Queueing Models and
Service Management
Vol. 6, No. 2, page 29–58, 2023

QMSM
©AU 2023

Asymptotic Optimal Scheduling of V-Systems with Deadlines and Customer Abandonment

Ping Cao[1] Junfei Huang[2] and Jingui Xie[3,*]

[1]School of Management
University of Science and Technology of China, 230026 Hefei, China.
[2]Department of Decisions, Operations and Technology
Chinese University of Hong Kong, Hong Kong, China
[3]School of Management
Technical University of Munich, 74076 Heilbronn, Germany.

(*Received November 2022; accepted June 2023*)

Abstract: We consider a V-structured queueing system with two classes of customers: class 1 customers cannot abandon but have a waiting-time deadline while class 2 customers may abandon. The objective is to minimize the number of abandonments of class 2 customers while meeting the deadline for class 1 customers. We consider the problem in an asymptotic framework, and prove that under the heavy traffic regime, the threshold policy that gives priority to class 1 customers if the age of the head-of-the-line class 1 customer exceeds a threshold is asymptotically optimal.

Keywords: Abandonment, V-structure, waiting deadline.

1. Introduction

The Internet has brought consumers increased access to information to make purchase decisions online. The rapid growth of online transactions in service industries raises new service mode. Many traditional service providers across different categories have offered their customers the "buy-online, pick-up-in-store" (BOPS) option [see 10, 11, and the references therein]. According to Retail Systems Research, as of June 2013, 64% of retailers have implemented BOPS [18]. To make the BOPS option attractive to customers, most service providers announce a deadline commitment after which the product is ready for pick-up.

For example, a customer can place an order online (named *online customers*) and pick up the food in a restaurant later. The restaurant promises that the food will be ready for pick-up 30 minutes later. Online customers are not sensitive to the virtual waiting time, but the deadline commitment. They expect the food to be prepared and packed before the deadline when they arrive at the restaurant and can take it away immediately. Traditional customers (named *offline customers*) may still prefer walking into the restaurant and having food inside.

* Corresponding author
Email: jingui.xie@tum.de

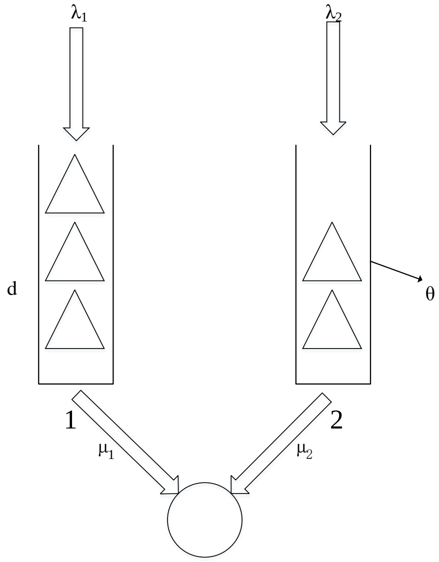

Figure 1. A V-queueing model

They are delay sensitive because they have to wait in queue. Customer with limited patience will abandon the queue if his patience ran out.

The goal of the service provider is to minimize the number of abandoned offline customers and guarantee the deadline commitment for online orders. The control decision faced by the system manager is the routing of the jobs: should the agent fulfill an online order or serve an offline customer after each service completion.

To tackle this problem, we model the system as a queueing model with two classes of customers. See Figure 1 for illustration. Since the waiting time is stochastic, the problem is not feasible for exact analysis. Therefore, we resort to heavy traffic approximation which has been extensively used in call centers. We propose a threshold policy and prove that the policy is asymptotically optimal in heavy traffic.

Our motivation is related to service systems with both online and offline demands, especially on “buy online and pick up offline”. [1] provide a comprehensive review on retailers with multiple channels. They address supply chain management issues and corresponding quantitative models specific to internet fulfillment in a multi-channel environment. [13] investigate airline ticket sales in both online and offline channels. Service systems can improve their performance substantially by offering multiple channels of service and hence managing the service for multiple channels becomes a critical operation. [9] analyze empirically the impact of implementing a BOPS project and discuss the integration of online and offline channels in retail. Different from their works, we consider a service provider rather than retailers. Retailers sell products and the inventory issue is relevant, while we focus on the service process.

Our model is related to queues with waiting deadlines and customer abandonments. There are several queueing models considering either service deadline or customer abandonments. For example, [14] study a multiclass queue with service deadline, and apply the queueing model to patient flow control in emergency departments. [12] study an N-queueing model with two classes of customers with possible abandonments. However, few works consider the trade-off between deadline and abandonment as we did in our paper.

Our queueing model is also related to queueing models with two types of customers where one type customers form a physical queue, while the other type customers form a virtual queue. For example in [15], a service provider offers customers the choice of either waiting in a line or going around and returning at a determined future time. But some customers with the later choice may not return for service at their appointed time. The authors discuss how to allocate service capacity between the two lines. Different from [15], we consider customers' abandonment in the physical queue ranter than the virtual queue. In our setting, the online customers have paid first and will pick up offline, they are unlikely to abandon. Moreover, [15] considers a static service allocation rule while our model considers dynamic service rule. Another related model is call center with call-back option [see 2, 3]. They consider a call center with two channels, one for real-time telephone service, and another for a postponed call-back service offered with a guarantee on the maximum delay until a reply is received. Different from their works, our model explicitly takes customer abandonments into consideration.

Our contributions are twofold: practically, we propose a queueing model to analyze service processes with both online and offline demand and try to balance between deadline commitment and customer abandonment; theoretically, we derive the asymptotic pathwise optimality of a threshold policy for the above queueing system. Note that this pathwise optimality is generally not true for systems with customer abandonment. As a result, we believe the pathwise optimality is itself interesting.

The remainder of this paper is organized as follows. Section 2 describes the queueing model and its heavy traffic framework. Section 3 proposes a threshold policy, and proves its asymptotical optimality. In Section 4, we test the proposed policy by numerical examples. Section 5 concludes the paper.

2. Model Formulation

2.1. Basic model

Consider a V-queueing model with a single server and two classes of customers. For $k = 1, 2$, class k customers arrive according to a renewal process $A_k = \{A_k(t), t \geq 0\}$, and are served based on the FCFS principle. The renewal process $S_k = \{S_k(t), t \geq 0\}$ represents the number of class k customers that can be served if the server works continuously and exclusively on class k customers in $[0, t]$. To be specific, for $k = 1, 2$,

$$A_k(t) = \max\left\{n \in \mathbb{N} : \sum_{i=1}^{n} u_k(i) \leq \lambda_k t\right\},$$
$$S_k(t) = \max\left\{n \in \mathbb{N} : \sum_{i=1}^{n} v_k(i) \leq \mu_k t\right\},$$

where $\{u_k(i), i \in \mathbb{N}\}$ are strictly positive i.i.d. random variables with mean 1 and variance $\alpha_k^2 \in [0, \infty)$, and $\{v_k(i), i \in \mathbb{N}\}$ are strictly positive i.i.d. random variables with mean 1 and variance $\beta_k^2 \in [0, \infty)$, λ_k and μ_k are the arrival rate and service rate of class k customers, respectively.

We assume that class 1 customers cannot abandon the system, while each class 2 customer will independently abandon the system if her/his waiting time exceeds her/his patience time, which is exponentially distributed with mean θ^{-1}.

A control policy is defined as $\pi = \{T_k, k = 1, 2\}$, where $T_k(t)$ is the cumulative amount of service time devoted to serving class k customers till time t. Then, $I(t) = t - T_1(t) - T_2(t)$ is the cumulative idle time of the server till time t. The number of class k customers at time t, denoted by $Q_k(t)$, satisfies the following equations:

$$Q_1(t) = Q_1(0) + A_1(t) - S_1(T_1(t)) \geq 0, \tag{1}$$
$$Q_2(t) = Q_2(0) + A_2(t) - S_2(T_2(t)) - R(t) \geq 0, \tag{2}$$

where

$$R(t) = N\left(\theta \int_0^t Q_2(s)ds\right) \tag{3}$$

is the cumulative number of class 2 customers who have abandoned till time t, N is a unit-rate Poisson process.

Remark 1. Here we take the same formulation of the abandonment process R as that in [19], which allows the customers in service to abandon, for convenience of notation. However, our results still hold when the customer in service is not allowed to abandon, i.e.,

$$R(t) = N\left(\theta \int_0^t (Q_2(s) - \mathbf{1}\{\dot{T}_2(s) > 0\})ds\right).$$

See, e.g., [12]. Here, **1** is the indicator function, $\dot{T}_2(t)$ is the right derivative of T_2 at time $t \geq 0$, and thus $\dot{T}_2(t) > 0$ denotes that a class 2 customer is being served at time t.

We define the workload of the system at time t to be

$$W(t) = \frac{Q_1(t)}{\mu_1} + \frac{Q_2(t)}{\mu_2}.$$

An *admissible* control policy π must satisfy (1)–(3) for all $t \geq 0$, and additionally, for $k = 1, 2,$

π is nonanticipating with respect to the queue length process $Q = (Q_1, Q_2)$,

T_k is continuous and nondecreasing with $T_k(0) = 0$,

I is continuous and nondecreasing with $I(0) = 0$.

We assume that each class 1 customer has a waiting time deadline d, that is, if we denote by $\tau_1(t)$ the "age" in the system of the head-of-the-line class 1 customer at time t, then we require $\tau_1(t) \leq d$ for all $t \geq 0$.

Let Π be the set of all admissible policies. The system manager's objective is to find an admissible control policy $\pi \in \Pi$ to minimize the expected cumulative abandonment number of class 2 customers meanwhile each class 1 customer receives service within d time, i.e., for any $T \geq 0$,

$$\begin{aligned} \min_{\pi \in \Pi} \quad & \mathbb{E}R(T) \\ \text{s.t.} \quad & \tau_1(t) \leq d, \quad 0 \leq t \leq T. \end{aligned} \tag{4}$$

Remark 2. Note that the total expected arrival of class 2 customers is fixed. Hence, the above problem is equivalent to minimizing the probability of abandonment.

However, the above problem is infeasible as the age process $\tau_1 = \{\tau_1(t), t \geq 0\}$ is stochastic. Thus, we will solve the above problem in an asymptotic framework, and will show the asymptotic optimality of a threshold hold policy which has a simple structure to use.

2.2. Heavy-traffic assumption

We consider a sequence of systems indexed by n. These systems all have the same structure as that described in the last section; however, relevant parameters and processes (except Poisson process N) may vary with n. We indicate the dependence of relevant parameters and processes on n by appending a superscript to them. Then, in the nth system, the arrival rate of class k customers is λ_k^n, the age deadline for class 1 customers is d^n, and the abandonment rate of class 2 customers is θ^n. We assume that the service rates are invariant with respect to n; hence there will be no superscript for terms relating to service rates.

Let $Q^n(t) = (Q_1^n(t), Q_2^n(t))$. We introduce the following diffusion scaled processes

$$\widehat{Q}^n(t) = \frac{Q^n(nt)}{\sqrt{n}}, \; \widehat{R}^n(t) = \frac{R^n(nt)}{\sqrt{n}},$$
$$\widehat{\tau}_1^n(t) = \frac{\tau_1^n(nt)}{\sqrt{n}}, \; \widehat{W}^n(t) = \frac{W^n(nt)}{\sqrt{n}}.$$

We assume that the following *heavy traffic* condition holds.

Assumption 1. *There exist strictly positive constants* λ_1, λ_2 *and constant* β *such that*

$$\lambda_k^n \to \lambda_k, \; k = 1, 2, \; \sqrt{n}\left(\frac{\lambda_1^n}{\mu_1} + \frac{\lambda_2^n}{\mu_2} - 1\right) \to \beta, \; \textit{as } n \to \infty.$$

We put the following assumption on the abandonment rate.

Assumption 2. *There exists a strictly positive constant θ such that $n\theta^n \to \theta$ as $n \to \infty$.*

Moreover, since under Assumption 1, the age of the head-of-the-line class 1 customer also should be of order $\sqrt{n}$. Hence, we take the following assumption on the age deadline of class 1 customers.

Assumption 3. *There exists a strictly positive constant d such that $d^n/\sqrt{n} \to d$, as $n \to \infty$.*

Finally, we assume that the following initial condition holds.

Assumption 4. *There exists a positive finite random variable $\widehat{W}_0$ such that $\widehat{W}^n(0) \Rightarrow \widehat{W}_0$ as $n \to \infty$, where $\Rightarrow$ denotes convergence in distribution.*

2.3. Asymptotic compliance and optimality

We will consider policies that are asymptotically compliant, which is a generalization of "feasibility". See e.g., [16].

Definition 1 (Asymptotic compliance)**.** A family of policies $\{\pi^n, n \in \mathbb{N}\}$ is called *asymptotic compliant* if, for any fixed $T \geq 0$,

$$\sup_{0\leq t\leq T}\left[\widehat{\tau}_1^n(t) - \frac{d^n}{\sqrt{n}}\right]^+ \Rightarrow 0, \text{ as } n \to \infty. \tag{5}$$

Definition 2 (Asymptotic optimality)**.** A family of control policies $\{\pi_*^n, n \in \mathbb{N}\}$ is *asymptotically optimal* if

1. It is asymptotically compliant and
2. For every $t > 0$ and every $x > 0$,

$$\limsup_{n\to\infty} \mathbb{P}\left\{\widehat{R}_*^n(t) > x\right\} \leq \liminf_{n\to\infty} \mathbb{P}\left\{\widehat{R}^n(t) > x\right\},$$

where $\widehat{R}_*^n(t)$ and $\widehat{R}^n(t)$ are the diffusion-scaled cumulative number of abandonments till time t under the family of control policies $\{\pi_*^n, n \in \mathbb{N}\}$ and any other asymptotically compliant family of policies $\{\pi^n, n \in \mathbb{N}\}$, respectively.

Remark 3. Note that a stochastic larger random variable has a larger expectation. Hence, here we consider a stronger criterion than that in (4) in fact.

3. The Proposed Policy

We propose the following family of work-conserving scheduling policy, which is denoted by $\{\pi_{th}^n, n \in \mathbb{N}\}$: Fix a sequence of $\{\epsilon^n, n \in \mathbb{N}\}$ such that $\epsilon^n/\sqrt{n} \to 0$ as $n \to \infty$. When becoming idle, the server uses a threshold rule to determine which class to serve next as follows:

1. If $\tau_1^n(t) \geq d^n - \epsilon^n$, give priority to class 1 customers.
2. Otherwise, give priority to class 2 customers.

Our main result is the following theorem.

Theorem 1. *Under Assumptions 1–4, the family of proposed policies $\{\pi^n_{th}, n \in \mathbb{N}\}$ is asymptotically optimal.*

Based on the above theorem, we suggest the following scheduling policy, denoted by π_{th}, for the original system: when becoming idle, the server uses a threshold rule to determine which class to serve next:

1. If $\tau_1(t) \geq d - \epsilon$, give priority to class 1 customers, where ϵ is small relative to d.
2. Otherwise, give priority to class 2 customers.

The proof of Theorem 1 takes two steps. First in Theorem 2 of Section 3.1, we prove that under any asymptotically "feasible" policy, diffusion-scaled cumulative abandonment numbers can be stochastically bounded from below by constructing an alternative system. Then, in Section 3.2 we show that, under the proposed policy, the lower bound can be achieved by establishing a state space collapse (SSC) result.

3.1. Lower bound of any asymptotic compliant policy

Theorem 2 (Lower bound). *Fix any asymptotically compliant family of policies $\{\pi^n, n \in \mathbb{N}\}$. For any $T, x > 0$,*

$$\begin{aligned}&\liminf_{n\to\infty} \mathbb{P}\{\widehat{R}^n(T) > x\}\\ &\geq \mathbb{P}\left\{\theta\mu_2 \int_0^T \left(\widehat{W}(t) - \frac{\lambda_1 d}{\mu_1}\right)^+ dt > x\right\},\end{aligned} \tag{6}$$

where $\widehat{W}$ is the unique solution to

$$\begin{aligned}&\widehat{W}(t) = \widehat{W}_0 + \widehat{X}(t) - \theta \int_0^t \left(\widehat{W}(s) - \frac{\lambda_1 d}{\mu_1}\right)^+ ds + \widehat{I}(t) \geq 0,\\ &\widehat{I}(t) \text{ is nondecreasing in } t, \widehat{I}(0) = 0,\\ &\int_0^\infty \mathbf{1}\{\widehat{W}(t) > 0\} d\widehat{I}(t) = 0.\end{aligned} \tag{7}$$

Here $\widehat{X}$ is a one-dimensional Brownian motion starting from the origin with drift rate β and variance

$$\sigma^2 := \frac{\lambda_1\alpha_1^2}{\mu_1^2} + \frac{\lambda_2\alpha_2^2}{\mu_2^2} + \frac{\lambda_1^2\beta_1^2}{\mu_1^3} + \frac{\lambda_2^2\beta_2^2}{\mu_2^3}.$$

The proof of Theorem 2 takes three steps. First, we claim that we only need to consider the work-conserving asymptotically compliant family of policies. Second, we construct an alternative system and a new policy so that the cumulative abandonment number in the alternative system is pathwise not less than that in the original system. Finally, we characterize the diffusion limit result of the alternative system, which provides a lower bound of the original system.

3.2. Asymptotic optimality of the proposed policy

For any $w \geq 0$, define

$$Q^*(w) = (\mu_1 \min(w, \lambda_1 d/\mu_1), \mu_2(w - \lambda_1 d/\mu_1)^+). \tag{8}$$

First, we have the following SSC result.

Proposition 1. *Assume that*

$$\widehat{Q}^n(0) \Rightarrow Q^*(\widehat{W}(0)), \text{ as } n \to \infty. \tag{9}$$

and Assumptions 1–4 hold. Then, under the family of proposed control policies $\{\pi_{th}^n, n \in \mathbb{N}\}$*, for any* $T > 0$*,*

$$\sup_{0 \leq t \leq T} |\widehat{Q}^n(t) - Q^*(\widehat{W}^n(t))| \Rightarrow 0, \text{ as } n \to \infty.$$

Thus, we have the following convergence result of the diffusion scaled processes.

Theorem 3. *Under the family of proposed policies* $\{\pi_{th}^n, n \in \mathbb{N}\}$*,*

$$(\widehat{W}^n, \widehat{Q}_1^n, \widehat{Q}_2^n) \Rightarrow (\widehat{W}, \widehat{Q}_1, \widehat{Q}_2), \text{ as } n \to \infty,$$

where $\widehat{W}(t)$ *is defined in* (7)*, and*

$$\widehat{Q}_1(t) = \mu_1 \min\left(\widehat{W}(t), \frac{\lambda_1 d}{\mu_1}\right), \; \widehat{Q}_2(t) = \mu_2 \left(\widehat{W}(t) - \frac{\lambda_1 d}{\mu_1}\right)^+.$$

Combining the results in Theorems 2 and 3, we can prove that the family of proposed policies $\{\pi_{th}^n, n \in \mathbb{N}\}$ is asymptotically optimal by verifying that it is asymptotically compliant and attains the lower bound stated in Theorem 2.

4. Numerical Study

The proposed policy is proved to be asymptotically optimal with a single server under heavy traffic regime. But we are not sure how well it will be comparing to other classical policies under more general settings. This numerical study is designed for this purpose.

Our base case is case 1 as shown in Table 2. The total arrival rate $\lambda = \lambda_1 + \lambda_2 = 60$, which means the average number of online and offline orders is 60 per hour. The average processing times of online and offline orders are equal, i.e., 6 minutes per order. The number of service agents in the system is 6. The deadline commitment for online orders is 30 minutes, and the average patience time for customers waiting in the physical queue is one hour.

Case 1 serves as the benchmark of the numerical study. We design other numerical examples and compare them with the base case to see the impact of different parameters on system performance under the optimal policy. Our optimal policy suggest the threshold to be $d - \epsilon$, where ϵ is of a smaller order than d, e.g., $\epsilon = d/10$. So, in this numerical study, ϵ is 3 minutes if d is 30 minutes. Please refer to Section 7.2 of [14] for more discussion on choosing ϵ.

Table 1. Notation Explanation

Notation	Explanation
λ_1	Average number of online orders per hour
λ_2	Average number of offline orders per hour
μ_1	Average process capacity for online orders per hour
μ_2	Average process capacity for offline orders per hour
N	Number of service agents
d	Deadline commitment for online orders (hr)
θ	Average patience time for waiting customers (hr)
P1	Work-conserving policy that gives priority to class 1
P2	Work-conserving policy that gives priority to class 2
Opt	Our proposed policy

Table 2. Parameter Setting

Notation	Case 1	Case 2	Case 3	Case 4	Case 5
λ_1	15	30	45	15	30
λ_2	45	30	15	45	90
μ_1	10	10	10	10	20
μ_2	10	10	10	10	20
N	6	6	6	6	6
d	.5	.5	.5	.5	.5
θ	1	1	1	3	1

4.1. λ_1/λ_2: proportion of online orders

Service providers may have different proportions of online orders. To study this impact, we set λ_1/λ_2 as 3, 1 and 1/3. When $\lambda_1 = 3\lambda_2$, it means that the service provider has a large proportion of online orders. When $\lambda_1 = \lambda_2$, it means that the service provider has an equal proportion of online and offline orders. When $\lambda_1 = \lambda_2/3$, it means that the service provider has a large proportion of offline orders.

From the numerical results, we can see that when the proportion of the online orders becomes large, the reduction on abandonment (comparing policy P1 and the optimal policy) becomes significant (from 2%,7% to 13%), but the percentage of orders meeting the deadline drops (from 96%, 93% to 83%). Though P2 minimizes the abandonment rate, the percentage of orders meeting the deadline is less than half, which may not be acceptable by online customers.

4.2. θ: patient and impatient customers

Let us change the average patience time for waiting customers θ from 1 hour to 20 minutes, which means the customers waiting in the physical queue is relatively impatient. Comparing case 1 and 4, the abandonment rate under each policy increases. The optimal policy guarantees 98% deadline commitment with 11% abandonment, saving 4% abandonment comparing to P1 policy. From the comparison of the numerical results, we can see that the optimal policy saves a lot when the customers are very impatient.

Table 3. Numerical examples

Case	λ_1/λ	**Policy**	$\mathbb{E}(W_1)$	$\mathbb{E}(W_2)$	$\mathbb{P}(W_1 < D)$	P_{Ab}	**Utilization**
1	0.25	P1	0.02	0.10	100%	10%	92%
	0.25	P2	1.02	0.05	41%	5%	97%
	0.25	Opt	0.27	0.08	96%	8%	95%
2	0.5	P1	0.03	0.15	100%	15%	93%
	0.5	P2	1.11	0.03	34%	3%	98%
	0.5	Opt	0.28	0.08	93%	8%	96%
3	0.75	P1	0.05	0.26	100%	26%	93%
	0.75	P2	1.07	0.02	41%	2%	98%
	0.75	Opt	0.31	0.13	83%	13%	97%
4	0.25	P1	0.02	0.05	100%	15%	89%
	0.25	P2	0.36	0.03	75%	9%	93%
	0.25	Opt	0.19	0.04	98%	11%	91%
5	0.25	P1	0.01	0.08	100%	8%	94%
	0.25	P2	0.92	0.03	47%	3%	97%
	0.25	Opt	0.28	0.05	100%	5%	97%

4.3. μ: process-and-pack v.s. pick-and-pack service

For products which require processing before packing, the service time is longer. There are some other products which the service agents can simply pick before packing, the service time is relatively shorter. In case 5, we assume the average service time is 3 minutes, and the arrival rate is 120 orders per hour. Let $\mu_1 = \mu_2 = 20$, and $\lambda = 120$. We see from the numerical results that the optimal policy can guarantee the deadline commitment while minimizing the abandonment rate (from 8% to 5% comparing to P1 policy).

5. Conclusion

We study a V model with two classes of customers in which class 1 customers cannot abandon but have waiting time deadlines while class 2 customers may abandon, with the objective of minimizing the cumulative number of class 2 customers who have abandoned while ensuring each class 1 customer receives the service within the waiting time deadline. We prove that under heavy traffic regime, the threshold policy that gives priority to class 1 customers if the age of the head-of-the-line class 1 customer exceeds a threshold is asymptotically optimal.

There are several directions worthy of further researching. Firstly, breaching a deadline commitment can be quantified by a cost, with the goal being to minimize cost or maximize profit. Furthermore, the deadline commitment may be random or dependent on the service requirement. Second, as mentioned in the introduction, our problem admits a pathwise optimal solution for the diffusion approximation. However, with other performance criterion, there might not exist a pathwise optimal solution. Hence, two questions arise: under what criterion a pathwise optimal solution is admitted? And if there is no pathwise optimal solution, is a threshold policy still asymptotically optimal? Third, we can generalize the V-queueing model to other queueing model with more general structure. Finally, in this paper, the waiting time deadline is given. We might consider the problem of how to choose the proper waiting time deadline to maximize the profit.

Acknowledgements

We express our gratitude to Prof. George Zhang, the Editors-in-Chief, the associate editor, and the reviewers for their insightful feedback and comments. The work of P. Cao is financially supported by the National Natural Science Foundation of China [Grants 72122019 and 71771202]. J. Huang's research is supported in part by the Hong Kong Research Grants Council [Grants 14501621 and 14505820] and the National Natural Science Foundation of China [Grant 72222023].

References

[1] Agatz, N. A., Fleischmann, M., & Van Nunen, J. A. (2008). E-fulfillment and multi-channel distribution: A review. *European Journal of Operational Research*, 187(2), 339–356.

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

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

[4] Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley, 2nd edition.

[5] Bramson, M. (1998). State Space Collapse with Application to Heavy Traffic Limits for Multiclass Queueing Networks. *Queueing Systems*, 30, 89–148.

[6] Chen, H., & Ye, H.-Q. (2012). Asymptotic Optimality of Balanced Routing. *Operations Research*, 60(1), 163–179.

[7] Dai, J. (1995). On Positive Harris Recurrence of Multiclass Queueing Networks: a Unified Approach Via Fluid Limit Models. *The Annals of Applied Probability*, 5(1), 49–77.

[8] Eschenfeldt, P., & Gamarnik, D. (2018). Join the shortest queue with many servers. The heavy-traffic asymptotics. *Mathematics of Operations Research*, 43(3), 867–886.

[9] Gallino, S., & Moreno, A. (2014). Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Science*, 60(6), 1434–1451.

[10] Gao, F., & Su, X. (2017). Omnichannel retail operations with buy-online-and-pick-up-in-store. *Management Science*, 63(8), 2478–2492.

[11] Gao, F., & Su, X. (2017). Online and offline information for omnichannel retailing. *Manufacturing & Service Operations Management*, 19(1), 84–98.

[12] Ghamami, S., & Ward, A. (2013). Dynamic scheduling of a two-server parallel server system with complete resource pooling and reneging in heavy traffic: Asymptotic optimality of a two-threshold policy. *Mathematics of Operations Research*, 38, 761–824.

[13] Granados, N., Gupta, A., & Kauffman, R. J. (2012). Online and offline demand and price elasticities: Evidence from the air travel industry. *Information Systems Research*, 23(1), 164–181.

[14] Huang, J., Carmeli, B., & Mandelbaum, A. (2015). Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4), 892–908.

[15] Kostami, V., & Ward, A. R. (2009). Managing service systems with an offline waiting option and customer abandonment. *Manufacturing & Service Operations Management*, 11(4), 644–656.

[16] Plambeck, E., Kumar, S., & Harrison, J. (2001). A Multiclass Queue in Heavy Traffic with Throughput Time Constraints: Asymptotically Optimal Dynamic Controls. *Queueing Systems*, 39, 23–54.

[17] Reed, J., & Ward, A. R. (2004). A diffusion approximation for a generalized Jackson network with reneging. In Skrikat, R., & Voulgaris, G., editors, *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing*. Monticello, IL.

[18] Rosenblum, P., & Kilcourse, B. (2013). Omni-Channel 2013: The Long Road To Adoption. *Benchmark Report, RSR Research, Miami*.

[19] Ward, A. R., & Kumar, S. (2008). Asymptotically Optimal Admission Control of a Queue with Impatient Customers. *Mathematics of Operations Research*, 33(1), 167–202.

[20] Whitt, W. (2002). *Stochastic-Process Limits*. Springer, New York.

[21] Ye, H.-Q., & Yao, D. D. (2008). Heavy-Traffic Optimality of a Stochastic Network Under Utility-Maximizing Resource Allocation. *Operations Research*, 56(2), 453–470.

Appendix

A. Preliminary Analysis

We introduce the following diffusion scaled processes

$$\widehat{A}_k^n(t) = \frac{A_k^n(nt) - \lambda_k^n nt}{\sqrt{n}},\ \widehat{S}_k^n(t) = \frac{S_k^n(nt) - \mu_k nt}{\sqrt{n}},\ k = 1, 2,$$
$$\widehat{N}^n(t) = \frac{N(nt) - nt}{\sqrt{n}},\ \widehat{I}^n(t) = \frac{I^n(nt)}{\sqrt{n}},$$

and fluid scaled processes

$$\bar{Q}^n(t) = \frac{Q^n(nt)}{n},\ \bar{A}_k^n(t) = \frac{A_k^n(nt)}{n},$$
$$\bar{T}_k^n(t) = \frac{T_k^n(nt)}{n},\quad k = 1, 2,$$
$$\bar{R}^n(t) = \frac{R^n(nt)}{n},\ \bar{I}^n(t) = \frac{I^n(nt)}{n},\ \bar{W}^n(t) = \frac{W^n(nt)}{n}.$$

It follows from the renewal process Functional Central Limit theorem [see e.g., Corollary 7.3.1 in 20] that

$$(\widehat{A}_k^n, \widehat{S}_k^n, \widehat{N}^n, k = 1, 2) \Rightarrow (\widehat{A}_k, \widehat{S}_k, \widehat{N}, k = 1, 2), \text{ as } n \to \infty, \tag{10}$$

where $\widehat{A}_k$, $\widehat{S}_k$, $\widehat{N}$, $k = 1, 2$ are mutually independent; $\widehat{A}_k$ is a one-dimensional Brownian motion that starts from the origin and has variance $\lambda_k \alpha_k^2$; $\widehat{S}_k$ is a one-dimensional Brownian motion that starts from the origin and has variance $\mu_k \beta_k^2$; $\widehat{N}$ is a standard one-dimensional Brownian motion.

The following result relates the age process of the head-of-line class 1 customers to the queue length process of class 1 customers on the diffusion scale.

Lemma 1. *Under any asymptotically compliant family of control policies, and for any $T > 0$,*

$$\sup_{0 \le t \le T} \left| \widehat{Q}_1^n(t) - \lambda_1^n \widehat{\tau}_1^n(t) \right| \Rightarrow 0, \quad \textit{as} \quad n \to \infty.$$

Corollary 1. *Under any asymptotically compliant family of control policies,*

$$\sup_{0 \le t \le T} \left[\widehat{Q}_1^n(t) - \lambda_1 d \right]^+ \Rightarrow 0, \quad \textit{as} \quad n \to \infty.$$

The proofs of the above lemma and corollary are exactly the same to those of Lemma EC.1.1 and Corollary 2 in [14], which are omitted for brevity.

The following result demonstrates that the result of Lemma 1 is also valid under any work-conserving policy.

Proposition 2. *Under any family of work-conserving policies, and for any $T > 0$,*

$$\sup_{0\le t\le T}\left|\widehat{Q}_1^n(t) - \lambda_1^n\widehat{\tau}_1^n(t)\right| \Rightarrow 0, \qquad as \quad n \to \infty. \tag{11}$$

The proof of Proposition 2 consists of three steps. First, in Lemma 2, we show that under any work-conserving policy, the family $\{\bar{Q}_k^n, \bar{T}_k^n, \bar{A}_k^n, \bar{R}^n, \bar{W}^n, \bar{I}^n, k = 1, 2\}$ is precompact, which has weak convergence limit. Second, in Lemma 3 we characterize the fluid limit of $(\bar{Q}_k^n, \bar{T}_k^n)$ under Assumption 4, by showing that the weak convergence limit is unique. Finally, we show (11) by using an inequality relating $Q_1(t)$ to $\tau_1(t)$ (see (22) below).

Lemma 2. *Suppose there exists a constant M such that $\bar{W}^n(0) \le M$ for all $n \in \mathbb{N}$. Under any work-conserving policy, the family $\{\bar{Q}_k^n, \bar{T}_k^n, \bar{A}_k^n, \bar{R}^n, \bar{W}^n, \bar{I}^n, k = 1, 2\}$ is C-tight, i.e., for any subsequence of $\{n\}$, there exists a further subsequence, denoted by $\mathcal{N}$, such that along $\mathcal{N}$,*

$$(\bar{Q}_k^n, \bar{T}_k^n, \bar{A}_k^n, \bar{R}^n, \bar{W}^n, \bar{I}^n, k = 1, 2)$$

converge uniformly on compact time sets (u.o.c.) to limit process

$$(\bar{Q}_k, \bar{T}_k, \bar{A}_k, \bar{R}, \bar{W}, \bar{I}, k = 1, 2),$$

which has continuous paths almost surely and satisfies the following equations:

$$\bar{A}_k(t) = \lambda_k t, k = 1, 2, \tag{12}$$

$$\bar{Q}_1(t) = \bar{Q}_1(0) + \bar{A}_1(t) - \mu_1\bar{T}_1(t), \tag{13}$$

$$\bar{Q}_2(t) = \bar{Q}_2(0) + \bar{A}_2(t) - \mu_2\bar{T}_2(t) - \bar{R}(t), \tag{14}$$

$$\bar{R}(t) = \theta\int_0^t \bar{Q}_2(s)ds, \tag{15}$$

$$\bar{W}(t) = \frac{\bar{Q}_1(t)}{\mu_1} + \frac{\bar{Q}_2(t)}{\mu_2}, \tag{16}$$

$$\bar{I}(t) = t - \bar{T}_1(t) - \bar{T}_2(t), \tag{17}$$

$$\int_0^\infty \mathbf{1}\{\bar{W}(t) > 0\}d\bar{I}(t) = 0. \tag{18}$$

Proof. Proof.The functional strong law of large numbers (FSLLN) implies

$$\begin{aligned} &\bar{A}_k^n(t) = \frac{A_k^n(nt)}{n} \to \bar{A}_k(t) := \lambda_k t,\\ &\frac{S_k(nt)}{n} \to \mu_k t, \quad \frac{N(nt)}{n} \to t, \end{aligned} \tag{19}$$

u.o.c., as $n \to \infty$.

Define $Y^n(t) = \int_0^t \bar{Q}_2^n(s)ds$. Then, we have

$$\bar{R}^n(t) \quad = \quad \frac{R^n(nt)}{n} = \frac{N\left(\theta^n\int_0^{nt} Q_2^n(s)ds\right)}{n}$$

$$= \frac{N\left(n\theta^n \int_0^t Q_2^n(ns)ds\right)}{n} = \frac{N\left(n^2\theta^n \int_0^t \bar{Q}_2^n(s)ds\right)}{n} = \frac{N\left(n^2\theta^n Y^n(t)\right)}{n}.$$

For any $0 \le t_1 < t_2$, $0 \le \bar{T}_k^n(t_2) - \bar{T}_k^n(t_1) \le t_2 - t_1$, which implies that $0 \le \bar{T}_k(t_2) - \bar{T}_k(t_1) \le t_2 - t_1$. Hence, it follows from Theorem 15.1 in [4] that $\{\bar{T}_k^n\}$ is C-tight. Moreover, for any $0 \le t_1 < t_2 \le 1$, we have

$$0 \le Y^n(t_2) - Y^n(t_1) = \int_{t_1}^{t_2} \check{Q}_2^n(s)ds \le (t_2 - t_1)(\check{Q}_2^n(0) + \check{A}_2^n(1)).$$

Again, it follows from Theorem 15.1 in [4] with a similar argument as that in the proof of Lemma 4.3 in [19] that $\{Y^n\}$ is C-tight.

Let $\mathcal{N}_1$ be any subsequence of $\{n\}$. Since $\bar{W}^n(0) \le M$ and $\{\bar{T}_k^n, Y^n\}$ is C-tight, we can find a subsequence $\mathcal{N}$ of $\mathcal{N}_1$, such that along $\mathcal{N}$,

$$\bar{Q}_k^n(0) \to \bar{Q}_k(0), \quad \bar{T}_k^n(t) \to \bar{T}_k(t), \quad Y^n(t) \to Y(t), \text{u.o.c.} \tag{20}$$

Hence, it follows from (19) and Assumption 2 that

$$\bar{R}^n(t) = \frac{N\left(n^2\theta^n Y^n(t)\right)}{n} \to \theta \int_0^t \bar{Q}_2(s)ds,$$

u.o.c., along $\mathcal{N}$.

It follows from

$$\begin{aligned}\bar{Q}_1^n(t) &= \bar{Q}_1^n(0) + \bar{A}_1^n(t) - \frac{S_1(n\bar{T}_1^n(t))}{n},\\ \bar{Q}_2^n(t) &= \bar{Q}_2^n(0) + \bar{A}_2^n(t) - \frac{S_2(n\bar{T}_2^n(t))}{n} - \bar{R}^n(t),\end{aligned}$$

(19), and (20) that that

$$\begin{aligned}\bar{Q}_1^n(t) &\to \bar{Q}_1(t) := \bar{Q}_1(0) + \bar{A}_1(t) - \mu_1\bar{T}_1(t),\\ \bar{Q}_2^n(t) &\to \bar{Q}_2(t) := \bar{Q}_2(0) + \bar{A}_2(t) - \mu_2\bar{T}_2(t) - \bar{R}(t),\end{aligned}$$

u.o.c., along $\mathcal{N}$.

It follows from the definition of $W(t)$ and $I(t)$ that

$$\bar{W}^n(t) \to \bar{W}(t) := \frac{\bar{Q}_1(t)}{\mu_1} + \frac{\bar{Q}_2(t)}{\mu_2}, \quad \bar{I}^n(t) \to \bar{I}(t) := t - \bar{I}_1(t) - \bar{I}_2(t)$$

u.o.c., along $\mathcal{N}$.

To prove (18), it suffices to show that given any interval $[t_1, t_2]$, if $\bar{W}(t) > 0$ for all $t \in [t_1, t_2]$, then $\bar{I}(t_2) - \bar{I}(t_1) = 0$. Note that $\bar{W}^n(t) > 0$ also holds for $t \in [t_1, t_2]$ (or

$W^n(t) > 0$ for all $t \in [nt_1, nt_2]$) when n is sufficiently large, because $\bar{W}^n(t) \to \bar{W}(t)$ u.o.c. Since under any family of work-conserving policies, it holds that

$$\int_0^\infty \mathbf{1}\{W^n(t) > 0\} dI^n(t) = 0,$$

we have $I^n(nt_2) - I^n(nt_1) = 0$, or $\bar{I}^n(t_2) - \bar{I}^n(t_1) = 0$. Letting $n \to \infty$ yields that $\bar{I}(t_2) - \bar{I}(t_1) = 0$.

Lemma 3. *Suppose that Assumption 4 holds. Then, we have*

$$(\bar{Q}^n, \bar{T}^n) \to (0, \bar{T}^*), \text{ as } n \to \infty, \tag{21}$$

where

$$\bar{T}^*(t) = \left(\frac{\lambda_1}{\mu_1} t, \frac{\lambda_2}{\mu_2} t\right).$$

Proof. Assumption 4 implies $\bar{W}^n(0)$ is stochastically bounded and thus the result in Lemma 2 holds for each sample path almost surely. Now fix any sample path such that the result in Lemma 2 holds. Then, $\bar{W}(0) = 0$ in view of Assumption 4. Moreover, (12)-(14) and Assumption 1 imply that

$$\bar{W}(t) = \bar{W}(0) + t - (\bar{T}_1(t) + \bar{T}_2(t)) - \frac{\theta}{\mu_2} \int_0^t \bar{Q}_2(s) ds = \bar{I}(t) - \frac{\theta}{\mu_2} \int_0^t \bar{Q}_2(s) ds.$$

Next, we show that $\bar{W}(t) = 0$ for all $t \geq 0$. Otherwise, by the Lipschitz continuity of $\bar{W}(t)$, there exist time points $t_1 < t_2$ such that $\bar{W}(t_1) = 0$ and $\bar{W}(t) > 0$ for all $t \in (t_1, t_2]$. Hence, it follows from (18) that $\bar{I}(t) = \bar{I}(t_1)$ for all $t \in (t_1, t_2]$. Therefore, we have

$$\begin{aligned} 0 = \bar{W}(t_1) &= \bar{I}(t_1) - \frac{\theta}{\mu_2} \int_0^{t_1} \bar{Q}_2(s) ds \\ &\geq \bar{I}(t_2) - \frac{\theta}{\mu_2} \int_0^{t_2} \bar{Q}_2(s) ds = \bar{W}(t_2) > 0, \end{aligned}$$

which reaches a contradiction. Hence, for all $t \geq 0$, $\bar{W}(t) = 0$ and thus $\bar{Q}_k(t) = 0$. $\bar{T}_k(t) = (\lambda_k/\mu_k)t$, $k = 1, 2$ follows from (12)–(14).

Therefore, due to the uniqueness of the weak convergence limit, we can conclude that (21) holds.

Finally, we give a proof of Proposition 2.

Proof. Proof of Proposition 2. Since the class 1 customers in queue at time t are those class 1 customers arriving between $[t - \tau_1^n(t), t]$, we have

$$|Q_1^n(t) - (A_1^n(t) - A_1^n((t - \tau_1^n(t))-)| \leq 1. \tag{22}$$

Hence, we have

$$\begin{aligned}
&|\bar{Q}_1^n(t) - \lambda_1 \bar{\tau}_1^n(t)| \\
&\leq |\bar{A}_1^n(t) - \bar{A}_1^n((t - \bar{\tau}_1^n(t))-) - \lambda_1 \bar{\tau}_1^n(t)| + \frac{1}{n}, && (23) \\
&|\widehat{Q}_1^n(t) - \lambda_1^n \widehat{\tau}_1^n(t)| \\
&\leq |\widehat{A}_1^n(t) - \widehat{A}_1^n((t - \bar{\tau}_1^n(t))-)| + \frac{1}{\sqrt{n}}. && (24)
\end{aligned}$$

It follows from the Functional Law of Large Numbers that

$$\sup_{0 \leq s \leq t \leq T} |\bar{A}_1^n(t) - \bar{A}_1^n(s) - \lambda_1 (t - s)| \Rightarrow 0$$

and thus

$$\sup_{0 \leq t \leq T} |\bar{A}_1^n(t) - \bar{A}_1^n((t - \bar{\tau}_1^n(t))-) - \lambda_1 \bar{\tau}_1^n(t)| \Rightarrow 0 \tag{25}$$

as $n \to \infty$.

Combining (21), (23) and (25), we have $\bar{\tau}_1^n \Rightarrow 0$ as $n \to \infty$. Hence, it follows from (10), (24), and the Random-Time-Change theorem that (11) holds.

Lemma 4. *Under any family of work-conserving policies, $\widehat{W}^n$ is stochastically bounded.*

Proof. It follows from (1)–(3) that

$$\begin{aligned}
\widehat{Q}_1^n(t) &= \widehat{Q}_1^n(0) + \widehat{A}_1^n(t) - \widehat{S}_1^n(\bar{T}_1^n(t)) \\
&\qquad + \sqrt{n}(\lambda_1^n t - \mu_1 \bar{T}_1^n(t)), && (26) \\
\widehat{Q}_2^n(t) &= \widehat{Q}_2^n(0) + \widehat{A}_2^n(t) - \widehat{S}_2^n(\bar{T}_2^n(t)) - \widehat{R}^n(t) \\
&\qquad + \sqrt{n}(\lambda_2^n t - \mu_2 \bar{T}_2^n(t)), && (27) \\
\widehat{R}^n(t) &= \widehat{N}^n\left((n\theta^n) \int_0^t (\bar{Q}_2^n(s))\, ds\right) \\
&\qquad + (n\theta^n) \int_0^t \widehat{Q}_2^n(s) ds. && (28)
\end{aligned}$$

Thus, we have

$$\widehat{W}^n(t) = \widehat{W}^n(0) + \widehat{X}^n(t) - \frac{\widehat{R}^n(t)}{\mu_2} + \widehat{I}^n(t), \tag{29}$$

$$\int_0^\infty \mathbf{1}\{\widehat{W}^n(t) > 0\} d\widehat{I}^n(t) = 0, \tag{30}$$

where

$$\widehat{X}^n(t) = \frac{\widehat{A}_1^n(t)}{\mu_1} + \frac{\widehat{A}_2^n(t)}{\mu_2} - \frac{\widehat{S}_1^n(\bar{T}_1^n(t))}{\mu_1} - \frac{\widehat{S}_1^n(\bar{T}_2^n(t))}{\mu_2}$$

$$+\sqrt{n}\left(\frac{\lambda_1^n}{\mu_1}+\frac{\lambda_2^n}{\mu_2}-1\right)t. \tag{31}$$

Therefore, it follows from (10), (21), (31), the Random-Time-Change theorem and Assumption 1 that

$$\widehat{X}^n \Rightarrow \widehat{X}, \text{ as } n \to \infty, \tag{32}$$

where $\widehat{X}$ is defined in Theorem 2.

Thus, $\widehat{X}^n$ is stochastically bounded. It follows from (29), (30) and $\widehat{R}^n(t) \geq 0$, $\widehat{W}^n(t) \geq 0$ for all $t \geq 0$ that $\widehat{W}^n(t) \leq \sup_{0\leq s\leq t}|\widehat{W}^n(0)+\widehat{X}^n(s)|$ and thus $\sup_{0\leq t\leq T}\widehat{W}^n(t) \leq \sup_{0\leq t\leq T}|\widehat{W}^n(0)+\widehat{X}^n(t)| \leq \widehat{W}^n(0)+\sup_{0\leq t\leq T}|\widehat{X}^n(t)|$. Therefore, $\widehat{W}^n$ is also stochastically bounded in virtue of Assumption 4.

B. A Generalized Reflection Map

In this part we give a generalized reflection map and its property of continuity in parameters, which is used in the proof of diffusion limit convergence result, i.e., Theorem 3 and Lemma 6. Below, we use $\mathbb{D}([0,\infty),\mathbb{R})$ to denote the set of all functions $x:[0,\infty)\to\mathbb{R}$ which are right continuous in $[0,\infty)$ and have finite left limits on $(0,\infty)$.

Definition 3. Given $\theta>0$, $d\geq 0$ and $x\in\mathbb{D}([0,\infty),\mathbb{R})$ with $x(0)\geq 0$, we define

$$(\phi^{\theta,d},\psi^{\theta,d}):\mathbb{D}([0,\infty),\mathbb{R})\to\mathbb{D}([0,\infty),[0,\infty)\times[0,\infty))$$

by $(\phi^{\theta,d},\psi^{\theta,d})(x)=(z,l)$, where

1) $z(t)=x(t)-\theta\int_0^t(z(s)-d)^+ds+l(t)\geq 0$ for all $t\geq 0$;
2) l is nondecreasing, $l(0)=0$, and $\int_0^\infty \mathbf{1}\{z(t)>0\}dl(t)=0$.

Define (ϕ,ψ) be the conventional one-sided refection map with lower barrier at 0. That is, for $x\in\mathbb{D}([0,\infty),\mathbb{R})$ with $x(0)\geq 0$, we have $z=\phi(x)$, $l=\psi(x)$, where $z(t)=x(t)+l(t)\geq 0$ for all $t\geq 0$; l is nondecreasing, $l(0)=0$, and $\int_0^\infty \mathbf{1}\{z(t)>0\}dl(t)=0$. This map has an explicit representation: $\phi(x)(t)=x(t)+\psi(x)(t)$ and $\psi(x)(t)=\sup_{0\leq s\leq t}(-x(s))^+$.

Define the map $\nu^{\theta,d}:\mathbb{D}([0,\infty),\mathbb{R})\to\mathbb{D}([0,\infty),\mathbb{R})$ as $\nu^{\theta,d}(x)=v$, where $(\phi^{\theta,d},\psi^{\theta,d})(x)=(z,l)$ satisfying

$$v(t)=x(t)-\theta\int_0^t(\phi(v)(s)-d)^+ds \tag{33}$$

with $v(0)=x(0)$. Then, we have

$$\phi^{\theta,d}(x)=\phi(\nu^{\theta,d}(x)),\psi^{\theta,d}(x)=\psi(\nu^{\theta,d}(x)) \tag{34}$$

Lemma 5. (i) *For each $x\in\mathbb{D}([0,\infty),\mathbb{R})$, there exists a unique function $\nu^{\theta,d}(x)=v$ satisfying* (33).

(ii) *$\nu^{\theta,d}(x)$ is continuous in (x,θ,d) with respect to the product topology, with $\mathbb{D}([0,\infty),\mathbb{R})$ equipped with the topology of uniform convergence over bounded intervals, and $[0,\infty)$ equipped with the order topology.*

Proof. (i). It follows immediately from Lemma 1 in [17] and the Lipschitz continuity of ϕ.

(ii). We use the same argument as that in the proof of Lemma 2 in [8]. Suppose that $(x^n, \theta^n, d^n) \to (x, \theta, d)$ as $n \to \infty$. Fix $\epsilon > 0$ and $t > 0$, and denote $v^n = \nu^{\theta^n, d^n}(x^n)$ and $v = \nu^{\theta,d}(x)$. There exists a strictly positive constant $\bar{\theta}$ and a number N such that for all $n \geq N$,

$$\begin{aligned}
&\theta^n \leq \bar{\theta},\\
&\|x^n - x\|_t + |\theta^n - \theta| \int_0^t (\phi(v)(s) - d)^+ ds\\
&\qquad\qquad +\theta^n |d^n - d| t < \delta
\end{aligned}$$

for some $\delta > 0$ which is yet to be determined, where we denote $\|x\|_t = \sup_{0 \leq s \leq t} |x(s)|$. We have

$$\begin{aligned}
&\|v^n - v\|_t\\
&\leq \|x^n - x\|_t + |\theta^n - \theta| \int_0^t (\phi(v)(s) - d)^+ ds\\
&+\theta^n \int_0^t (|(\phi(v^n)(s) - d^n)^+ - (\phi(v^n)(s) - d)^+|\\
&\qquad +|(\phi(v^n)(s) - d)^+ - (\phi(v)(s) - d)^+|) ds\\
\leq\ & \|x^n - x\|_t + |\theta^n - \theta| \int_0^t (\phi(v)(s) - d)^+ ds\\
&+\theta^n |d^n - d| t + 2\bar{\theta} \int_0^t \|v^n - v\|_s ds\\
\leq\ & \delta + 2\bar{\theta} \int_0^t \|v^n - v\|_s ds,
\end{aligned}$$

where the last inequality is due to $|a^+ - b^+| \leq |a - b|$ for any $a, b \in \mathbb{R}$ and $\|\phi(x) - \phi(x')\|_t \leq 2\|x - x'\|_t$ for any $x, x' \in \mathbb{D}([0, \infty), \mathbb{R})$, and $\theta^n \leq \bar{\theta}$. It follows from Gronwall's inequality that $\|v^n - v\|_t \leq \delta e^{2\bar{\theta} t}$. The desired continuity is obtained by setting $\delta = \epsilon e^{-2\bar{\theta} t}$.

Now we state several properties of the maps $\phi^{\theta,d}$ and $\psi^{\theta,d}$.

Proposition 3. (i) *For each* $x \in \mathbb{D}([0, \infty), \mathbb{R})$ *with* $x(0) \geq 0$, *there exists a unique pair of functions* $(\phi^{\theta,d}, \psi^{\theta,d})(x) = (z, l)$ *satisfying Definition 3.*

(ii) $\phi^{\theta,d}(x)$ *is continuous in* (x, θ, d) *with respect to the product topology, with* $\mathbb{D}([0, \infty), \mathbb{R})$ *equipped with the topology of uniform convergence over bounded intervals, and* $[0, \infty)$ *equipped with the order topology.*

Proof. (i). The existence and uniqueness follows from (34), Lemma 5 (i) and the existence and uniqueness of the conventional one-sided refection map (ϕ, ψ).

(ii). It follows from (34), Lemma 5 (ii) and the Lipschitz continuity of ϕ.

C. Proof of Theorem 2

Under the proposed policy, the following equations hold additionally:

$$\int_0^\infty \mathbf{1}\{W^n(t) > 0\} dI^n(t) = 0, \tag{35}$$

$$\int_0^\infty \mathbf{1}\left\{\tau_1^n(t) \le d^n - \epsilon^n, Q_2^n(t) > 0\right\} dT_1^n(t) = 0, \tag{36}$$

$$\int_0^\infty \mathbf{1}\left\{\tau_1^n(t) > d^n - \epsilon^n\right\} dT_2^n(t) = 0. \tag{37}$$

Fix any asymptotically compliant family of policies $\{\pi^n, n \in \mathbb{N}\}$. If π^n is not work-conserving, we construct a work-conserving policy $\pi^{n,w}$ such that the server serves the same class of customers as that under policy π^n, unless there is no customer of that class. Besides, if the server under policy π^n is idle, give priority to serve class 2 customers under policy $\pi^{n,w}$. By the standard coupling argument, we can couple the two systems such that on each sample path, $Q_k^n(t) \ge Q_k^{n,w}(t)$ for all $t \ge 0$, where $Q_k^{n,w}(t)$ is the number of class k customers under policy $\pi^{n,w}$, $k = 1, 2$. Hence, $\{\pi^{n,w}, n \in \mathbb{N}\}$ is also an asymptotically compliant family of policies. Moreover, $R^n(t) = R(\int_0^t Q_2^n(s)ds) \ge_{st} R^{n,w}(t) = R(\int_0^t Q_2^{n,w}(s)ds)$, where $\ge_{st}$ is the standard stochastic order and $R^{n,w}(t)$ is the cumulative abandonment number of class 2 customers till time t under policy $\pi^{n,w}$. Therefore, $\mathbb{P}\{\widehat{R}^n(T) > x\} \ge \mathbb{P}\{\widehat{R}^{n,w}(T) > x\}$ for all $x > 0$ and thus we can consider work-conserving policy only.

Now fix any asymptotically compliant family of work-conserving policies $\{\pi^n, n \in \mathbb{N}\}$. We will construct an alternative system and a new policy, and prove that the cumulative abandonment number in the alternative system under the new policy is pathwise not less than that in the original system with a probability nearly 1.

First, it follows from Corollary 1 and Assumption 3 that there exists a sequence of numbers ϵ^n satisfying $\epsilon^n/\sqrt{n} \to 0$ as $n \to \infty$ and

$$\mathbb{P}\left(\sup_{0\le t\le T}\left[\widehat{Q}_1^n(t) - \frac{\lambda_1^n d^n}{\sqrt{n}}\right]^+ > \frac{\epsilon^n}{\sqrt{n}}\right) \le \frac{\epsilon^n}{\sqrt{n}}, \text{ for all } n \in \mathbb{N}.$$

That is,

$$\mathbb{P}\left(\sup_{0\le t\le nT}\left[Q_1^n(t) - \lambda_1^n d^n\right]^+ > \epsilon^n\right) \le \frac{\epsilon^n}{\sqrt{n}}, \text{ for all } n \in \mathbb{N}.$$

Define

$$\Gamma^n(T) = \left\{\sup_{0\le t\le nT}\left[Q_1^n(t) - \lambda_1^n d^n\right]^+ \le \epsilon^n\right\}. \tag{38}$$

Then, we have $\lim_{n\to\infty} \mathbb{P}\{\Gamma^n(T)\} = 1$.

We construct a new (alternative) system which is the same as the original one, except that the abandonment rate of class 2 customers is now state dependent, rather than a constant

θ^n. To be specific, the abandonment rate of class 2 customers at time t is now

$$\frac{\theta^n \mu_2 \left(W^{n,a}(t) - \frac{\lambda_1^n d^n + \epsilon^n}{\mu_1} \right)^+}{Q_2^{n,a}(t)},$$

where superscript "a" stands for alternative and

$$W^{n,a}(t) = \frac{Q_1^{n,a}(t)}{\mu_1} + \frac{Q_2^{n,a}(t)}{\mu_2}$$

is the workload in the alternative system at time t, $Q_k^{n,a}(t)$ is the number of class k customers in the alternative system at time t, $k = 1, 2$.

Now we will construct a new policy $\pi^{n,a}$ for the new system, based on a pathwise modification of the original system under policy π^n, such that

$$Q_1^{n,a}(t) = Q_1^n(t),\ Q_2^{n,a}(t) \geq Q_2^n(t),\ \text{for all } t \in [0, nT] \\ \text{on the event } \Gamma^n(T). \tag{39}$$

We construct policy $\pi^{n,a}$ so that when there is a new service starting in the original system, there is a customer with the same class starting service in the new system (maybe preemptive; it follows from (39) that there must be a customer with the same class in the new system). Furthermore, we impose that $\pi^{n,a}$ is work-conserving.

Next we show (39) holds by induction on time event. Obviously, (39) holds at $t = 0$ since both systems have the same initial state. Suppose that it holds at time t. There are three possibilities for the next time event t':

- *Arrival:* there is an arrival to the original system. We can couple two systems so that there is also an arrival of the same class to the new system. (39) holds at time t'.

- *Service completion:* there is a service completion in the original system. By the construction of policy $\pi^{n,a}$ we can couple two systems so that there is also a service completion of the same customer class in the new system. (39) still holds at time t'.

- *Customer abandonment:* if $Q_2^{n,a}(t) > Q_2^n(t)$, then $Q_2^{n,a}(t') \geq Q_2^{n,a}(t) - 1 \geq Q_2^n(t) \geq Q_2^n(t')$ and thus (39) holds at time t'. If $Q_2^{n,a}(t) = Q_2^n(t)$, then we can couple two systems so that there is a class 2 customer abandoning the original system, while there is a class 2 customer abandoning the new system with probability

$$\frac{\mu_2 \left(W^{n,a}(t') - \frac{\lambda_1^n d^n + \epsilon^n}{\mu_1} \right)^+}{Q_2^{n,a}(t')} \\ \leq \frac{\mu_2 \left(W^{n,a}(t') - \frac{Q_1^{n,a}(t')}{\mu_1} \right)}{Q_2^{n,a}(t')} = 1,$$

  where the inequality is due to $Q^n(t') - \lambda_1^n d^n \leq \epsilon^n$ on the set $\Gamma^n(T)$ and $Q_1^{n,a}(t') = Q_1^n(t')$. Hence, (39) still holds at time t'.

By the construction of policy $\pi^{n,a}$ and (39), when the server in the original system serves class 1 customers, the server in the new system also serves class 1 customers, and vice versa; when the server in the original system serves class 2 customers, the server in the new system will also serve class 2 customers; however, it is possible that the original system is empty and there are class 2 customers in the new system. In this case, the server in the new system continues serving class 2 customers. Therefore, we have

$$T_1^{n,a}(t) = T_1^n(t),\ T_2^{n,a}(t) \geq T_2^n(t),\ \text{for all } t \in [0, nT] \\ \text{on the event } \Gamma^n(T), \tag{40}$$

where $T_k^{n,a}(t)$ is the cumulative amount of service time devoted to serving class k customers till time t in the new system.

We have the following equations:

$$Q_1^{n,a}(t) = Q_1^{n,a}(0) + A_1^n(t) - S_1^n(T_1^{n,a}(t)), \tag{41}$$

$$Q_2^{n,a}(t) = Q_2^{n,a}(0) + A_2^n(t) - S_2^n(T_2^{n,a}(t)) - R^{n,a}(t). \tag{42}$$

where

$$R^{n,a}(t) = N\left(\int_0^t \theta^n \mu_2 \left(W^{n,a}(s) - \frac{\lambda_1^n d^n + \epsilon^n}{\mu_1}\right)^+ ds\right). \tag{43}$$

Therefore, we have

$$\frac{R^{n,a}(t)}{\mu_2} = W^{n,a}(0) - W^{n,a}(t) + \frac{A_1^n(t)}{\mu_1} + \frac{A_2^n(t)}{\mu_2} \\ - \frac{S_1^n(T_1^{n,a}(t))}{\mu_1} - \frac{S_2^n(T_2^{n,a}(t))}{\mu_2}.$$

Besides, it follows from (1)–(3) that

$$\frac{R^n(t)}{\mu_2} = W^n(0) - W^n(t) + \frac{A_1^n(t)}{\mu_1} + \frac{A_2^n(t)}{\mu_2} \\ - \frac{S_1^n(T_1^n(t))}{\mu_1} - \frac{S_2^n(T_2^n(t))}{\mu_2}.$$

Hence, it follows from (39) and (40) that

$$R^{n,a}(t) \leq R^n(t),\ \text{for all } t \in [0, nT] \text{ on } \Gamma^n(T). \tag{44}$$

Similar to the original system, we can also define the fluid scaled processes and the diffusion scaled processes for the alternative system. The detailed definitions are omitted for brevity.

Next, we show the following diffusion limit result of the alternative system.

Lemma 6. *Under any family of work-conserving policies,*

$$\widehat{W}^{n,a} \Rightarrow \widehat{W},\ \widehat{R}^{n,a} \Rightarrow \widehat{R},\ as\ n \to \infty, \tag{45}$$

where $\widehat{W}$ is defined in (7) and

$$\widehat{R}(t) = \theta\mu_2 \int_0^t \left(\widehat{W}(s) - \frac{\lambda_1 d}{\mu_1}\right)^+ ds.$$

Proof. Since the policy is work-conserving, additionally it holds that

$$\int_0^\infty \mathbf{1}\{W^{n,a}(t) > 0\} dI^{n,a}(t) = 0.$$

With the same argument as the proof of Lemma 2, we can show that the family

$$\{\bar{Q}_k^{n,a}, \bar{T}_k^{n,a}, \bar{A}_k^n, \bar{R}^{n,a}, k = 1, 2\}$$

is precompact, and the corresponding weak limit processes $(\bar{Q}_k^a, \bar{T}_k^a, \bar{A}_k, \bar{R}^a, k = 1, 2)$ satisfy the following equations:

$$\bar{A}_k(t) = \lambda_k t, k = 1, 2, \tag{46}$$

$$\bar{Q}_1^a(t) = \bar{Q}_1^a(0) + \bar{A}_1(t) - \mu_1 \bar{T}_1^a(t), \tag{47}$$

$$\bar{Q}_2^a(t) = \bar{Q}_2^a(0) + \bar{A}_2(t) - \mu_2 \bar{T}_2^a(t) - \bar{R}^a(t), \tag{48}$$

$$\bar{R}^a(t) = \theta \int_0^t \bar{W}^a(s) ds, \tag{49}$$

$$\int_0^\infty \mathbf{1}\{\bar{W}^a(t) > 0\} d\bar{I}^a(t) = 0. \tag{50}$$

With a similar argument as that in the proof of Lemma 3, we can conclude that

$$(\bar{Q}^{n,a}, \bar{T}^{n,a}) \to (0, \bar{T}^{*,a}),\ \text{as } n \to \infty, \tag{51}$$

where $\bar{T}^{*,a}(t) = \left(\frac{\lambda_1}{\mu_1} t, \frac{\lambda_2}{\mu_2} t\right)$.

It follows from (41)–(43) that

$$\begin{aligned}\widehat{Q}_1^{n,a}(t) &= \widehat{Q}_1^{n,a}(0) + \widehat{A}_1^n(t) - \widehat{S}_1^n(\bar{T}_1^{n,a}(t))\\ &+\sqrt{n}(\lambda_1^n t - \mu_1 \bar{T}_1^{n,a}(t)),\end{aligned} \tag{52}$$

$$\begin{aligned}\widehat{Q}_2^{n,a}(t) &= \widehat{Q}_2^{n,a}(0) + \widehat{A}_2^n(t) - \widehat{S}_2^n(\bar{T}_2^{n,a}(t)) - \widehat{R}^{n,a}(t)\\ &+\sqrt{n}(\lambda_2^n t - \mu_2 \bar{T}_2^{n,a}(t)),\end{aligned} \tag{53}$$

$$\widehat{R}^{n,a}(t) = \widehat{N}^n \left((n\theta^n)\mu_2 \int_0^t \left(\bar{W}^{n,a}(s) - \frac{\lambda_1^n d^n + \epsilon^n}{n\mu_1}\right)^+ ds\right)$$

$$+(n\theta^n)\mu_2 \int_0^t \left(\widehat{W}^{n,a}(s) - \frac{\lambda_1^n d^n + \epsilon^n}{\sqrt{n}\mu_1} \right)^+ ds. \tag{54}$$

Thus, we have

$$\widehat{W}^{n,a}(t) = \widehat{W}^{n,a}(0) + \widehat{X}^{n,a}(t) - \frac{\widehat{R}^{n,a}(t)}{\mu_2} + \widehat{I}^{n,a}(t) \geq 0, \tag{55}$$

$$\widehat{I}^{n,a} \text{ is nondecreasing with } \widehat{I}^{n,a}(0) = 0, \tag{56}$$

$$\int_0^\infty \mathbf{1}\{\widehat{W}^{n,a}(t) > 0\} d\widehat{I}^{n,a}(t) = 0, \tag{57}$$

where

$$\widehat{X}^{n,a}(t) = \frac{\widehat{A}_1^n(t)}{\mu_1} + \frac{\widehat{A}_2^n(t)}{\mu_2} - \frac{\widehat{S}_1^n(\bar{T}_1^{n,a}(t))}{\mu_1} - \frac{\widehat{S}_1^n(\bar{T}_2^{n,a}(t))}{\mu_2}$$
$$+\sqrt{n} \left(\frac{\lambda_1^n}{\mu_1} + \frac{\lambda_2^n}{\mu_2} - 1 \right) t. \tag{58}$$

Therefore, it follows from (10), (51), (58), the Random-Time-Change theorem and Assumption 1 that

$$\widehat{X}^{n,a} \Rightarrow \widehat{X}, \text{ as } n \to \infty, \tag{59}$$

where $\widehat{X}$ is defined in Theorem 2.

It follows from (54), (55), (57) and Definition 3 that

$$\widehat{W}^{n,a} = \phi^{n\theta^n, \frac{\lambda_1^n d^n + \epsilon^n}{\sqrt{n}\mu_1}}(Y^{n,a}),$$

where

$$Y^{n,a}(t) = \widehat{W}^{n,a}(0) + \widehat{X}^{n,a}(t)$$
$$- \frac{\widehat{N}^n \left((n\theta^n)\mu_2 \int_0^t \left(\bar{W}^{n,a}(s) - \frac{\lambda_1^n d^n + \epsilon^n}{n\mu_1} \right)^+ ds \right)}{\mu_2}.$$

It follows from (10), (51), (59), Assumptions 2 and 4, and the Random-Time-Change theorem that

$$Y^{n,a} \Rightarrow \widehat{W}_0 + \widehat{X}, \text{ as } n \to \infty.$$

Moreover, it follows from Assumptions 1, 2 and 3 that

$$n\theta^n \to \theta, \text{ and } \frac{\lambda_1^n d^n + \epsilon^n}{\sqrt{n}\mu_1} \to \frac{\lambda_1 d}{\mu_1}, \text{ as } n \to \infty.$$

Therefore, it follows from Proposition 3 and the Continuous Mapping theorem that

$$\widehat{W}^{n,a} \Rightarrow \phi^{\theta, \frac{\lambda_1 d}{\mu_1}}(\widehat{W}_0 + \widehat{X}), \text{ as } n \to \infty.$$

Comparing (7) with Definition 3, we have $\widehat{W}^{n,a} \Rightarrow \widehat{W}$, where $\widehat{W}$ is defined in (7). It follows from (54) that $\widehat{R}^{n,a} \Rightarrow \widehat{R}$ as $n \to \infty$, where

$$\widehat{R}(t) = \theta\mu_2 \int_0^t \left(\widehat{W}(s) - \frac{\lambda_1 d}{\mu_1}\right)^+ ds.$$

Therefore, it follows from (44), $\lim_{n\to\infty} \mathbb{P}\{\Gamma^n(T)\} = 1$ and Lemma 6 that

$$\begin{aligned}
&\mathbb{P}\{\widehat{R}^n(T) > x\} \geq \mathbb{P}\{\widehat{R}^n(T) > x, \Gamma^n(T)\} \\
&\geq \mathbb{P}\{\widehat{R}^{n,a}(T) > x, \Gamma^n(T)\} \\
&\to \mathbb{P}\left\{\theta\mu_2 \int_0^T \left(\widehat{W}(t) - \frac{\lambda_1 d}{\mu_1}\right)^+ dt > x\right\}, \text{ as } n \to \infty,
\end{aligned}$$

which completes the proof of Theorem 2.

D. Proof of Proposition 1

In order to show the SSC result, we use the framework of [5]. We mention that customer abandonment is not evolved in the framework of [5], and thus the results in [5] cannot be directly used. To ease the argument, we adopt a sample-path approach based on the Skorohod representation theorem, which has been used in [21] and [6]. The sample path approach turns the weak convergence into a probability one convergence of suitable copies of the processes on a common probability space.

In the rest of this section, we focus on a given sample path for which the above u.o.c. convergence holds.

As in [21] and [6], we consider a time interval $[\tau, \tau+\delta]$, where $\tau \geq 0$ and $\delta > 0$. Let $T > 0$ be a fixed time to be specified later. Divide the time interval $[\tau, \tau+\delta]$ into a total of $\lceil \sqrt{n}\delta/T \rceil$ segments with equal length $T/\sqrt{n}$ (except the last one) and define

$$\bar{\bar{W}}^{n,j}(u) := \widehat{W}^n\left(\tau + \frac{jT+u}{\sqrt{n}}\right) = \frac{W^n(n\tau + \sqrt{n}(jT+u))}{\sqrt{n}} \tag{60}$$

for $u \geq 0$ and $j = 0, 1, \cdots, \lfloor \sqrt{n}\delta/T \rfloor$. Similarly, we define the hydrodynamically scaled processes

$$\begin{aligned}
\bar{\bar{Q}}^{n,j}(u) &= \widehat{Q}^n\left(\tau + \frac{jT+u}{\sqrt{n}}\right), \\
\bar{\bar{T}}^{n,j}(u) &= \widehat{T}^n\left(\tau + \frac{jT+u}{\sqrt{n}}\right) - \widehat{T}^n\left(\tau + \frac{jT}{\sqrt{n}}\right), \\
\bar{\bar{A}}^{n,j}(u) &= \widehat{A}^n\left(\tau + \frac{jT+u}{\sqrt{n}}\right) - \widehat{A}^n\left(\tau + \frac{jT}{\sqrt{n}}\right), \\
\bar{\bar{R}}^{n,j}(u) &= \widehat{R}^n\left(\tau + \frac{jT+u}{\sqrt{n}}\right) - \widehat{R}^n\left(\tau + \frac{jT}{\sqrt{n}}\right),
\end{aligned}$$

$$
\begin{aligned}
\bar{\bar{I}}^{n,j}(u) &= \widehat{I}^n\left(\tau + \frac{jT+u}{\sqrt{n}}\right) - \widehat{I}^n\left(\tau + \frac{jT}{\sqrt{n}}\right),\\
\bar{\bar{\tau}}_1^{n,j}(u) &= \widehat{\tau}_1^n\left(\tau + \frac{jT+u}{\sqrt{n}}\right).
\end{aligned}
$$

Proposition 4. *Let M be a given positive constant and j_n be some integer with $j_n \in [0, \sqrt{n}\delta/T]$. Suppose $|\bar{\bar{W}}^{n,j_n}(0)| \le M$ for sufficiently large n. Then, for any subsequence of $\{n\}$, there exists a further subsequence, denoted by $\mathcal{N}$, such that along $\mathcal{N}$, the processes $(\bar{\bar{Q}}^{n,j_n}, \bar{\bar{T}}^{n,j_n}, \bar{\bar{A}}^{n,j_n}, \bar{\bar{R}}^{n,j_n}, \bar{\bar{W}}^{n,j_n}, \bar{\bar{I}}^{n,j_n}, \bar{\bar{\tau}}_1^{n,j_n})$ converge u.o.c. to limit processes $(\bar{\bar{Q}}, \bar{\bar{T}}, \bar{\bar{A}}, \bar{\bar{R}}, \bar{\bar{W}}, \bar{\bar{I}}, \bar{\bar{\tau}}_1)$, which satisfy the following equations:*

$$
\begin{aligned}
&\bar{\bar{A}}_k(t) = \lambda_k t, \quad k = 1, 2, &(61)\\
&\bar{\bar{Q}}_1(t) = \bar{\bar{Q}}_1(0) + \bar{\bar{A}}_1(t) - \mu_1 \bar{\bar{T}}_1(t), &(62)\\
&\bar{\bar{Q}}_2(t) = \bar{\bar{Q}}_2(0) + \bar{\bar{A}}_2(t) - \mu_2 \bar{\bar{T}}_2(t) - \bar{\bar{R}}(t), &(63)\\
&\bar{\bar{R}}(t) = 0, &(64)\\
&\bar{\bar{N}}(t) = t, &(65)\\
&\bar{\bar{I}}(t) = t - \bar{\bar{T}}_1(t) - \bar{\bar{T}}_2(t), &(66)\\
&\bar{\bar{W}}(t) = \frac{\bar{\bar{Q}}_1(t)}{\mu_1} + \frac{\bar{\bar{Q}}_2(t)}{\mu_2}, &(67)\\
&\bar{\bar{\tau}}_1(t) = \frac{\bar{\bar{Q}}_1(t)}{\lambda_1}, &(68)\\
&\int_0^\infty \mathbf{1}\{\bar{\bar{W}}(t) > 0\} d\bar{\bar{I}}(t) = 0, &(69)\\
&\int_0^\infty \mathbf{1}\left\{\bar{\bar{\tau}}_1(t) < d, \bar{\bar{Q}}_2(t) > 0\right\} d\bar{\bar{T}}_1(t) = 0, &(70)\\
&\int_0^\infty \mathbf{1}\left\{\bar{\bar{\tau}}_1(t) > d\right\} d\bar{\bar{T}}_2(t) = 0. &(71)
\end{aligned}
$$

Proof. Most results can be obtained by using the same argument as the proof of Lemma 2. For example, it follows from

$$
\bar{\bar{R}}^{n,j_n}(u) = \frac{N\left(\theta^n \int_{n\tau+\sqrt{n}j_nT}^{n\tau+\sqrt{n}(j_nT+u)} Q_2^n(u)du\right)}{\sqrt{n}} = \frac{N\left(n\theta^n \int_0^u \bar{\bar{Q}}_2^{n,j_n}(s)ds\right)}{\sqrt{n}},
$$

the stochastic boundedness of $\bar{\bar{Q}}_2^{n,j_n}$ (as implied by (2)) and Assumption 2 that $\bar{\bar{R}}^{n,j_n} \to 0$, u.o.c. as $n \to \infty$.

Similar to Proposition 2, we have the following result, which relates the age process of the head-of-line class 1 customers to the queue length process of class 1 customers on the hydrodynamic scale. Since the proof is quite similar to that of Proposition 2, we omit it for brevity.

Lemma 7. *For any $T > 0$, $\sup_{0\le t\le T} |\lambda_1^n \bar{\bar{\tau}}_1^{n,j_n}(t) - \bar{\bar{Q}}_1^n(t)| \Rightarrow 0$ as $n \to \infty$.*

Equation (67) follows immediately by virtue of Lemma 7. Relations (69)–(71) follow from (35)–(37), using a similar argument as the proof of Proposition 3 (relation (24) therein) in [6].

We mention that (61)–(67) hold for any policy, (68) and (69) hold for any work-conserving policy, and (70) and (71) hold for our proposed policy.

Any processes $(\bar{\bar{Q}}, \bar{\bar{T}}, \bar{\bar{A}}, \bar{\bar{R}}, \bar{\bar{W}}, \bar{\bar{I}}, \bar{\bar{\tau}}_1)$ satisfying (61)–(71) is called a hydrodynamic model solution, which is obviously Lipschitz continuous. Hence, they are differentiable at almost all time $t \ge 0$. Below, when we write the derivative of such processes with respect to time t, we assume by default that such a time is regular, i.e., all the related processes are differentiable at this time t.

The following lemma, which is Lemma 5.2 in [7], is needed for proving Proposition 5, presenting a uniform attraction property of the hydrodynamic limit.

Lemma 8. *Let $f : [0,\infty) \to [0,\infty)$ be an absolutely continuous function and $\delta > 0$. Assume that whenever $f(t) > 0$ and f is differentiable at t, $\dot{f}(t) \le -\delta$. Then $f(t) = 0$ for all $t \ge f(0)/\delta$.*

Recall that Q^* is defined in (8).

Proposition 5 (Uniform attraction). *Assume that $\bar{\bar{W}}(0) < M$ for some constant $M > 0$. Then, there exists a time constant T_M such that, for all $t \ge T_M$, $\bar{\bar{Q}}(t) = Q^*(\bar{\bar{W}}(t))$, and $\bar{\bar{W}}(t) = \bar{\bar{W}}(T_M)$. Moreover, if $\bar{\bar{Q}}(0) = Q^*(\bar{\bar{W}}(0))$, then $T_M = 0$ and $\bar{\bar{Q}}(t) = Q^*(\bar{\bar{W}}(0))$ for all $t \ge 0$.*

Proof. First, we prove that there exists a time T'_M such that $\bar{\bar{Q}}_1(t) \le \lambda_1 d$ for all $t \ge T'_M$. Define $f(t) = (\bar{\bar{Q}}_1(t) - \lambda_1 d)^+$. If $f(t) > 0$, then it follows from (68) and (71) that $\dot{\bar{\bar{T}}}_2(t) = 0$. Moreover, $\bar{\bar{W}}(t) > 0$ and thus it follows from (69) that $\dot{\bar{\bar{I}}}(t) = 0$. Hence, (66) implies $\dot{\bar{\bar{T}}}_1(t) = 1$ and thus $f'(t) = \dot{\bar{\bar{Q}}}_1(t) = \lambda_1 - \mu_1 < 0$ by virtue of (62). Hence, it follows from Lemma 8 that there exists a time T'_M such that $f(t) = 0$ for all $t \ge T'_M$.

Next, we prove that there exists a time $T_M \ge T'_M$ such that $\bar{\bar{Q}}_1(t) = \mu_1 \min\left(\bar{\bar{W}}(t), \lambda_1 d/\mu_1\right)$. For $t \ge T'_M$, it holds that $\bar{\bar{Q}}_1(t) \le \mu_1 \min\left(\bar{\bar{W}}(t), \lambda_1 d/\mu_1\right)$. Define

$$g(t) = \min\left(\bar{\bar{Q}}_2(t), \lambda_1 d - \bar{\bar{Q}}_1(t)\right).$$

If $g(t) > 0$, then $\bar{\bar{Q}}_2(t) > 0$ and $\bar{\bar{Q}}_1(t) < \lambda_1 d$. Hence, it follows from (68) and (70) that $\dot{\bar{\bar{T}}}_1(t) = 0$. Moreover, $\bar{\bar{W}}(t) > 0$ and thus it follows from (69) that $\dot{\bar{\bar{I}}}(t) = 0$. Hence, (66) implies that $\dot{\bar{\bar{T}}}_2(t) = 1$. Thus, $\dot{\bar{\bar{Q}}}_2(t) = \lambda_2 - \mu_2$ and $\dot{\bar{\bar{Q}}}_1(t) = \lambda_1$. Hence, $g'(t) \le \max(\lambda_2 - \mu_2, -\lambda_1) < 0$. It follows from Lemma 8 that there exists a time T_M such that $g(t) = 0$ for all $t \ge T_M$. Thus, for $t \ge T_M$, we have $\min(\bar{\bar{Q}}_2(t), \lambda_1 d - \bar{\bar{Q}}_1(t)) = 0$

and $\bar{\bar{Q}}_1(t) \le \lambda_1 d$. It follows from (67) that $\bar{\bar{Q}}_1(t) = \mu_1 \min\left(\bar{\bar{W}}(t), \lambda_1 d/\mu_1\right)$ and then $\bar{\bar{Q}}_2(t) = \mu_2 \left(\bar{\bar{W}}(t) - \lambda_1 d/\mu_1\right)^+$ for all $t \ge T_M$.

It follows from (62), (63), (64), (66) and (67) that $\dot{\bar{\bar{W}}}(t) = \dot{\bar{\bar{I}}}(t)$ for all $t \ge T_M$. If $\bar{\bar{W}}(t) > 0$, then it follows from (69) that $\dot{\bar{\bar{I}}}(t) = 0$. Hence, $\dot{\bar{\bar{W}}}(t) = 0$ and thus $\bar{\bar{W}}$ is a constant hereafter. If $\bar{\bar{W}}(t) = 0$, then it follows from (69) that $\bar{\bar{W}}$ is 0 hereafter. In either case, $\bar{\bar{W}}(t) = \bar{\bar{W}}(T_M)$ for all $t \ge T_M$.

The second part of Proposition 5 is obtained immediately by observing that $T_M = 0$.

Proposition 6. *Consider the time interval* $[\tau, \tau+\delta]$*, with* $\tau \ge 0$ *and* $\delta > 0$*; choose a constant* $C > 0$ *such that*

$$\sup_{\tau \le t_1 < t_2 \le \tau+\delta} |\widehat{X}(t_1) - \widehat{X}(t_2)| \le C,$$

and suppose that

$$\lim_{n\to\infty} \widehat{W}^n(\tau) = \chi \text{ and } \lim_{n\to\infty} \widehat{Q}^n(\tau) = Q^*(\chi) \tag{72}$$

for some $\chi \ge 0$*. Let* $\epsilon > 0$ *be any given number. Then, there exists a sufficiently large* T *such that, for sufficiently large* n*, the following results hold for all integers* $j \in [0, \sqrt{n}\delta/T]$*:*

(a) $|\bar{\bar{Q}}^{n,j}(u) - Q^*(\bar{\bar{W}}^{n,j}(u))| \le \epsilon$ *for all* $u \in [0, T]$*;*

(b) $\bar{\bar{W}}^{n,j}(u) \le \chi + C + 1$ *for all* $u \in [0, T]$.

Proof. Let $T = T_{\chi+C+1}$. This time length T is sufficiently long so that in any hydrodynamic limit, $\bar{\bar{Q}}(t)$ will approach the fixed-point state from an initial state $\bar{\bar{Q}}(0)$ with $\bar{\bar{W}}(0) \le \chi + C + 1$.

We prove that properties (a) and (b) for $j = 0$ first. By the definition of $\bar{\bar{W}}^{n,j}(u)$, we have $(\bar{\bar{W}}^{n,0}(0), \bar{\bar{Q}}^{n,0}(0)) = (\widehat{W}^n(\tau), \widehat{Q}^n(\tau))$ and thus $(\bar{\bar{W}}^{n,0}(0), \bar{\bar{Q}}^{n,0}(0)) \to (\chi, Q^*(\chi))$ as $n \to \infty$, in view of (72). Hence, it follows from Propositions 4 and 5 that as $n \to \infty$,

$$(\bar{\bar{W}}^{n,0}(u), \bar{\bar{Q}}^{n,0}(u)) \to (\bar{\bar{W}}(u), \bar{\bar{Q}}(u)) = (\chi, Q^*(\chi)), \text{ u.o.c. in } u \in [0, T]. \tag{73}$$

Here the convergence is along the whole sequence of n because the limit is unique. Let n be sufficiently large such that $|\bar{\bar{W}}^{n,0}(u) - \chi| \le \epsilon/\max(\mu_1, \mu_2)$ and $|\bar{\bar{Q}}^{n,0}(u) - Q^*(\chi)| \le \epsilon/2$ for all $u \in [0, T]$. Then, we have

$$\begin{aligned} |\bar{\bar{Q}}^{n,0}(u) - Q^*(\bar{\bar{W}}^{n,0}(u))| &\le |\bar{\bar{Q}}^{n,0}(u) - Q^*(\chi)| + |Q^*(\bar{\bar{W}}^{n,0}(u)) - Q^*(\chi)| \\ &\le \frac{\epsilon}{2} + \max(\mu_1, \mu_2) \cdot |\bar{\bar{W}}^{n,0}(u) - \chi| \le \epsilon, \end{aligned}$$

for all $u \in [0, T]$. Hence, property (a) holds for $j = 0$ when n is sufficiently large.

It follows from (73) that $\bar{\bar{W}}^{n,0}(u)$ is close to χ for all $u \in [0, T]$ when n is sufficiently large, which leads to property (b) for $j = 0$.

Next, we proceed to verify properties (a) and (b) for $j = 1, \ldots, \lfloor\sqrt{n}\delta/T\rfloor$. Suppose, to the contrary, there exists a subsequence $\mathcal{N}_1$ of $\{n\}$ such that for any $n \in \mathcal{N}_1$, at least one of the properties (a, b) fails to hold for some integers $j \in [1, \sqrt{n}\delta/T]$. Then, for any $n \in \mathcal{N}_1$, there exists a smallest integer, denoted by j_n, in the interval $[1, \sqrt{n}\delta/T]$, such that at least one of the properties (a, b) fails to hold. To reach a contradiction, it suffices to construct an infinite subsequence $\mathcal{N}_2 \subset \mathcal{N}_1$, such that the properties (a, b) hold for $j = j_n$ for sufficiently large $n \in \mathcal{N}_2$.

From the contradictory assumption, we know that properties (a) and (b) hold for $j = 0, 1, \ldots, j_n - 1$, $n \in \mathcal{N}_1$. Specifically, for $j = j_n - 1$, we have $\bar{\bar{W}}^{n,j_n-1}(0) \leq \chi + C + 1$, for all $n \in \mathcal{N}_1$. Hence, it follows from Proposition 4, there exits a further subsequence $\mathcal{N}_2 \subset \mathcal{N}_1$, such that $(\bar{\bar{W}}^{n,j_n-1}(u), \bar{\bar{Q}}^{n,j_n-1}(u)) \rightarrow (\bar{\bar{W}}(u), \bar{\bar{Q}}(u))$, u.o.c., as $n \rightarrow \infty$ along $\mathcal{N}_2$ with $\bar{\bar{W}} \leq \chi + C + 1$. It follows from Proposition 5 that $\bar{\bar{Q}}(u) = Q^*(\bar{\bar{W}}(u))$ for all $u \geq T$. Hence, for sufficiently large $n \in \mathcal{N}_2$, $|\bar{\bar{Q}}^{n,j_n}(u) - Q^*(\bar{\bar{W}}^{n,j_n}(u))| = |\bar{\bar{Q}}^{n,j_n-1}(u + T) - Q^*(\bar{\bar{W}}^{n,j_n-1}(u + T))| < \epsilon$ for all $u \in [0, T]$. Hence, property (a) holds with $j = j_n$ for sufficiently large $n \in \mathcal{N}_2$.

Property (b) holds in view of (29), (30) and the relation between the hydrodynamic scale and diffusion scale. (See also the proof of Lemma 4). Hence, we have shown that properties (a) and (b) holds for $j = j_n$ when $n \in \mathcal{N}_2$ is sufficiently large , which contradicts the definition of the subsequence $\mathcal{N}_2$.

Note that Assumption 4 and (9) imply that (72) holds for $\tau = 0$. Hence, it follows from property (a) in Proposition 6 and the relation between the hydrodynamic scale and diffusion scale that $|\hat{Q}^n(t) - Q^*(\hat{W}^n(t))| \leq \epsilon$ for all $t \in [0, \delta]$ when n is sufficiently large, which leads to Proposition 1 by letting $\delta = T$.

E. Other Omitted Proofs

Proof. Proof of Theorem 3. We only need to prove that $\widehat{W}^n \Rightarrow \widehat{W}$ as $n \rightarrow \infty$, since then the result that $(\widehat{Q}_1^n, \widehat{Q}_2^n) \Rightarrow (\widehat{Q}_1, \widehat{Q}_2)$ as $n \rightarrow \infty$ follows immediately from Proposition 1.

It follows from (28)–(30) that

$$\widehat{W}^n = \phi^{n\theta^n, \frac{\lambda_1 d}{\mu_1}}(Y^n),$$

where

$$Y^n(t) = \widehat{W}^n(0) + \widehat{X}^n(t) - \frac{\widehat{N}^n\left((n\theta^n)\int_0^t \left(\bar{Q}_2^n(s)\right) ds\right)}{\mu_2}$$
$$-\frac{(n\theta^n)\int_0^t \left(\widehat{Q}_2(s) - \mu_2\left(\widehat{W}^n(s) - \frac{\lambda_1 d_1}{\mu_1}\right)^+\right) ds}{\mu_2}.$$

It follows from (10), (21), (32), Assumptions 2 and 4, Proposition 1 and the Random-Time-Change theorem that

$$Y^n \Rightarrow \widehat{W}_0 + \widehat{X}, \text{ as } n \rightarrow \infty.$$

Therefore, it follows from Proposition 3 and the Continuous Mapping theorem that

$$\widehat{W}^n \Rightarrow \phi^{\theta,\frac{\lambda_1 d}{\mu_1}}(\widehat{W}_0 + \widehat{X}), \text{ as } n \to \infty,$$

which completes the proof.

Proof. Proof of Theorem 1. It follows from Proposition 1 that

$$\sup_{0\le t\le T} \left[\widehat{Q}_1^n(t) - \lambda_1 d\right]^+ \Rightarrow 0, \text{ as } n \to \infty.$$

Thus, it follows from Proposition 2 and Assumptions 1 and 3 that (5) holds, which implies that the family of proposed policies $\{\pi_{th}^n, n \in \mathbb{N}\}$ is asymptotically compliant.

It follows from (28) and Theorem 3 that

$$\widehat{R}^n(t) \Rightarrow \theta \int_0^t \widehat{Q}_2(s)ds = \theta\mu_2 \int_0^t \left(\widehat{W}(s) - \frac{\lambda_1 d}{\mu_1}\right)^+ ds$$

as $n \to \infty$.

Hence, the lower bound in Theorem 2 is attained under $\{\pi_{th}^n, n \in \mathbb{N}\}$. Therefore, the family of the proposed policies is asymptotically optimal.