s^{(m-1)}$ does exist, then we use (67) to find $t^{(m-1)}$ and follow the remaining procedures in the second period to show $u^*(t) = 0$. The termination of the analysis happens when one finds the first $s^k$ or $t^l$ does not exist.

### 7.2. Proofs of Lemmas 4.1 and 4.2

Before providing the proofs of Lemmas 4.1 and 4.2, we introduce a process $\{Y^n(t); t \geq 0\}$, which serves as an upper bound for the fluid scaled queue-length process $\{\bar{Q}^n(t); t \geq 0\}$. Let $\{\lambda_p^n(t); t \in [0, T]\}$ be an admissible production rate. We note that, from (1) and (2), for $t \geq 0$,

$$\begin{aligned}\bar{Q}^n(t) = \bar{Q}^n(0) + \bar{N}_1^n\left(\int_0^t \bar{\lambda}_p^n(s)ds\right) - \bar{N}_2^n\left(\int_0^t \bar{\lambda}_d^n\left(s, \bar{Q}^n(s)\right)ds\right)\\ - \bar{N}_3^n\left(\theta_p^n \int_0^t \bar{Q}^{n,+}(s)ds\right) + \bar{N}_4^n\left(\theta_d^n \int_0^t \bar{Q}^{n,-}(s)ds\right).\end{aligned}$$

For $t \geq 0$, let $N^n(t) = \sum_{i=1}^4 N_i^n(t)$ and $\bar{N}^n(t) = n^{-1}N^n(nt)$. Noting that $\bar{\lambda}_p^n(t) \leq \bar{\Lambda}_p$ for each $t \in [0, T]$, and using (25), there exists a positive constant $C_1$ such that for $t \geq 0$,

$$1 + |\bar{Q}^n(t)| \leq 1 + |\bar{Q}^n(0)| + \bar{N}^n\left(C_1 \int_0^t (1 + |\bar{Q}^n(u)|)du\right). \tag{71}$$

Define for $t \geq 0$,

$$Y^n(t) = 1 + |\bar{Q}^n(0)| + \bar{N}^n\left(C_1 \int_0^t Y^n(u)du\right). \tag{72}$$

Then, we have

$$1 + |\bar{Q}^n(t)| \leq Y^n(t),\ t \geq 0. \tag{73}$$

(This is because $1 + |\bar{Q}^n(t)|$ and $Y^n(t)$ have the same initial value, and if $\{\tau_k^n\}_{k\geq 1}$ are the jump points of $N^n$, it can be shown that $1 + |\bar{Q}^n(t)| = Y^n(t)$ for $t \in [0, \tau_1^n)$, and using (71), $1 + |\bar{Q}^n(\tau_1^n)| \leq Y^n(\tau_1^n)$, and eventually, $1 + |\bar{Q}^n(t)| \leq Y^n(t)$ for $t \geq 0$.) In (72), using Ito's formula for semimartingales, we observe that $\{Y^n(t)e^{-C_1 t}; t \geq 0\}$ is a positive martingale, and so

$$\mathbb{E}\left[Y^n(t)\right] = e^{C_1 t}\left(1 + \mathbb{E}|\bar{Q}^n(0)|\right),\ t \geq 0. \tag{74}$$

Furthermore, from theorem 2.2 in Kurtz (1978), almost surely,

$$\lim_{n\to\infty} \sup_{0\leq s\leq t} |Y^n(s) - y(s)| = 0, \tag{75}$$

where for $t \geq 0$, $y(t) = (1 + |q_0|)e^{C_1 t}$, satisfying the integral equation $y(t) = 1 + |q_0| + C_1 \int_0^t y(u)du$. Furthermore, the process $Y^n(t)$ is defined as a random time change of the Poisson process $\bar{N}^n(t)$, and it has the following martingale representation (see Pang et al. 2007 and the references therein for more details). For $t \geq 0$, define

$$\mathcal{G}_t^n = \sigma\left(\bar{Q}^n(0), \bar{N}^n\left(C_1 \int_0^s Y^n(u)du\right) : 0 \leq s \leq t\right),$$

augmented by including all null sets, and let

$$\mathcal{D}^n(t) = \bar{N}^n\left(C_1 \int_0^t Y^n(u)du\right) - C_1 \int_0^t Y^n(u)du.$$

Then, $\{\mathcal{D}^n(t)\}$ is a $\{\mathcal{G}_t^n\}_{t\geq 0}$-martingale, and its associated predictable quadratic variation is

$$\langle \mathcal{D}^n \rangle(t) = C_1 \int_0^t Y^n(u)du.$$

Consequently,

$$\left[\mathcal{D}^n(t)\right]^2 - \langle \mathcal{D}^n \rangle(t) = \left(\bar{N}n^n\left(C_1 \int_0^T Y^n(u)du\right)\right)^2 - C_1 \int_0^T Y^n(u)du, \tag{76}$$

is also a $\{\mathcal{G}_t^n\}$-martingale.

**Proof of Lemma 4.1.** We note that from (72) and (73),

$$\mathbb{E}\left[\sup_{0\le t\le T}\left(\bar{Q}^n(t)\right)^2\right] \le \mathbb{E}\left[\sup_{0\le t\le T}\left(Y^n(t)\right)^2\right]$$
$$\le 2\mathbb{E}\left[\left(1+\left|\bar{Q}^n(0)\right|\right)^2\right] + 2\mathbb{E}\left[\left(\bar{N}^n\left(C_1\int_0^T Y^n(u)du\right)\right)^2\right], \tag{77}$$
$$\le 2\mathbb{E}\left[\left(1+\left|\bar{Q}^n(0)\right|\right)^2\right] + 2C_1\int_0^T \mathbb{E}\left[\sup_{0\le t\le u}\left(Y^n(t)\right)^2\right]du, \tag{78}$$

where (77) follows from an inequality $(a+b)^2 \le 2a^2+2b^2$, and (78) follows from the fact in (76). From Gronwall's inequality, it is seen that

$$\sup_{n\in\mathbb{N}}\mathbb{E}\left[\sup_{0\le t\le T}\left(\bar{Q}^n(t)\right)^2\right] \le \sup_{n\in\mathbb{N}}\mathbb{E}\left[\sup_{0\le t\le T}\left(Y^n(t)\right)^2\right] \le 2\sup_{n\in\mathbb{N}}\mathbb{E}\left[\left(1+\left|\bar{Q}^n(0)\right|\right)^2\right]e^{2C_1T}, \tag{79}$$

which concludes the proof.

**Proof of Lemma 4.2.** Let $\lambda_p^n$ be an admissible production rate and $\tilde{q}^n$ be defined by (45). The proof can be divided into two steps. The first step is to show that the fluid-scaled queue length is bounded, and the second is to show fluid approximation under any given admissible production rate.

**Step I.** We show that $\bar{Q}^n$ can be formulated as in (80), where $\bar{N}_5^{n,c}$ converges to zero in mean.

The fluid-scaled queue-length process is given by

$$\bar{Q}^n(t) = \bar{Q}^n(0) + \bar{N}_5^{n,c}(t) + \int_0^t \bar{\lambda}_p^n(s)ds - \int_0^t \bar{\lambda}_d^n\left(s,\bar{Q}^n(s)\right)ds - \theta_p^n\int_0^t \bar{Q}^{n,+}(s)ds + \theta_d^n\int_0^t \bar{Q}^{n,-}(s)ds, \tag{80}$$

where

$$\bar{N}_5^{n,c}(t) = \bar{N}_1^n\left(\int_0^t \bar{\lambda}_p^n(s)ds\right) - \int_0^t \bar{\lambda}_p^n(s)ds$$
$$- \bar{N}_2^n\left(\int_0^t \bar{\lambda}_d^n\left(s,\bar{Q}^n(s)\right)ds\right) + \int_0^t \bar{\lambda}_d^n\left(s,\bar{Q}^n(s)\right)ds$$
$$- \bar{N}_3^n\left(\theta_p^n\int_0^t \bar{Q}^{n,+}(s)ds\right) + \theta_p^n\int_0^t \bar{Q}^{n,+}(s)ds$$
$$+ \bar{N}_4^n\left(\theta_d^n\int_0^t \bar{Q}^{n,-}(s)ds\right) - \theta_d^n\int_0^t \bar{Q}^{n,-}(s)ds.$$

We next show that $\mathbb{E}[\sup_{0\le t\le T}|\bar{N}_5^{n,c}(t)|] \to 0$. For $t\ge 0$, define $\bar{N}_6^{n,c}(t) = \sum_{i=1}^4 |\bar{N}_i^n(t)-t|$. From Doob's inequality for submartingales and Hölder's inequality, we have

$$\mathbb{E}\left[\sup_{0\le t\le T}\bar{N}_6^{n,c}(t)\right] = \sum_{i=1}^4 \mathbb{E}\left[\sup_{0\le t\le T}|\bar{N}_i^n(t)-t|\right] \le \sum_{i=1}^4\sqrt{\mathbb{E}\left[\left(\sup_{0\le t\le T}|\bar{N}_i^n(t)-t|\right)^2\right]}$$
$$= \sum_{i=1}^4\sqrt{\mathbb{E}\left[\sup_{0\le t\le T}|\bar{N}_i^n(t)-t|^2\right]} \le 2\sum_{i=1}^4\sqrt{\mathbb{E}\left[|\bar{N}_i^n(T)-T|^2\right]} = 2\sum_{i=1}^4\sqrt{nT/n^2} \to 0. \tag{81}$$

Thus, from (73), (74), and (81), for any $\epsilon > 0$,

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}\left(\sup_{0\le s\le T} |\bar{N}_5^{n,c}(s)| > \epsilon\right) \\
&\le \limsup_{n\to\infty} \mathbb{P}\left(\sup_{0\le s\le C_1\int_0^T |Y^n(u)|du} \left|\bar{N}_6^{n,c}(s)\right| > \epsilon\right) \\
&\le \lim_{K\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{0\le s\le C_1\int_0^T |Y^n(u)|du} \left|\bar{N}_6^{n,c}(s)\right| > \epsilon, \int_0^T |Y^n(u)|du > K\right) \\
&\quad + \lim_{K\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{0\le s\le C_1\int_0^T |Y^n(u)|du} \left|\bar{N}_6^{n,c}(s)\right| > \epsilon, \int_0^T |Y^n(u)|du \le K\right) \\
&\le \lim_{K\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\int_0^T |Y^n(u)|du > K\right) \\
&\qquad + \lim_{K\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\sup_{0\le s\le C_1 K} \left|\bar{N}_6^{n,c}(s)\right| > \epsilon\right) = 0,
\end{aligned}$$

which establishes that $\sup_{0\le s\le T} |\bar{N}_5^{n,c}(s)| \to 0$, in probability. We further note that similar to (71), for the same $\bar{N}^n$ and positive constant $C_1$ as in (71),

$$\begin{aligned}
&\sup_n \mathbb{E}\left[\sup_{0\le s\le T} |\bar{N}_5^{n,c}(s)|^2\right] \\
&\le \sup_n \mathbb{E}\left[\bar{N}^n\left(C_1\int_0^T \left(1 + |\bar{Q}^n(u)|\right)du\right) + \int_0^T C_1\left(1 + |\bar{Q}^n(u)|\right)du\right]^2 \\
&\le \sup_n \mathbb{E}\left[Y^n(T) + C_1\int_0^T Y^n(u)du\right]^2 < \infty,
\end{aligned}$$

where the last two inequalities follow from (72), (73), and Lemma 4.1. The above uniform integrability of $\sup_{0\le s\le T} |\bar{N}_5^{n,c}(s)|$ yields that

$$\mathbb{E}\left[\sup_{0\le s\le T} |\bar{N}_5^{n,c}(s)|\right] \to 0. \tag{82}$$

**Step II.** We show that $\mathbb{E}[\sup_{0\le t\le T} |\bar{Q}^n(t) - \tilde{q}^n(t)|] \to 0$.
We now note that for $t \ge 0$,

$$\begin{aligned}
|\bar{Q}^n(t) - \tilde{q}^n(t)| \le |\bar{Q}^n(0) - q_0| + |\bar{N}_5^{n,c}(t)| + \int_0^t |\bar{\lambda}_d^n\left(s, \bar{Q}^n(s)\right) - \bar{\lambda}_d\left(s, \tilde{q}^n(s)\right)|ds \\
+ |\theta_p^n - \bar{\theta}_p| \int_0^t \bar{Q}^{n,+}(s)ds + |\theta_d^n - \bar{\theta}_d| \int_0^t \bar{Q}^{n,-}(s)ds \\
+ \left(\bar{\theta}_p + \bar{\theta}_d\right) \int_0^t |\bar{Q}^n(s) - \tilde{q}^n(s)|.
\end{aligned}$$

For a positive constant $C_2$, we divide the following term into three parts.

$$\begin{aligned}
&\int_0^t |\bar{\lambda}_d^n(s,\bar{Q}^n(s)) - \bar{\lambda}_d(s,\tilde{q}^n(s))|ds \\
&\quad = \int_0^t |\bar{\lambda}_d(s,\bar{Q}^n(s)) - \bar{\lambda}_d(s,\tilde{q}^n(s))| 1_{\left\{\sup_{0\le u\le t}|\bar{Q}^n(u)|\le C_2\right\}} ds \\
&\qquad + \int_0^t |\bar{\lambda}_d(s,\bar{Q}^n(s)) - \bar{\lambda}_d(s,\tilde{q}^n(s))| 1_{\left\{\sup_{0\le u\le t}|\bar{Q}^n(u)|> C_2\right\}} ds \\
&\qquad + \int_0^t |\bar{\lambda}_d^n(s,\bar{Q}^n(s)) - \bar{\lambda}_d(s,\bar{Q}^n(s))| ds.
\end{aligned}$$

Noting that $\tilde{q}^n$ is uniformly bounded on $[0,T]$ (see (12)), and using (39) in Assumption 3(ii), we have for some $C_3 > 0$,

$$\int_0^T |\bar{\lambda}_d(s,\bar{Q}^n(s)) - \bar{\lambda}_d(s,\tilde{q}^n(s))| 1_{\left\{\sup_{0\le s\le T}|\bar{Q}^n(s)|\le C_2\right\}} ds \le C_3 \int_0^T |\bar{Q}^n(s) - \tilde{q}^n(s)| ds. \tag{83}$$

Hence, using (83) and (73), we have

$$\begin{aligned}
|\bar{Q}^n(t) - \tilde{q}^n(t)| \le |\bar{Q}^n(0) - q_0| + |\bar{N}_5^n(t)| + \left(\left|\theta_p^n - \bar{\theta}_p\right| + |\theta_d^n - \bar{\theta}_d|\right) \int_0^t Y^n(s) ds \\
+ \left(\bar{\theta}_p + \bar{\theta}_d + C_3\right) \int_0^t \left|\bar{Q}^n(s) - \tilde{q}^n(s)\right| ds + O^n(t),
\end{aligned}$$

where

$$\begin{aligned}
O^n(t) = \int_0^t |\bar{\lambda}_d(u,\bar{Q}^n(u)) - \bar{\lambda}_d(u,\tilde{q}^n(u))| 1_{\left\{\sup_{0\le u\le t}|\bar{Q}^n(u)|> C_2\right\}} du \\
+ \int_0^t |\bar{\lambda}_d^n(u,\bar{Q}^n(u)) - \bar{\lambda}_d(u,\bar{Q}^n(u))| du.
\end{aligned} \tag{84}$$

Gronwall's inequality yields that

$$\begin{aligned}
\sup_{0\le s\le T}\left|\bar{Q}^n(s) - \tilde{q}^n(s)\right| \le \Bigg(\left|\bar{Q}^n(0) - q_0\right| + \sup_{0\le s\le T}\left|\bar{N}_5^n(t)\right| + \sup_{0\le s\le T} O^n(s) \\
+ \left(\left|\theta_p^n - \bar{\theta}_p\right| + \left|\theta_d^n - \bar{\theta}_d\right|\right) \int_0^T Y^n(s) ds\Bigg) e^{\left(\bar{\theta}_p + \bar{\theta}_d + C_3\right)T}.
\end{aligned} \tag{85}$$

Given part (i) of Assumption 3, the assumption on initial queue length, (82) in step one and (73) in Lemma 4.1, it suffices to show that $\mathbb{E}(\sup_{0\le s\le T} O^n(s)) \to 0$. Indeed from (38), (73), (75), and (12), we have for some $C_4 > 0$,

$$\begin{aligned}
&\int_0^T \mathbb{E}\left(|\bar{\lambda}_d(s,\bar{Q}^n(s)) - \bar{\lambda}_d(s,\tilde{q}^n(s))| 1_{\left\{\sup_{0\le s\le T}|\bar{Q}^n(s)|> C_2\right\}}\right) dt \\
&\le \int_0^T \mathbb{E}\left(L_1(C_4 + Y^n(s)) 1_{\{Y^n(T)>C_2\}}\right) dt \\
&\to \int_0^T L_1(C_4 + y(s)) 1_{\{y(T)>C_2\}} ds, \text{ as n} \to \infty,
\end{aligned}$$

and for $C_5 > 0$,

$$\begin{aligned}
&\int_0^T \mathbb{E}\left|\bar{\lambda}_d^n\left(s,\bar{Q}^n(s)\right) - \bar{\lambda}_d\left(s,\bar{Q}^n(s)\right)\right| ds \\
&= \int_0^T \mathbb{E}\left(\left|\bar{\lambda}_d^n\left(s,\bar{Q}^n(s)\right) - \bar{\lambda}_d\left(s,\bar{Q}^n(s)\right)\right| 1_{\left\{\sup_{0\le s\le t}|\bar{Q}^n(s)|\le C_5\right\}}\right) ds \\
&\quad + \int_0^t \mathbb{E}\left(\left|\bar{\lambda}_d^n\left(s,\bar{Q}^n(s)\right) - \bar{\lambda}_d\left(s,\bar{Q}^n(s)\right)\right| 1_{\left\{\sup_{0\le s\le T}|\bar{Q}^n(s)|> C_5\right\}}\right) ds \\
&\le \int_0^T \sup_{|x|\le C_5}\left|\bar{\lambda}_d^n(s,x) - \bar{\lambda}_d(s,x)\right| ds + \int_0^T \mathbb{E}\left(2L_1 Y^n(t) 1_{\{Y^n(T)>C_4\}}\right) dt \\
&\to \int_0^T 2L_1 y(t) 1_{\{y(T)>C_5\}} dt.
\end{aligned}$$

To summarize, we have shown that

$$\limsup_{n\to\infty} \mathbb{E}\left(\sup_{0\le s\le T}|O^n(s)|\right) \le \int_0^T L_1\left(C_4 + y(s)\right) 1_{\{y(T)>C_2\}} dt + \int_0^T 2L_1 y(s) 1_{\{y(T)>C_5\}} dt.$$

Letting $C_2$ and $C_5$ be larger than $y(T)$, the result follows.

## 8. Conclusions

The decisions on when and how much to produce in manufacturing impact the overall system. In continuous manufacturing setup, flexibility of changing production rates plays an important role in both designing optimal production rate and reducing total cost. Based on a double-ended queueing model, we formulate a stochastic queueing control problem that takes into account (1) inventory holding and perishment cost, (2) back-order and lost sale cost, and (3) production flexibility cost. The direct analysis of the QCP is intractable. We then develop a deterministic fluid-control problem that is shown to be a performance lower bound for the QCP. Furthermore, we show that the FCP lower bound can be achieved asymptotically for large-scale systems and propose an asymptotically optimal production rate for the QCP.

The aforementioned FCP is hard to solve, given the nonstationary demand process. We develop a linearized discrete time problem which is an LP and can be effectively solved by commercial solvers. Simulations of systems with various scales are conducted to validate model effectiveness as scale increases. Numerical examples are also presented to show that the controlled production is able to achieve production-cost reduction compared with constant production rate and total flexible production rate.

We intend to extend the current model to consider more realistic abandonment assumptions, such as deterministic patience/goods expiry times. We are currently working on a related fluid-limit model for performance analysis and have conducted some preliminary simulations to verify the accuracy of fluid approximations. Another extension is to consider a network with multiclass customers and multipart assembly lines. Furthermore, one can also consider the joint optimization for pricing and production capacity. The double-ended queue will serve as a base model, and more realistic features are required to expand its applicability in manufacturing and other service systems.

### Acknowledgments

The authors thank Dr. Ping Cao (University of Science and Technology of China) for providing many constructive comments. The authors also thank the associate editor and the two anonymous referees for valuable comments.

## Appendix. A FCP Numerical Solution: An LP Method

From Definition 3.1, we write out the complete fluid-limit continuous time-control problem as follows:

$$\min_{\{\bar{\lambda}_p(t);t\in[0,T]\}} \quad \bar{\mathcal{R}}(\bar{\lambda}_p), \tag{A.1}$$

$$\text{s.t.} \quad q'(t) = \bar{\lambda}_p(t) - \bar{\lambda}_d\big(t, q(t)\big) - \bar{\theta}_p q^+(t) + \bar{\theta}_d q^-(t), \tag{A.2}$$

$$0 \le \bar{\lambda}_p(t) \le \bar{\Lambda}_p, \tag{A.3}$$

$$q^+(t) = \max\big(q(t),\ 0\big), \tag{A.4}$$

$$q^-(t) = \max\big(-q(t),\ 0\big), \tag{A.5}$$

$$q(0) = q_0 \in \mathbb{R}, \quad 0 \le t \le T. \tag{A.6}$$

The objective function (A.1) and the fluid-conservation constraint (A.2) are direct results from Definition 3.1. The quantities $q^+(t)$ and $q^-(t)$ are the fluid version of $Q^+(t)$ and $Q^-(t)$, respectively. We consider a finite time interval $[0, T]$—for example, $T = 24$. Other parameters follow Assumption 1.

We resort to time discretization and introduce a linear reformulation, which not only can be solved efficiently, but also can be extended to consider linearly formulated realistic constraints. Section A.1 specifies the discrete-time problem and linear reformulation. Section A.2 provides two examples of realistic constraints.

### A.1. Discrete-Time Fluid-Optimization Problem

Let the time-discretization epoch be $\Delta t$, which can be treated as the interval where decisions are made and/or information is collected. Then, we have $q_n = q(n\Delta t)$, $\lambda_{p,n} = \bar{\lambda}_p(n\Delta t)$, $\lambda_{d,n} = \bar{\lambda}_d(n\Delta t, q(n\Delta t))$, $q_n^+ = q^+(n\Delta t)$, $q_n^- = q^-(n\Delta t)$, and linear cost $\bar{C}(\lambda_{p,n}, n\Delta t) = C_n\lambda_{p,n}$, where $n = 1, \ldots, N \equiv T/\Delta t$. The initial condition is $q_0$. All cost coefficients $h_p$, $h_d$, $c_p, c_d$, $c_f$ and the abandonment rates $\bar{\theta}_d$, $\bar{\theta}_p$ are as in the continuous time problem. Note that in the discretization of the FCP, we only allow linear terms of $q_n$ of the demand function and linear cost function, both of which are supported by Assumption 3.

Discretizing the objective function $\bar{\mathcal{R}}(\bar{\lambda}_p)$ yields:

$$\sum_{n=1}^{N}\big((h_p + c_p\bar{\theta}_p)q_n^+ + (h_d + c_d\bar{\theta}_d)q_n^- + C_n\lambda_{p,n}\big)\Delta t + c_f\sum_{n=2}^{N}|\lambda_{p,n} - \lambda_{p,n-1}|.$$

To linearize the absolute value function, define auxiliary variables $Z_n^+ \ge 0$ and $Z_n^- \ge 0$, for $n = 2, \ldots, N$, and add constraints $\lambda_{p,n} - \lambda_{p,n-1} = Z_n^+ - Z_n^-$, for $n = 2, \ldots, N$. In the objective function, we replace $|\lambda_{p,n} - \lambda_{p,n-1}| = Z_n^+ + Z_n^-$, for $n = 2, \ldots, N$.

For the constraint (A.2), use difference $(q_n - q_{n-1})/\Delta t$ to approximate the derivative $q'(n\Delta t)$ with initial condition $q_0$,

$$\frac{q_n - q_{n-1}}{\Delta t} = \lambda_{p,n} - \lambda_{d,n} - \bar{\theta}_p q_n^+ + \bar{\theta}_d q_n^-, \quad n = 1, \ldots N.$$

Instead of formulating constraints (A.4) and (A.5) by definition, we consider a new set of constraints combining $q(t)$, $q^+(t)$, and $q^-(t)$:

$$q(t) = q^+(t) - q^-(t), \quad q^+(t) \ge 0, \quad q^-(t) \ge 0, \quad 0 \le t \le T. \tag{A.7}$$

The reason that this reformulation is equivalent to constraints (A.4) and (A.5) is the following.

Note that the constraint (A.7) characterizes exactly the same underlying dynamics of the system—that is, no nonzero $q^+(t)$ and $q^-(t)$ simultaneously. Otherwise, assume there is an optimal solution containing positive $\tilde{q}^+(t)$ and $\tilde{q}^-(t)$ at an interval $[t_1, t_2]$. Then, we can construct another solution $\check{q}$ such that $\check{q}^+(t) = \tilde{q}^+(t)$ and $\check{q}^-(t) = \tilde{q}^-(t)$ for $t \in [0, t_1) \cup (t_2, T]$. For $t \in [t_1, t_2]$, define $\check{q}^+(t) = \tilde{q}^+(t) - \epsilon(t)$ and $\check{q}^-(t) = \tilde{q}^-(t) - \epsilon(t)$, where $\epsilon(t) = \min(\tilde{q}^+(t), \tilde{q}^-(t))$. Notice that constraint (A.7) is not violated under $\check{q}$, but in the objective function there is a positive deduction $\epsilon(t)((h_p + c_p\bar{\theta}_p) + (h_d + c_d\bar{\theta}_d))$. Therefore, the solution under $\tilde{q}$ is not optimal. This is a standard argument in LP.

We now provide the complete formulation for the discrete-time fluid-optimization model, which is an LP problem, as follows:

$$\begin{aligned}
\min_{\lambda_{p,n}, n=1,\ldots,N} \quad & \sum_{n=1}^{N}\big((h_p + c_p\bar{\theta}_p)q_n^+ + (h_d + c_d\bar{\theta}_d)q_n^- + C_n\lambda_{p,n}\big)\Delta t + c_f\sum_{n=2}^{N}\big(Z_n^+ + Z_n^-\big);\\
\text{s.t.} \quad & \frac{q_n - q_{n-1}}{\Delta t} = \lambda_{p,n} - \lambda_{d,n} - \bar{\theta}_p q_n^+ + \bar{\theta}_d q_n^-\\
& q_n = q_n^+ - q_n^-\\
& \lambda_{p,n} - \lambda_{p,n-1} = Z_n^+ - Z_n^-\\
& 0 \le q_n^+, \quad 0 \le q_n^-, \quad 0 \le \bar{\lambda}_p(t) \le \bar{\Lambda}_p\\
& 0 \le Z_n^+, \quad 0 \le Z_n^-\\
& n = 1, \ldots, N.
\end{aligned} \tag{A.8}$$

System inputs are $T$, $\Delta t$, $h_p$, $c_p$, $\bar{\theta}_p$, $h_d$, $c_d$, $\bar{\theta}_d$, $c_f$, $\lambda_{d,n}$, $\bar{\Lambda}_p$, $C_n$, and $q_0$. The decision variables of this LP are $q_n^+$, $q_n^-$, $Z_n^+$, $Z_n^-$ and $\lambda_{p,n}$. The control of the problem is $\lambda_{p,n}$, for $n = 1, \ldots, N$. This is an LP and can be solved very efficiently.

### A.2. Realistic Constraints

The original stochastic optimal control problem (3) is a general formulation for the double-ended system. However, the LP reformulation can be extended to incorporate other realistic features, which significantly increases the implementability and practicality of the obtained optimal solution. As an example, we present two additional linear constraints that capture different production- and service-quality requirements.

The first constraint is that the production-rate changes can only be made on prespecified (equidistant) time epochs. That is, the time length between the adjacent production-rate adjustments is fixed $L$ (e.g., $L = 8$ *hours*, a typical length of one work shift). We divide the entire time horizon $T$ into $\hat{N} = T/L$ slots, each having the length of $L$. Control variables within each interval are set to remain unchanged. Let $K = L/\Delta t$, and we have the following constraints:

$$\lambda_{p,k+i\cdot K} = \lambda_{p,k+1+i\cdot K}, \quad k = 1, \ldots, K-1, \quad i = 0, \ldots, \hat{N}-1.$$

The second kind of realistic constraint is on the total fulfillment rate. In addition to minimizing total production cost, in order to ensure a desired level of quality of service, one natural practice is to set a fulfillment target. Define $\beta$ as the minimum fulfillment requirement:

$$\beta \int_0^T \bar{\lambda}_d\big(t, q(t)\big)dt \le \int_0^T \bar{\lambda}_p(t)dt.$$

For instance, the manager may require that at least 90% of the total potential demand is covered by the scheduled production over $[0, T]$. The above fulfillment requirement constraint can easily be translated into a set of constraints in LP reformulation problem.

## References

Afeche P, Diamant A, Milner J (2014) Double-sided batch queues with abandonment: Modeling crossing networks. *Oper. Res.* 62(5):1179–1201.

Aras AK, Chen X, Liu Y (2018) Many-server Gaussian limits for overloaded non-Markovian queues with customer abandonment. *Queueing Systems* 89:1–45.

Arcidiacono P, Ellickson PB, Landry P, Ridley DB (2013) Pharmaceutical followers. *Internat. J. Indust. Organ.* 31(5):538–553.

Armony M, Atar R, Honnappa H (2019) Asymptotically optimal appointment schedules. *Math. Oper. Res.* 44(4):1345–1380.

Ata B, Kumar S (2005) Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies. *Ann. Appl. Probability* 15(1A):331–391.

Atar R, Keslassy I, Mendelson G (2019) Subdiffusive load balancing in time-varying queueing systems. *Oper. Res.* 67(6):1678–1698.

Bassamboo A, Harrison JM, Zeevi A (2005) Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems* 51(3):249–285.

Bassamboo A, RandhawaRS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Oper. Res.* 58(5):1398–1413.

Boxma OJ, David I, Perry D, Stadje W (2011) A new look at organ transplantation models and double matching queues. *Probab. Engrg. Inform. Sci.* 25:135–155.

Budhiraja A, Ghosh AP (2006) Diffusion approximations for controlled stochastic networks: An asymptotic bound for the value function. *Ann. Appl. Probab.* 16(4):1962–2006.

Corstjens M, Doyle P (1981) A model for optimizing retail space allocations. *Management Sci.* 27(7):822–833.

Cudina M, Ramanan K (2011) Asymptotically optimal controls for time-inhomogeneous networks. *SIAM J. Control Optim.* 49(2):611–645.

Feichtinger G, Hartl RF (1985) On the use of Hamiltonian and maximized Hamiltonian in nondifferentiable control theory. *J. Optim. Theory Appl.* 46(4):493–504.

Gallego G, Van Ryzin G (1994) Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Sci.* 40(8):999–1020.

Gurvich I, Ward A (2014) On the dynamic control of matching queues. *Stochastic Systems* 4(2):479–523.

Harrison JM (2000) Brownian models of open processing networks: Canonical representation of workload. *Ann. Appl. Probab.* 10(1):75–103.

He Q-C, Nie T, Shen Z-JM (2018) Beyond rebalancing: Crowd-sourcing and geo-fencing for shared-mobility systems. Working paper, Southern University of Science and Technology, Shenzhen, China.

Kabe DG, Gouranga Rao UL (1986) Direct solution to an optimal control problem of econometric systems. *Optimal Control Appl. Methods* 7(3):327–331.

Kaspi H, Perry D (1983) Inventory systems of perishable commodities. *Adv. Appl. Probab.* 15(3):674–685.

Kaspi H, Perry D (1984) Inventory systems of perishable commodities with Poisson input and renewal output. *Adv. Appl. Probab.* 16(2):402–421.

Khademi A, Liu X (2019) Asymptotically optimal allocation policies for transplant queueing systems. Working paper, Clemson University, Clemson, SC.

Khmelnitsky E, Gerchak Y (2002) Optimal control approach to production systems with inventory-level-dependent demand. *IEEE Trans. Automatic Control* 47(2):289–292.

Kurtz TG (1978) Strong approximation theorems for density dependent Markov chains. *Stochastic Processes Appl.* 6(3):223–240.

Liu X (2019) Diffusion approximations for double-ended queues with reneging in heavy traffic. *Queueing Systems* 91(1-2):49–87.

Liu X, Gong Q, Kulkarni VG (2015) Diffusion models for double-ended queues with renewal arrival processes. *Stochastic Systems* 5(1):1–61.

Liu Y (2018) Staffing to stabilize the tail probability of delay in service systems with time-varying demand. *Oper. Res.* 66(2):514–534.
Liu Y, Sun X, Hovey K (2021) Scheduling to differentiate service in a multiclass service system. *Oper. Res.* Forthcoming.
Liu Y, Whitt W (2011) A network of time-varying many-server fluid queues with customer abandonment. *Oper. Res.* 59(4):835–846.
Liu Y, Whitt W (2012a) The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems* 71(4):405–444.
Liu Y, Whitt W (2012b) A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Oper. Res. Lett.* 40(5):307–312.
Liu Y, Whitt W (2012c) Many-server heavy-traffic limits for queues with time-varying parameters. *Ann. Appl. Probab.* 24(1):378–421.
Liu Y, Whitt W (2014) Algorithms for time-varying networks of many-server fluid queues. *INFORMS J. Comput.* 26(1):59–73.
Mandelbaum A, Pats G (1998) State-dependent stochastic networks. Part I. Approximations and applications with continuous diffusion limits. *Ann. Appl. Probab.* 8(2):569–646.
Mandelbaum A, Massey WA, Reiman MI (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30:149–201.
Matula J (1987) On an extremum problem. *J. Australian Math. Soc. B Appl. Math.* 28(3):376–392.
Milgrom P, Segal I (2002) Envelope theorems for arbitrary choice sets. *Econometrica* 70(2):583–601.
Niyirora J, Pender J (2016) Optimal staffing in nonstationary service centers with constraints. *Naval Res. Logist.* 63(8):615–630.
Ozkan E, Ward A (2020) Dynamic matching for real-time ridesharing. *Stochastic Systems* 10(1):29–70.
Pang G, Talreja R, Whitt W (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys* 4:193–267.
Pender J (2016) Risk measures and their application to staffing nonstationary service systems. *Eur. J. Oper. Res.* 254(1):113–126.
Perry D, Stadje W (1999) Perishable inventory systems with impatient demands. *Math. Methods Oper. Res.* 50(1):77–90.
Pontryagin LS (2018) *Mathematical Theory of Optimal Processes* (Routledge, Abingdon, UK).
Prabhakar B, Bambos N, Mountford TS (2000) The synchronization of Poisson processes and queueing networks with service and synchronization nodes. *Adv. Appl. Probab.* 32(3):824–843.
Seierstad A, Sydsaeter K (1986) *Optimal Control Theory with Economic Applications*, *Advanced Textbooks in Economics*, vol. 24 (Elsevier North-Holland, Inc., Amsterdam).
Whitt W (2006) Fluid models for multiserver queues with abandonments. *Oper. Res.* 54(1):37–54.
Yu L (2016) Continuous manufacturing has a strong impact on drug quality. *FDA Voice* (April 13), 12.
Zenios SA (1999) Modeling the transplant waiting list: A queueing model with reneging. *Queueing Systems* 31:239–251.