APRIL 8, 2019
OPTIMAL CONTROL OF A TIME-VARYING DOUBLE-ENDED PRODUCTION QUEUEING MODEL

By Chihoon Lee[*], Xin Liu[†],
Yunan Liu[‡], and Ling Zhang[‡],

Stevens Institute of Technology[*],Clemson University[†], and North Carolina State University [‡]

Motivated by production systems with nonstationary stochastic demand, we study a double-ended queueing model having backorders and customer abandonment. One side of our model stores backorders, and the other side represents inventory. We assume first-come-first-serve 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, perishment) and demand-related (waiting, 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 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.

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, time-since-entry), over-the-counter offerings, on-going market competition may lead to nonstationary demand behavior over time; for instance, Figure 1 in [3] shows monthly prescriptions fluctuation 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, the continuous manufacturing can produce more

Keywords and phrases: time-varying demand, asymptotic optimality, double-ended queue, abandonment, optimal control, diffusion refinement

reliable products, while achieving higher efficiency in cost. Recognizing its value, the Food and Drug Administration (FDA) approved in April 2016, for the first time, Janssen Products, LP's change in their production method from batch to continuous manufacturing; see [35]. Moreover, continuous manufacturing allows manufacturers to adjust for nonstationary (i.e., time-varying) demand much quicker, and hence preventing a potential shortage or large backorder 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, backorders are subject to cancellations; (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 Fig. 1), in which goods are produced according to a Poisson process with time-varying production rate, and demands arrive according to another Poisson process 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 3.1). 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 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 system capacity can scale up *large* in response to the demand. This enables us to pursue approximate solutions based on asymptotic analysis. We show 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, e.g., 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 a 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 $o(n)$.

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

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. [14, 1, 36, 32, 7, 13]). In particular, [14] 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 [15, 30]. When renewal arrivals and/or generally distributed patience times are considered, exact analysis becomes intractable. In [18], rigorous fluid and diffusion models were developed for double-ended queues with renewal arrivals and exponential patience times. Later on, [17] establishes the heavy traffic asymptotics for the system with renewal arrivals and generally distributed patience times.

Staffing and control of time-varying queues. Heavy-traffic fluid and diffusion limits were developed by [24] for time-varying Markovian queueing networks with Poisson arrivals and exponential service times. Gaussian approximation methods for Markovian queues have been developed by [27, 29]. Adopting a two-parameter queue length descriptor, the pioneering work by [34] studied the $G/GI/s+GI$ fluid model having non-exponential service and abandonment times. Extending the work by [34], [20] developed a fluid approximation for the $G_t/GI/s_t+GI$ queue with time-varying arrivals and non-exponential distributions; they later extended it to the framework of fluid networks [19, 23]. A functional weak law of large numbers (FWLLN) [21] was established to substantiate the fluid approximation in [20] and a functional central limit theorem (FCLT) [22] 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. [12, 4], and [8]). For time-inhomogeneous queueing systems, the asymptotic control is usually considered under fluid scaling, cf. [5, 9, 28], 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. Without assuming any asymptotics, in many situations, deterministic models can also be shown to be the best performance bound for the expected stochastic performance.

To the best of our knowledge, all existing results are established for systems with linear drift, e.g., in [11], 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 [6] 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.

Organization of the paper. In §2 we introduce our double-ended queueing model with abandonment at both sides and formulate a finite time horizon queueing control problem (QCP). In §3 we develop a deterministic fluid control problem (FCP) that provides a lower bound for the QCP. In §4 we consider high-volume systems, 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 §5. Additional proofs are given in §6, and conclusions are given in §7. At 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 first-come-first-serve (FCFS) principle. We further assume that the products are perishable, and their lifetimes are identically and independently distributed (i.i.d.) exponential random variables, and 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.

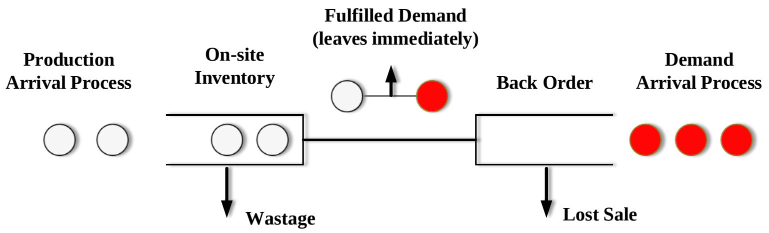

Fig 1: A production/inventory system modeled by a double-ended queue.

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 $\mathbb{E}$.

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 which represent the production rate, and demand arrival rate, respectively. Parameters $\theta_p \geq 0$ and $\theta_d \geq 0$ are the goods perishment and backorder abandonment rates (the reciprocals of the means of product shelf life and demand patience time). For the production system with non-perishable goods or infinitely patient customers, we can let $\theta_p = 0$ or $\theta_d = 0$, which is a special case of our model here.

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

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

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

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

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 σ-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 a 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), monotonicity of total variation (Proposition 3.2) are presented. The optimal solution of the FCP will later 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 [21]). We note that from (2), for $t \ge 0$,

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

and

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

The objective function (3) can then be written as

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

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

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

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

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

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

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

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

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 (FCP) 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*

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

subject to

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

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

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

(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 λ_d and the production cost function C.

Assumption 1. (i) *There exists a positive constant* L_1 *such that for* $t \geq 0$ *and* $x \in \mathbb{R}$,

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

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*

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

(ii) *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 [25].

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 [25]. 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 $\mu(t) \leq f(q(t), \lambda_p(t))$ for any admissible pair (q, λ_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; (iv) there exists an integrable function $m(\cdot)$ such that $|g(t, q(t), \lambda_p(t))| \leq m(t)$ for each admissible pair (q, λ_p) and for almost all $t \in [0,T]$.

Clearly f is continuous, and has a lower bound which can be taken to be 0, and furthermore, g is continuous, and from (10) in Assumption 1,

$|g(t,q(t),u(t))| \leq \Lambda_p + L_1(1+|q(t)|) + (\theta_d+\theta_p)|q(t)|$. To show the integrability of q, we observe that

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

and from Gronwall's inequality (see Section 1 of online appendix of [2] for a reference), we have

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

This verifies the sufficient conditions. □

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 $\mathcal{M}$, i.e.,

$$\mathcal{V}(\mathcal{M}) \equiv \min_{\bar{\lambda}_p} \bar{\mathcal{R}}(\bar{\lambda}_p;\mathcal{M}).$$

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

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

PROPOSITION 3.1 (Linear scalability property). *Assume that the production cost $C(\cdot)$ is linear, i.e., for $x \in \mathbb{R}_+$, $C(x) = C_0 x$ for some $C_0 > 0$, and the demand arrival rate is bounded and state independent, that is for $t \geq 0$ and $x \in \mathbb{R}$, $\lambda_d(t,x) \equiv \lambda_d(t)$ and $\sup_{t\geq 0}\lambda_d(t) < \infty$. Then 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,

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

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

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

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 $\mathcal{V}(\mathcal{M}^\kappa)$,

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

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

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

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_κ^* satisfies

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

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

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

where the state process is q_κ^*/κ. 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 κq^*. Additionally, due to (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 Fig. 3 in Section 5, which shows that when c_f is within a moderate range the total variation term affects the optimal fluid production rate in a non-trivial manner.

PROPOSITION 3.2 (Monotonicity of Total Variation). *Under Assumption 1, fix all the input data in $\mathcal{M}$ except the flexiblity 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

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

We parameterize the problem with c_f. Define an optimal value function parameterized by c_f as

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

which can be also written as

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

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 [26], which concerns about 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

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

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. *Under the assumption of Proposition 3.1, 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 λ_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

$$\begin{aligned}&\int_0^T \Big[(h_p + c_p\theta_p)\mathbb{E}[Q^+(t)] + (h_d + c_d\theta_d)\mathbb{E}[Q^-(t)]\Big]dt \\ &\qquad \geq \int_0^T \Big[(h_p + c_p\theta_p)q^+(t) + (h_d + c_d\theta_d)q^-(t)\Big]dt,\end{aligned} \tag{19}$$

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

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

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:

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

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

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

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

where Δ_0 is a given constant and $d(t) = \mathbb{E}[\lambda_p(t)] - \lambda_d(t)$.

LEMMA 3.1. *Assume that $\theta_p \neq \theta_d$ and fix an admissible control $\lambda_p \in \mathcal{U}$. Then $u^*(t) = 0, t \in [0, T]$, is an optimal control for the problem* (21) *–* (23)*.*

The proof of Lemma 3.1 is given in Section 6.

In the following we finish the proof of Theorem 3.2. Let

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

Now the constraint (7) becomes

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

Furthermore, Lemma 4.1 shows that $\sup_{0 \leq t \leq T} \mathbb{E}[|Q(t)|] < \infty$, which yields that there exists $\Delta_0 > 0$ such that $u_1(t) \leq \Delta_0$ for all $t \in [0, T]$. This shows that $u_1(t)$ is an admissible control to the auxiliary control problem in Lemma 3.1, and the corresponding state process is $\{\mathbb{E}[Q(t)]; t \in [0, T]\}$.

From Lemma 3.1, $u^*(t) = 0, t \in [0, T]$ is an optimal solution, and we have

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

where

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

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 the Pontryagin's Maximum Principle; see [31, 33]. 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. See Section A in Appendix for details of the LP reformulation.*

The FCP provides a convenient performance lower bound. Theorem 3.2 states the existence of such lower bound for any admissible control. In §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 4.1), 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.

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 §2, and for the n^{th} system, we append a superscript n to the quantities introduced in §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 λ_d^n is $\mathcal{O}(n)$ and θ_p^n and θ_d^n are $\mathcal{O}(1)$ (see Assumption 2 (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^{th} system is to choose $\{\lambda_p^n(t); t \in [0, T]\} \in \mathcal{U}^n$ with an objective of minimizing

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

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^{th} system. A production rate function $\{\lambda_p^n(t); t \in [0, T]\}$ is *admissible* in the n^{th} system if it satisfies the following conditions:

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

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

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

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

To study the asymptotic properties of the system, we 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

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

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

$$\begin{aligned}
\bar{\mathcal{R}}^n(\bar{\lambda}_p^n; \mathcal{M}^n) \equiv \mathbb{E}\bigg(\int_0^T \Big[h_p \bar{Q}^{n,+}(t) + h_d \bar{Q}^{n,-}(t) + \bar{C}^n(\bar{\lambda}_p^n(t))\Big] dt \\
+ c_p \bar{G}_p^n(T) + c_d \bar{G}_d^n(T) + c_f V_T(\bar{\lambda}_p^n)\bigg).
\end{aligned} \tag{25}$$

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

$$\lim_{n\to\infty} \bar{\mathcal{R}}^n(\bar{\lambda}_p^{n,*}; \mathcal{M}^n) = \inf \liminf_{n\to\infty} \bar{\mathcal{R}}^n(\bar{\lambda}_p^n; \mathcal{M}^n), \tag{26}$$

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 2.

(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* $L_0 > 0$,

$$\sup_{|x|\le L_0} |\bar{\lambda}^n_d(t,x) - \bar{\lambda}_d(t,x)| \to 0, \quad \textit{as } n\to\infty. \tag{27}$$

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

$$\bar{\lambda}^n_d(t,x) \le L_1(1+|x|), \tag{28}$$

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*

$$\sup_{t\in K_1, x,y\in K_2} |\bar{\lambda}_d(t,x)-\bar{\lambda}_d(t,y)| \le L_2|x-y|. \tag{29}$$

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

$$\sup_{|x|\le L_3} |\bar{C}^n(x)-\bar{C}(x)| \to 0, \quad \textit{as } n\to\infty. \tag{30}$$

Remark 4.1.

(i) *From Assumption 2 (i) and (ii), we see that the goods perishment and backorder abandonment rates* θ^n_p *and* θ^n_d *are* $\mathcal{O}(1)$*, while the arrival rate of demands is* $\mathcal{O}(n)$*.*

(ii) *The assumptions* (28) *and* (29) *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* (30) *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.*

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 $\bar{\lambda}^n_p(\cdot)$ 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 $t\in[0,T]$,

$$q(t) = q_0 + \int_0^t \Big[\bar{\lambda}_p(s) - \bar{\lambda}_d(s,q(s)) - \bar{\theta}_p q^+(s) + \bar{\theta}_d q^-(s)\Big] ds, \tag{31}$$

and the associated fluid scaled cost $\bar{\mathcal{R}}^n(\lambda^n_p;\mathcal{M}^n)$ 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 2, 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 4.1 (Asymptotic Optimality of FCP). *Under Assumption 2, $\{n\bar{\lambda}_p^*\}_{n\geq 1}$ is asymptotically optimal for the control problem* (24) *of the n^{th} system under the fluid scaling, namely,*

$$\lim_{n\to\infty} \bar{\mathcal{R}}^n(\bar{\lambda}_p^*; \mathcal{M}^n) = \mathcal{V}(\bar{\mathcal{M}}), \tag{32}$$

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

$$\liminf_{n\to\infty} \bar{\mathcal{R}}^n(\bar{\lambda}_p^n; \mathcal{M}^n) \geq \mathcal{V}(\bar{\mathcal{M}}). \tag{33}$$

An immediate consequence of Theorem 4.1 gives the following corollary, and we omit its proof.

COROLLARY 4.1. *Under Assumption 2, considering a production rate λ_p^n such that*

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

then λ_p^n is asymptotically optimal under the fluid scaling.

From Remark 4.1(iii), the limit production cost $\bar{C}(\cdot)$ is linear. Assume that the limit arrival rate $\bar{\lambda}_d$ is state-independent, i.e., $\bar{\lambda}_d(t,x) \equiv \bar{\lambda}_d(t)$ for $t \in [0,T]$. Define $\tilde{\mathcal{M}}^n = (nq_0, n\bar{\lambda}_d, n\bar{\Lambda}_p, \bar{\theta}_p, \bar{\theta}_d, \bar{C}, h_p, h_d, c_p, c_d, c_f)$. The following corollary follows from Theorem 3.1, Proposition 3.1, and (32) of Theorem 4.1, and its proof will be omitted.

COROLLARY 4.2. *For $n \in \mathbb{N}$, the FCP associated with $\tilde{\mathcal{M}}^n$ admits an optimal solution $\lambda_p^{n,*} = n\bar{\lambda}_p^*$. Furthermore, for each $n \in \mathbb{N}$,*

$$\mathcal{R}^n(\lambda_p^{n,*}; \tilde{\mathcal{M}}^n) \geq \mathcal{V}(\tilde{\mathcal{M}}^n),$$

and as $n \to \infty$,

$$\mathcal{R}^n(\lambda_p^{n,*}; \tilde{\mathcal{M}}^n) = \mathcal{V}(\tilde{\mathcal{M}}^n) + o(n).$$

REMARK 4.2. *Theorems 3.2 and 4.1 provide a useful framework to construct asymptotically optimal production rate functions for systems with large demand rates. Specifically, Corollary 4.2 says that the optimal solution of the FCP is optimal for the QCP with error $o(n)$ as the system scale*

n grows. In §5, we illustrate how the FCP solution achieves asymptotic optimality. See §A of Appendix for the detailed solution procedure for the FCP.

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 [20] for similar observation).

Proof of Theorem 4.1. The following two lemmas will be used in the proof of Theorem 4.1. In particular, Lemma 4.1 establishes the uniform integrability of $\bar{Q}^n$, and Lemma 4.2 is essentially the fluid approximation under an arbitrary admissible production rate λ_p^n, in which the process $\tilde{q}^n$ can be interpreted as the fluid limit under λ_p^n. For proofs of these two lemmas, see §6.

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

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

Given a deterministic $q_0 \in \mathbb{R}$, under an admissible production rate λ_p^n, define the following stochastic process.

$$\tilde{q}^n(t) = q_0 + \int_0^t \left[\bar{\lambda}_p^n(s) - \bar{\lambda}_d(s, \tilde{q}^n(s)) - \bar{\theta}_p \tilde{q}^{n,+}(s) + \bar{\theta}_d \tilde{q}^{n,-}(s)\right] ds. \tag{34}$$

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

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

Let $n\bar{\lambda}_p^*$ be the production rate for the n^{th} system. From Lemma 4.2, under the production rate $n\bar{\lambda}_p^*$, we have

$$\mathbb{E}\left(\sup_{0\le t\le T} |\bar{Q}^n(t) - q^*(t)|\right) \to 0, \tag{35}$$

where for $t \geq 0$,

$$q^*(t) = q_0 + \int_0^t \left[\bar{\lambda}_p^*(s) - \bar{\lambda}_d(s, q^*(s)) - \bar{\theta}_p q^{*,+}(s) + \bar{\theta}_d q^{*,-}(s)\right] ds. \tag{36}$$

We proceed to prove Theorem 4.1 using Lemmas 4.1 and 4.2. We first show (32) in Theorem 4.1. Note that from (25), we have

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

From (35), we have that

$$\begin{aligned}&\mathbb{E}\left(\sup_{0\leq t\leq T} |\bar{Q}^{n,+}(t) - q^{*,+}(t)|\right) + \mathbb{E}\left(\sup_{0\leq t\leq T} |\bar{Q}^{n,-}(t) - q^{*,-}(t)|\right) \\ &\leq 2\mathbb{E}\left(\sup_{0\leq t\leq T} |\bar{Q}^n(t) - q^*(t)|\right) \to 0. \qquad (37)\end{aligned}$$

We next note that

$$\begin{aligned}&\mathbb{E}\left[\sup_{0\leq t\leq T}\left|\bar{N}_3^n\left(\theta_p^n \int_0^t \bar{Q}^{n,+}(s) ds\right) - \bar{\theta}_p \int_0^t q^{*,+}(s) ds\right|\right] \\ &\leq \mathbb{E}\left[\sup_{0\leq t\leq T}\left|\bar{N}_3^n\left(\theta_p^n \int_0^t \bar{Q}^{n,+}(s) ds\right) - \bar{\theta}_p^n \int_0^t \bar{Q}^{n,+}(s) ds\right|\right] \qquad (38) \\ &\quad + |\theta_p^n - \bar{\theta}_p| \int_0^T q^{*,+}(s) ds + \theta_p^n \int_0^T \mathbb{E}\left[\left|\bar{Q}^{n,+}(s) - q^{*,+}(s)\right|\right] ds \qquad (39)\end{aligned}$$

From (37) and Assumption 2 (i), the summation in (39) converges to 0. The expectation in (38) also converges to 0, which follows from the proof of Lemma 4.1 (see (62)). Using a similar argument, we have

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

Now (32) follows from the above convergence and Assumption 2 (iii). To show (33), let $\{\lambda_p^n\}_{n\geq 1}$ be an arbitrary admissible sequence of production rates, and define $\tilde{q}^n$ as in (34). We first note that

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

which follows from Lemma 4.1, and the fact that

$$\limsup_{n\to\infty}\Big[\mathbb{E}(\bar{G}_p^n(T))+\mathbb{E}(\bar{G}_d^n(T))\Big]\le \limsup_{n\to\infty}\mathbb{E}(\bar{N}_1^n(\bar{\Lambda}_p T))+\mathbb{E}\left[\bar{N}_2^n\left(\int_0^T L_1(1+|\bar{Q}^n(s)|)\right)\right]<\infty,$$

which holds due to the boundedness of queue length by Lemma 4.1. Next from Lemma 4.2, we have

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

and using similar arguments in the proof of (38) and (39), it can be shown that

$$\begin{aligned}&\mathbb{E}\left[\left|\bar{N}_2^n\left(\theta_p^n\int_0^T\bar{Q}^{n,+}(t)dt\right)-\bar{\theta}_p\int_0^T\tilde{q}^{n,+}(t)dt\right|\right]\to 0,\\ &\mathbb{E}\left[\left|\bar{N}_4^n\left(\theta_d^n\int_0^T\bar{Q}^{n,-}(t)dt\right)-\bar{\theta}_d\int_0^T\tilde{q}^{n,-}(t)dt\right|\right]\to 0.\end{aligned}\tag{40}$$

It follows that

$$\left|\bar{\mathcal{R}}^n(\bar{\lambda}_p^n;\mathcal{M}^n)-\mathbb{E}(\bar{\mathcal{R}}(\bar{\lambda}_p^n;\bar{\mathcal{M}}))\right|\to 0,\tag{41}$$

where

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

Since $\mathcal{V}(\bar{\mathcal{M}})$ is the optimal value of the FCP, we must have $\mathbb{E}(\bar{\mathcal{R}}(\bar{\lambda}_p^n;\bar{\mathcal{M}}))\ge\mathcal{V}(\bar{\mathcal{M}})$ for each n. Thus from (41), we have $\liminf_{n\to\infty}\bar{\mathcal{R}}^n(\bar{\lambda}_p^n;\mathcal{M}^n)\ge\mathcal{V}(\bar{\mathcal{M}})$. □

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 §A of Appendix.

5.1. *Asymptotic Optimality.* 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 lost 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 $c_f = 1$. As an example to extend the LP reformulation, we also consider a requirement that changes made to production rate can only occur every ΔT amount of time. Here we let $\Delta T = 4$. Detailed formulation and another example with the extension of realistic constraints are presented in §A.2.

The optimal production rate $\bar{\lambda}_p^*(t)$ of the FCP associated with $\bar{\mathcal{M}}$ is the dash line in panel one of Fig. 2, where the solid line is the demand rate $\bar{\lambda}_d(t)$. We denote the optimal inventory and backorder 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 (36). For scale n system, we implement production rate $n\bar{\lambda}_p^*$ and set parameters to be of $\tilde{\mathcal{M}}^n$ (recall that $\tilde{\mathcal{M}}^n$ is defined above Corollary 4.2). In Fig. 2, we plot $\mathbb{E}[\bar{Q}^{n,+}(t)]$ and $\mathbb{E}[\bar{Q}^{n,-}(t)]$ (solid lines in the last two 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, backorder and total cost from simulations can be better approximated by their fluid counterparts. See the second and third panel of Fig. 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 backorder size (negative part of the queue length) are proportional to the total inventory holding cost and the total backorder 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.

Furthermore, although the true optimal solution to QCP is unknown, we can still validate the optimality of FCP solution. Based on Corollary 4.2, it is clear that for each n^{th} system the FCP associated with $\tilde{\mathcal{M}}^n$ serves as a lower bound, and as $n \to \infty$, the gap between the QCP and the FCP lower bound under the proposed production rate $n\bar{\lambda}_p^*$ is $o(n)$. Thus the proposed production rate $n\bar{\lambda}_p^*$ is asymptotically optimal with error $o(n)$.

5.2. *Impact of the Flexibility Cost.* Now we discuss the impact of production flexibility cost. In Fig. 3, the total costs of three cases with different

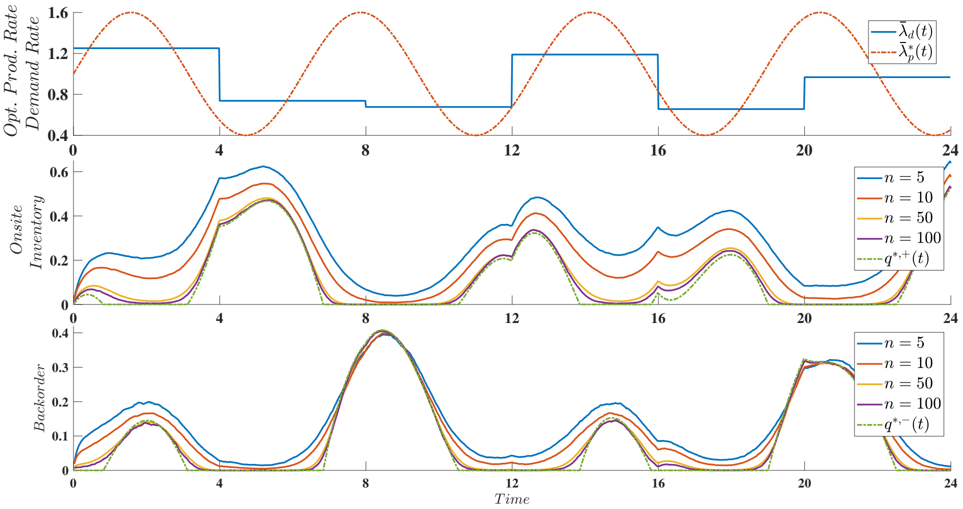

Fig 2: $n = 5, 10, 50, 100$. Let $T = 24$, demand rate $\lambda_d^n(t) = n\bar{\lambda}_d(t)$, $\bar{\lambda}_d(t) = 1+0.6\sin(t)$, and $\theta_p^n = 0.5$, $\theta_d^n = 2$, and unit costs are $c_d = 5$, $c_p = 3$, $h_d = 5$, $h_d = 3$, $c_f = 1$, $C_0 = 2$. Allow production rate change occur every $\Delta T = 4$.

Table 1
Relative difference between FCP and simulation results under $n\bar{\lambda}_p^*$

System Scale	$n = 5$	$n = 10$	$n = 25$	$n = 50$	$n = 100$	$n = 1000$
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%
Backorder 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 w/o Prod.	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%

level 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, i.e., $\bar{\lambda}_p^*(t) = \bar{\lambda}_d(t)$. 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 non-linear fashion as c_f increases. In the left panel of Fig. 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 Fig. 3, we illustrate the monotonicity of total variation as a function of c_f, which is proved in proposition 3.2.

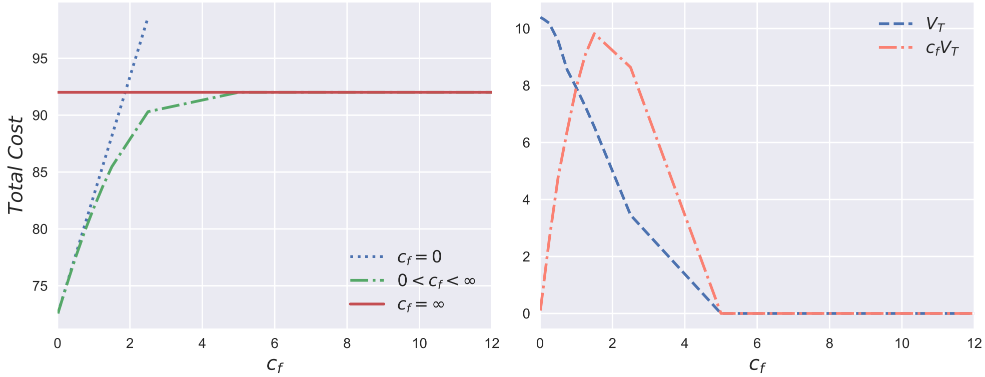

Fig 3: Total cost of three cases with different production flexibility as c_f varies.

6. Additional Proofs.

6.1. *Proof of Lemma 3.1.* To solve the control problem in Lemma 3.1, we remark that it is non-smooth as it involves the positive and negative parts of system state, which calls on an extension of the general Maximum principle. See [10]. Furthermore, the structure of the mean queue length process has a 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 stay 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 reach back to zero and repeatedly move away from and then back to zero. The main ideas of this proof is to show the optimality of $u^*(t) = 0$

for all possible paths.

Following Pontryagin's Maximum Principle (see [33]), 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

$$\begin{aligned}\mathcal{H}(q(t), u(t), t) = &-c_1 q^+(t) - c_2 q^-(t) - (c_1 + c_2)u(t) \\ &+ \lambda(t)\left(d(t) - \theta_p q^+(t) + \theta_d q^-(t) - (\theta_p - \theta_d)u(t)\right),\end{aligned} \tag{42}$$

where $\lambda(t)$ is the co-state variable. Notice that this Hamiltonian is non-smooth only at discrete time points. In order to define the co-state, we divide the time interval $[0, T]$ into four disjoint subsets A_T, B_T, C_T and D_T, i.e., $A_T \subset [0, T]$, $B_T \subset [0, T]$, $C_T \subset [0, T]$, $D_T \subset [0, T]$, $A_T \cup B_T \cup C_T \cup D_T = [0, T]$, 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 0, $t^{(i)} = s^{(i)}$. 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 non-differentiable point of $\mathcal{H}$ in q.

If $(q^*(t), u^*(t))$ is the optimal control for problem (21)–(23), following Theorem 2.1 of [10],

$$\begin{aligned}\frac{d\lambda(t)}{dt} =& -\frac{d\mathcal{H}(q(t), u(t), t)}{dq(t)} = \begin{cases} c_1 + \theta_p\lambda(t), & \text{if } t \in A_T, \\ -c_2 + \theta_d\lambda(t), & \text{if } t \in B_T, \\ 0, & \text{if } t \in C_T, \\ \partial_q\mathcal{H}(q, u, t), & \text{if } t \in D_T, \end{cases} \\ \lambda(T) =& 0,\end{aligned} \tag{43}$$

where ∂_q denotes the generalized gradient with respect to q, and

$$\mathcal{H}(q^*(t), u^*(t), t) = \max_u \mathcal{H}(q(t), u(t), t). \tag{44}$$

Notice that $\mathcal{H}$ is linear in u, therefore the optimizer of (44) is

$$u^*(t) = \begin{cases} 0, & \text{if } -(c_1 + c_2) - (\theta_p - \theta_d)\lambda(t) < 0, \\ \Delta_0, & \text{if } -(c_1 + c_2) - (\theta_p - \theta_d)\lambda(t) > 0, \\ \delta \in [0, \Delta_0], & \text{if } -(c_1 + c_2) - (\theta_p - \theta_d)\lambda(t) = 0. \end{cases} \tag{45}$$

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 λ_p, the auxiliary function $\beta(t) = -(c_1 + c_2) - (\theta_p - \theta_d)\lambda(t) < 0$, which in essence requires to solve for co-state 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 0. Roughly speaking, there are four parts. The first period which starts from the terminal time T is the last period where $q^*(t)$ is non-zero. 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 last period where $q^*(t)$ is non-zero. The last period is the rest of the total horizon and the analysis here repeats the second and third periods.

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 (43), solving $\lambda(t)$ in $t \in (s^{(m)}, T]$, we obtain

$$\lambda(t) = \frac{c_1}{\theta_p}\left(e^{\theta_p(t-T)} - 1\right), \qquad \text{if } (s^{(m)}, T] \subset A_T, \tag{46}$$

$$\lambda(t) = \frac{c_2}{\theta_d}\left(1 - e^{\theta_d(t-T)}\right), \qquad \text{if } (s^{(m)}, T] \subset B_T. \tag{47}$$

Checking the sign of the auxiliary function $\beta(t)$ in $t \in (s^{(m)}, T]$ under (46) and (47) respectively,

$$\beta(t) = -c_2 - c_1\left(1 - \left(1 - e^{\theta_p(t-T)}\right)\left(1 - \frac{\theta_d}{\theta_p}\right)\right) < 0,\ (s^{(m)}, T] \subset A_T, \tag{48}$$

$$\beta(t) = -c_1 - c_2\left(1 - \left(1 - e^{\theta_d(t-T)}\right)\left(1 - \frac{\theta_p}{\theta_d}\right)\right) < 0,\ (s^{(m)}, T] \subset B_T. \tag{49}$$

If $s^{(m)}$ does not exist, we conclude that $q(t)$ does not change sign throughout the entire decision horizon. Therefore (48) and (49) 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.

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 $(t^{(m)}, T] \subset C_T$ and from (43), $\lambda(t) = 0$ and $\beta(t) = -(c_1 + c_2)$ for $t \in (t^{(m)}, T]$. For the interval $[0, t^{(m)})$, the analysis returns to cases in (46) and (47). 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 non-trivial $s^{(m)} < T$, let

$$t^{(m)} = \sup\{t \in [0, s^{(m)}];\ q^*(t) \neq 0\}. \tag{50}$$

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 non-smooth point at $t = s^{(m)}$, where the state process $q(t)$ becomes 0 from non-zero value. Due to the continuity of the co-state process, at $t = s^{(m)}$, depending on the sign of $q^*(t)$ in the first period, we have

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

and we can check that $\beta(s^{(m)}) < 0$ for both cases from (48) and (49). Following (43), the co-state is constant on the interval $[t^{(m)}, s^{(m)}]$, i.e.,

$$\lambda(t) = \lambda(s^{(m)}),\ t \in [t^{(m)}, s^{(m)}]. \tag{51}$$

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 (48), (49) and (51) and hence $u^*(t) = 0$ for $t \in [0,T]$.

The third period. Assuming the existence of nontrivial first and second periods, i.e., $0 < t^{(m)} < s^{(m)} < T$, we define

$$s^{(m-1)} = \sup\{t \in [0, t^{(m)}]; q^*(t) = 0\}. \tag{52}$$

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 co-state process $\lambda(\cdot)$, it follows (43) and the terminal value is now $\lambda(t^{(m)})$ as in (51) 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

$$\lambda(t) = \frac{a}{b}\left(e^{b(t-t^{(m)})} - 1\right) + \lambda(s^{(m)})e^{b(t-t^{(m)})}, \tag{53}$$

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 $\beta(t) < 0$, e.g, if $q^*(t) > 0$ on both intervals,

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

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

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

If $s^{(m-1)}$ does not exist, then we must have $q^*(t) \neq 0$ for $t \in [0, t^{(m))})$, $q^*(t) = 0$ 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 (50) 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.

□

6.2. *Proof of Lemma 4.1 .* Let λ_p^n be an admissible production rate. We note that, from (1) and (2), for $t \geq 0$,

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

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 $\bar{\lambda}_p^n \leq \bar{\Lambda}_p$, and using Assumption 2 (ii), there exists a positive constant C_1 such that for $t \geq 0$,

$$1 + |\bar{Q}^n(t)| \leq 1 + |\bar{Q}^n(0)| + \bar{N}^n\left(C_1 \int_0^t \left(1 + |\bar{Q}^n(u)|\right) du\right). \tag{54}$$

Define for $t \geq 0$,

$$Y^n(t) = 1 + |\bar{Q}^n(0)| + \bar{N}^n\left(C_1 \int_0^t Y^n(u)du\right). \tag{55}$$

Then we have

$$1+|\bar{Q}^n(t)| \leq Y^n(t),\ t\geq 0. \tag{56}$$

(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 (54), $1+|\bar{Q}^n(\tau_1^n)|\leq Y^n(\tau_1^n)$, and eventually, $1+|\bar{Q}^n(t)|\leq Y^n(t)$ for $t\geq 0$.) In (55), using Ito's formula for semimartingales, we observe that $\{Y^n(t)e^{-C_1t}; t\geq 0\}$ is a positive martingale, and so

$$\mathbb{E}[Y^n(t)] = e^{C_1t}(1+\mathbb{E}|\bar{Q}^n(0)|),\ t\geq 0. \tag{57}$$

Furthermore, from Theorem 2.2 in [16], almost surely,

$$\lim_{n\to\infty}\sup_{0\leq s\leq t}|Y^n(s)-y(s)|=0, \tag{58}$$

where for $t\geq 0$, $y(t)=(1+|q_0|)e^{C_1t}$, satisfying the integral equation $y(t)=1+|q_0|+C_1\int_0^t y(u)du$. The process Y^n serves as an upper bound for the fluid scaled queue length process $\bar{Q}^n$, and will be used extensively in the following proofs.

We note that from (55) and (56),

$$\begin{aligned}
\mathbb{E}&\left[\sup_{0\leq t\leq T}(\bar{Q}^n(t))^2\right] \leq \mathbb{E}\left[\sup_{0\leq t\leq T}(Y^n(t))^2\right]\\
&\leq 2\mathbb{E}\left[(1+|\bar{Q}^n(0)|)^2\right] + 2\mathbb{E}\left[\left(\bar{N}^n\left(C_1\int_0^t Y^n(u)du\right)\right)^2\right]\\
&= 2\mathbb{E}\left[\left(\bar{N}^n\left(C_1\int_0^T Y^n(u)du\right)\right)^2 - C_1\int_0^T Y^n(u)du\right]\\
&\quad + 2\mathbb{E}\left[(1+|\bar{Q}^n(0)|)^2\right] + 2C_1\int_0^T \mathbb{E}(Y^n(u))du\\
&\leq 2\mathbb{E}\left[(1+|\bar{Q}^n(0)|)^2\right] + 2C_1\int_0^T \mathbb{E}\left[\sup_{0\leq t\leq u}(Y^n(t))^2\right]du.
\end{aligned}$$

Now using Gronwall's inequality, we conclude that

$$\begin{aligned}
\sup_{n\in\mathbb{N}}\mathbb{E}&\left[\sup_{0\leq t\leq T}(\bar{Q}^n(t))^2\right] \leq \sup_{n\in\mathbb{N}}\mathbb{E}\left[\sup_{0\leq t\leq T}(Y^n(t))^2\right]\\
&\leq 2\sup_{n\in\mathbb{N}}\mathbb{E}\Big[(1+|\bar{Q}^n(0)|)^2\Big]e^{2C_1T}.
\end{aligned} \tag{59}$$

From (54) and (59), the lemma follows. □

6.3. *Proof of Lemma 4.2.* Let λ_p^n be an admissible production rate, and $\tilde{q}^n$ be defined by (34). 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 (60), where $\bar{N}_5^n$ converges to 0 in mean.

The fluid-scaled queue length process is given by

$$\begin{aligned}\bar{Q}^n(t) = \bar{Q}^n(0) + \bar{N}_5^{n,c}(t) + \int_0^t \bar{\lambda}_p^n(s)ds - \int_0^t \bar{\lambda}_d^n(s, \bar{Q}^n(s))ds \\ - \theta_p^n \int_0^t \bar{Q}^{n,+}(s)ds + \theta_d^n \int_0^t \bar{Q}^{n,-}(s)ds,\end{aligned} \tag{60}$$

where

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

We next show that $\mathbb{E}\left[\sup_{0\le t\le T} |\bar{N}_5^{n,c}(t)|\right] \to 0$. 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

$$\begin{aligned}&\mathbb{E}\left[\sup_{0\le t\le T} \bar{N}_6^{n,c}(t)\right] \\ &= \sum_{i=1}^4 \mathbb{E}\left[\sup_{0\le t\le T} |\bar{N}_i^n(t) - t|\right] \le \sum_{i=1}^4 \sqrt{\mathbb{E}\left[\left(\sup_{0\le t\le T} |\bar{N}_i^n(t) - t|\right)^2\right]} \\ &= \sum_{i=1}^4 \sqrt{\mathbb{E}\left[\sup_{0\le t\le T} |\bar{N}_i^n(t) - t|^2\right]} \le 2\sum_{i=1}^4 \sqrt{\mathbb{E}\left[|\bar{N}_i^n(T) - T|^2\right]} \\ &= 2\sum_{i=1}^4 \sqrt{nT/n^2} \to 0.\end{aligned} \tag{61}$$

Thus from (56), (57) and (61), 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 (54), for the same $\bar{N}^n$ and positive constant C_1 as in (54),

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

where the last two inequalities follow from (55), (56), and Lemma 4.1. The above uniform integrability of $\sup_{0\le s\le T} |\bar{N}_5^{n,c}(s)|$ yields that

$$\mathbb{E}\left[\sup_{0\le s\le T} |\bar{N}_5^{n,c}(s)| \right] \to 0. \tag{62}$$

Step II. We show that $\mathbb{E}\left[\sup_{0\le t\le T} |\bar{Q}^n(t) - \tilde{q}^n(t)|\right] \to 0$.

We now note that for $t \geq 0$,

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

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

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

Noting that $\tilde{q}^n$ is uniformly bounded on $[0, T]$ (see (12)), and using (29) in Assumption 2 (ii), we have for some $C_3 > 0$,

$$
\begin{aligned}
\int_0^T |\bar{\lambda}_d(s, \bar{Q}^n(s)) - \bar{\lambda}_d(s, \tilde{q}^n(s))| 1_{\{\sup_{0 \leq s \leq T} |\bar{Q}^n(s)| \leq C_2\}} ds \\
\leq C_3 \int_0^T |\bar{Q}^n(s) - \tilde{q}^n(s)| ds.
\end{aligned}
\tag{63}
$$

Hence using (63) and (56), we have

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

where

$$
\begin{aligned}
O^n(t) = \int_0^t |\bar{\lambda}_d(u, \bar{Q}^n(u)) - \bar{\lambda}_d(u, \tilde{q}^n(u))| 1_{\{\sup_{0 \leq u \leq t} |\bar{Q}^n(u)| > C_2\}} du \\
+ \int_0^t |\bar{\lambda}_d^n(u, \bar{Q}^n(u)) - \bar{\lambda}_d(u, \bar{Q}^n(u))| du.
\end{aligned}
\tag{64}
$$

Gronwall's inequality yields that

$$\begin{aligned}\sup_{0\le s\le T}|\bar{Q}^n(s)-\tilde{q}^n(s)| \le \Bigg(&|\bar{Q}^n(0)-q_0| + \sup_{0\le s\le T}|\bar{N}_5^n(t)| + \sup_{0\le s\le T} O^n(s)\\ &+(\left|\theta_p^n-\bar{\theta}_p\right| + |\theta_d^n-\bar{\theta}_d|)\int_0^T Y^n(s)ds\Bigg)\, e^{(\bar{\theta}_p+\bar{\theta}_d+C_3)T}. \end{aligned} \tag{65}$$

Given part (i) of Assumption 2, the assumption on initial queue length, (62) in step one and (56) in Lemma 4.1, it suffices to show that $\mathbb{E}(\sup_{0\le s\le T} O^n(s)) \to 0$. Indeed from (28), (56), (58), and (12), we have for some $C_4 > 0$,

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

and for $C_5 > 0$,

$$\begin{aligned}&\int_0^T \mathbb{E}|\bar{\lambda}_d^n(s,\bar{Q}^n(s)) - \bar{\lambda}_d(s,\bar{Q}^n(s))|ds\\ &= \int_0^T \mathbb{E}\left(|\bar{\lambda}_d^n(s,\bar{Q}^n(s)) - \bar{\lambda}_d(s,\bar{Q}^n(s))|1_{\{\sup_{0\le s\le t}|\bar{Q}^n(s)|\le C_5\}}\right)ds\\ &\quad+\int_0^t \mathbb{E}\left(|\bar{\lambda}_d^n(s,\bar{Q}^n(s)) - \bar{\lambda}_d(s,\bar{Q}^n(s))|1_{\{\sup_{0\le s\le T}|\bar{Q}^n(s)|> C_5\}}\right)ds\\ &\le \int_0^T \sup_{|x|\le C_5}|\bar{\lambda}_d^n(s,x)-\bar{\lambda}_d(s,x)|ds + \int_0^T \mathbb{E}\left(2L_1Y^n(t)1_{\{Y^n(T)>C_4\}}\right)dt\\ &\to \int_0^T 2L_1 y(t)1_{\{y(T)>C_5\}}dt.\end{aligned}$$

To summarize, we have shown that

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

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

7. 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 (QCP) that takes into account (1) inventory holding and perishment cost, (2) backorder 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 (FCP) 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 the controlled production is able to achieve production cost reduction compared to 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 multi-class customers and multi-part 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. We thank Dr. Ping Cao (University of Science and Technology of China) for providing constructive comments.

APPENDIX

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:

$$\min_{\{\lambda_p(t);t\in[0,T]\}} \quad \bar{\mathcal{R}}(\bar{\lambda}_p) \tag{66}$$
$$\text{s.t.} \quad q'(t) = \bar{\lambda}_p(t) - \bar{\lambda}_d(t, q(t)) - \bar{\theta}_p q^+(t) + \bar{\theta}_d q^-(t), \tag{67}$$
$$0 \le \bar{\lambda}_p(t) \le \Lambda_p \tag{68}$$
$$q^+(t) = \max(q(t),\ 0) \tag{69}$$
$$q^-(t) = \max(-q(t),\ 0) \tag{70}$$
$$q(0) = q_0 \in \mathbb{R}, \quad 0 \le t \le T. \tag{71}$$

The objective function (66) and the fluid conservation constraint (67) 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 $[0, T]$, e.g., $T = 24$. 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 Δt, 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 $C_n = C(n\Delta t)$, 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.

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

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

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

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

Instead of formulating constraints (69) and (70) by definition, we consider a new set of constraints combining $q(t)$, $q^+(t)$ and $q^-(t)$:

$$q(t) = q^+(t) - q^-(t), \quad q^+(t) \geq 0, \quad q^-(t) \geq 0, \quad 0 \leq t \leq T \tag{72}$$

The reason that this reformulation is equivalent to constraint (69) and (70) is the following.

Note that the constraint (72) characterizes exactly the same underlying dynamics of the system, i.e., no non-zero $q^+(t)$ 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 (72) 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:

$$\begin{aligned}
&\min_{\lambda_{p,n=1,\dots,N}} \sum_{n=1}^{N} \left((h_p + c_p\bar{\theta}_p) q_n^+ + (h_d + c_d\bar{\theta}_d) q_n^- + C_n \lambda_{p,n} \right) \Delta t + c_f \sum_{n=2}^{N} (Z_n^+ + Z_n^-) \\
&\quad \text{s.t. } \frac{q_n - q_{n-1}}{\Delta t} = \lambda_{p,n} - \lambda_{d,n} - \bar{\theta}_p q_n^+ + \bar{\theta}_d q_n^- \\
&\qquad q_n = q_n^+ - q_n^- \\
&\qquad \lambda_{p,n} - \lambda_{p,n-1} = Z_n^+ - Z_n^- \\
&\qquad 0 \leq q_n^+, \quad 0 \leq q_n^-, \quad 0 \leq \bar{\lambda}_p(t) \leq \Lambda_p \\
&\qquad 0 \leq Z_n^+, \quad 0 \leq Z_n^- \\
&\qquad n = 1, \cdots, N.
\end{aligned} \tag{73}$$

System inputs are T, Δt, h_p, c_p, $\bar{\theta}_p$, h_d, c_d, $\bar{\theta}_d$, c_f, $\lambda_{d,n}$, Λ_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 $n = 1, \dots, N$. 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 pre-specified (equi-distant) time epochs. That is, the time length between the adjacent production rate adjustments is fixed ΔT (e.g., $\Delta T = 8$ *hours*, a typical length of one work shift). We divide the entire time horizon T into $\hat{N} = T/\Delta T$ slots, each having the length of ΔT. Control variables within each interval are set to remain unchanged. Let $K = \Delta T/\Delta t$ 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 minimize total production cost, in order to ensure a desired level of quality of service, one natural practice is to set a fulfillment target. Define β as the minimum fulfillment requirement:

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

For instance, the manager may require 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

[1] P. Afeche, A. Diamant, and J. Milner. Double-sided batch queues with abandonment: Modeling crossing networks. *Operations Research*, 62(5):1179–1201, 2014.

[2] A Korhan Aras, Xinyun Chen, and Yunan Liu. Many-server Gaussian limits for overloaded non-Markovian queues with customer abandonment. *Queueing Systems*, pages 1–45, 2018.

[3] P. Arcidiacono, P. B. Ellickson, P. Landry, and D. B. Ridley. Pharmaceutical followers. *International Journal of Industrial Organization*, 31(5):538 – 553, 2013.

[4] Baris Ata and Sunil Kumar. Heavy traffic analysis of open processing networks with complete resource pooling: Asymptotic optimality of discrete review policies. *Ann. Appl. Probab.*, 15(1A):331–391, 02 2005.

[5] Achal Bassamboo, J. Michael Harrison, and Assaf Zeevi. Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems*, 51(3):249–285, Dec 2005.

[6] Achal Bassamboo and Ramandeep S. Randhawa. On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Research*, 58(5):1398–1413, 2010.

[7] O.J. Boxma, I. David, D. Perry, and W. Stadje. A new look at organ transplantation models and double matching queues. *Probability in the Engineering and Informational Sciences*, 25:135–155, 2011.

[8] Amarjit Budhiraja and Arka Prasanna Ghosh. Diffusion approximations for controlled stochastic networks: An asymptotic bound for the value function. *Ann. Appl. Probab.*, 16(4):1962–2006, 11 2006.

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

[10] Gustav Feichtinger and Richard F Hartl. On the use of Hamiltonian and maximized Hamiltonian in nondifferentiable control theory. *Journal of Optimization Theory and Applications*, 46(4):493–504, 1985.

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

[12] J. Michael Harrison. Brownian models of open processing networks: canonical representation of workload. *Ann. Appl. Probab.*, 10(1):75–103, 02 2000.

[13] Qiao-Chu He, Tiantian Nie, and Zuo-Jun Max Shen. Beyond rebalancing: Crowdsourcing and geo-fencing for shared-mobility systems. *Available at SSRN*, 2018.

[14] H. Kaspi and D. Perry. Inventory systems of perishable commodities. *Adv. Appl. Prob.*, 15:674–685, 1983.

[15] H. Kaspi and D. Perry. Inventory systems of perishable commodities with poisson input and renewal output. *Adv. Appl. Prob.*, 16:402–421, 1984.

[16] Thomas G Kurtz. Strong approximation theorems for density dependent markov chains. *Stochastic Processes and their Applications*, 6(3):223–240, 1978.

[17] X. Liu. Diffusion approximations for double-ended queues with reneging in heavy traffic. *Submitted*, 2017.

[18] Xin Liu, Qi Gong, and Vidyadhar G. Kulkarni. Diffusion models for double-ended queues with renewal arrival processes. *Stoch. Syst.*, 5(1):1–61, 2015.

[19] Yunan Liu and Ward Whitt. A network of time-varying many-server fluid queues with customer abandonment. *Operations Research*, 59(4):835–846, 2011.

[20] Yunan Liu and Ward Whitt. The $G_t/GI/s_t+GI$ many-server fluid queue. *Queueing Systems*, 71(4):405–444, 2012.

[21] Yunan Liu and Ward Whitt. A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Operations Research Letters*, 40(5):307–312, 2012.

[22] Yunan Liu and Ward Whitt. Many-server heavy-traffic limits for queues with time-varying parameters. *Annals of Applied Probability*, 24(1):378–421, 2012.

[23] Yunan Liu and Ward Whitt. Algorithms for time-varying networks of many-server fluid queues. *INFORMS Journal on Computing*, 26(1):59–73, 2014.

[24] Avishai Mandelbaum and Martin I. Massey, William A.and Reiman. Strong approximations for markovian service networks. *Queueing Systems*, 30:149–201, 1998.

[25] Joanna Matula. On an extremum problem. *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, 28(03):376–392, 1987.

[26] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

[27] Jerome Niyirora and Jamol Pender. Optimal staffing in nonstationary service centers with constraints. *Naval Research Logistics (NRL)*, 63(8):615–630, 2016.

[28] Erhun Ozkan and Amy R Ward. Dynamic matching for real-time ridesharing. *Available at SSRN: https://ssrn.com/abstract=2844451*, 2017.

[29] Jamol Pender. Risk measures and their application to staffing nonstationary service systems. *European Journal of Operational Research*, 254(1):113–126, 2016.

[30] D. Perry and W. Stadje. Perishable inventory systems with impatient demands. *Mathematical Methods of Operations Research*, 50(1):77–90, 1999.

[31] Lev Semenovich Pontryagin. *Mathematical theory of optimal processes.* Routledge, 2018.

[32] B. Prabhakar, N. Bambos, and T. S. Mountford. The synchronization of Poisson processes and queueing networks with service and synchronization nodes. *Advances in Applied Probability*, 32(3):824–843, 2000.

[33] Atle Seierstad and Knut Sydsaeter. *Optimal control theory with economic applications.* Elsevier North-Holland, Inc., 1986.

[34] Ward Whitt. Fluid models for multiserver queues with abandonments. *Operations Research*, 54(1):37–54, 2006.

[35] Lawrence Yu. Continuous manufacturing has a strong impact on drug quality. *FDA Voice*, 12, 2016.

[36] S. A. Zenios. Modeling the transplant waiting list: A queueing model with reneging. *Queueing Systems*, 31:239–251, 1999.