IISE Transactions

Quality & Reliability Engineering

IISE Transactions

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uiie21

Mind your own customers and ignore the others: Asymptotic optimality of a local policy in multi-class queueing systems with customer feedback

Jiankui Yang, Junfei Huang & Yunan Liu

To cite this article: Jiankui Yang, Junfei Huang & Yunan Liu (2021): Mind your own customers and ignore the others: Asymptotic optimality of a local policy in multi-class queueing systems with customer feedback, IISE Transactions, DOI: 10.1080/24725854.2021.1952358

To link to this article: https://doi.org/10.1080/24725854.2021.1952358

Published online: 01 Sep 2021.

IISE TRANSACTIONS
https://doi.org/10.1080/24725854.2021.1952358

Mind your own customers and ignore the others: Asymptotic optimality of a local policy in multi-class queueing systems with customer feedback

Jiankui Yang[a], Junfei Huang[b], and Yunan Liu[c]

[a]School of Science, Beijing University of Posts and Telecommunications, Beijing, China; [b]Department of Decision Sciences and Managerial Economics, CUHK Business School, The Chinese University of Hong Kong, Shatin, Hong Kong; [c]Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC, USA

ABSTRACT
This work contributes to the investigation of optimal routing and scheduling policies in multi-class multi-server queueing systems with customer feedback. We propose a new policy, dubbed *local policy* that requires access to only local queue information. Our new local policy specifies how an idle server chooses the next customer by using the queue length information of not all queues, but only those this server is eligible to serve. To gain useful insights and mathematical tractability, we consider a simple W model with customer feedback, and we establish limit theorems to show that our local policy is asymptotically optimal among all policies that may use the global system information, with the objective of minimizing the cumulative queueing costs measured by convex functions of the queue lengths. Numerical experiments provide convincing engineering confirmations of the effectiveness of our local policy for both W model and a more general non-W model.

ARTICLE HISTORY
Received 17 October 2020
Accepted 30 June 2021

KEYWORDS
Multi-class queueing systems; asymptotic optimality; customer feedback; local information; convex delay cost; $c\mu$-rule

1. Introduction

Stochastic Processing Networks (SPNs) have been widely used to model complex manufacturing, telecommunications, computer systems, and service networks; see Harrison (2000). General SPNs consist of flexible servers, multi-class arrivals, customer feedback, and a set of activities, where an activity refers to the processing of a class of customers by a specific server. In SPNs, two important decisions are: (i) *routing*: Paring a new arrival to one of the available servers; and (ii) *scheduling*: Paring a newly available server to the next waiting customer. In this article, we consider a multi-class multi-server queueing system with customer feedback; we study the optimal routing and scheduling decisions with the objective of minimizing the cumulative congestion cost (measured by a convex function of the queue lengths).

Customer feedback commonly occur in many service systems, such as healthcare, call centers, mobile networks, and computer systems (Tran-Gia and Mandjes, 1997; Yom-Tov and Mandelbaum, 2014; Wang *et al.* 2020). In realistic settings, a customer's statistical parameters (e.g., retrial probability, service rate, etc.) may indeed depend on the customer's entire service history (Liu and Whitt, 2017). Nevertheless, due to the complex nature of queueing systems with customer feedback, researchers often resort to approximating models with a Markovian feedback structure to gain model tractability (Yom-Tov and Mandelbaum, 2014).

Customer feedback may also have a significant influence on the design of an SPN's routing and scheduling policies, because when a server is deciding on which one of the next customers to serve, it has to take into account the customer's potential future behavior (e.g., which queue to join if the customer decides to retry for more service). This seems to suggest that it will be beneficial for the server to have access to the real-time state of the entire SPN (e.g., queue lengths and customer delay at all queues). However, real-time information retrieval in SPNs may be costly, slow, or inconvenient. For example, in emergency departments, it may be inconvenient to keep track of the status of patients waiting to be treated in other units. In web service systems and data centers, acquiring timely global system states requires high communication overhead (Lu *et al.*, 2011), so that researchers are motivated to develop policies that only use local information, such as join-the-idle-queue and the power-of-d policies (Vvedenskaya *et al.*, 1996; Lu *et al.*, 2011).

In this article, *we propose a new scheduling rule that is based on only local information, and we study its optimality*. To gain insights into the operational control for multi-class multi-server SPNs with customer feedback, we focus on a simple W model (see Figure 1) having three customer classes and two servers. Server 1 can serve class 1 and 2 customers, whereas server 2 can serve class 2 and 3 customers. After finishing the current service, a customer may be internally transferred to join another customer queue, or leave the system permanently.

A new "local" policy

Our proposed new policy, dubbed the *local policy*, is a generalized $c\mu$-type policy which requires access of queue length

CONTACT Junfei Huang junfeih@cuhk.edu.hk
All authors contributed to the manuscript equally.

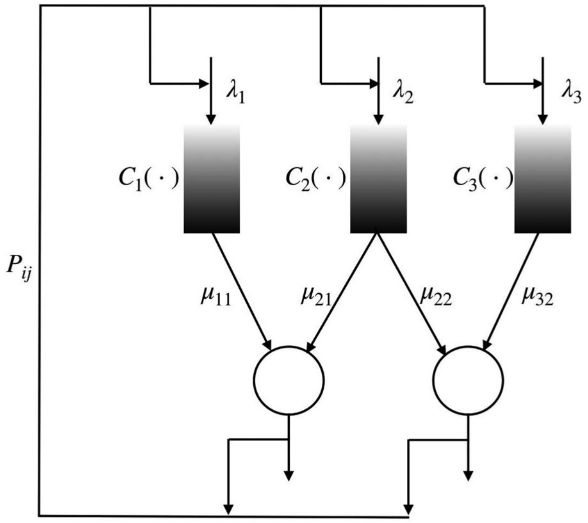

Figure 1. A W queueing system.

information of not all queues, but only those eligible to the deciding server. Specifically, let $C_i(\cdot)$ and Q_i be the cost rate function and queue length and $\mathcal{C}(j)$ be the set of customer classes that server j is eligible to serve. When server j is ready to serve a new customer, he/she will choose the head-of-line customer from class

$$i^* \in \text{argmax}_{i\in\mathcal{C}(j)} \frac{C_i'(Q_i(t))}{y_i^*}, \qquad (1)$$

where the constant y_i^* is some properly selected weight for class i customers (see (10) for details of y_i^*). The intuition behind this is that, by appropriately controlling local queues, we expect to achieve some local *State-Space Collapse* (SSC) result for each server. As class 2 customers serve as a connecting class, the local SSC results for both servers would then combine to form a global SSC (thus, though implicitly, class 3 customers influence the decision made on server 1's end, and so do class 1 customers on server 2's end). The idea is to make the two servers work collaboratively as a "super-server" (as in Mandelbaum and Stolyar (2004)), so that the scheduling policy is expected to play the same role as in a system with only one server (Huang *et al.*, 2015). Here, the properly designed weight $y^* = (y_i^*)$ measures the role of different customer classes in the "super-server" system.

Following this idea, we prove that, under the conventional heavy traffic condition, i.e., the traffic intensity is near one, this policy is asymptotically optimal among all "global" policies (policies that may use all queue lengths). Hence, our new local scheduling rule is not only conceptually simpler than the policy proposed in Mandelbaum and Stolyar (2004), it is also easier to implement in certain applications - especially those in which less information about queue lengths can be accessed.

1.1. Related literature

The literature on SPNs and parallel-server queueing systems is well established. We sketch related studies in the literature without the intention of being exhaustive. Analysis of the Brownian control problem introduced in Harrison (1988) initiates a series of papers suggesting "good" control policies of SPN, which are summarized in Bell and Williams (2001). Although there are various heuristic schemes in the extant literature, very few have been theoretically substantiated with rigorious asymptotic optimality analysis. Difficulties arise in two aspects: network topology and realistic cost/utility function. A specific SPN with simple topology (a single-server multi-class queueing system) is considered by Van Mieghem (1995), where the generalized $c\mu$-rule is proved to be asymptotically optimal. Over the past two decades, some progress has been made in complete resource pooling structured networks (Harrison and Lopez, 1999). Based on linear cost functions, Harrison (1998) proves asymptotic optimality in a two-server case. Harrison and Lopez (1999) extend Harrison (1998) to a general system and heuristically derive conditions for discrete-review optimality; whereas Ata and Kumar (2005) prove the asymptotic optimality for the stochastic network with feedback (activity-based routing). For a mean cumulative discounted cost of holding jobs in the system, Bell and Williams (2001) consider an "N" structured network and prove the asymptotic optimality of a threshold-type policy. Stolyar (2004) considers a discrete time generalized switch system with parallel servers, which even allows for arbitrary dependence between servers. The max-weight scheduling policy has been proved to be asymptotically optimal for a polynomial-type cost function. For a general SPN Dai and Lin (2008) propose a max-pressure policy, which is also asymptotically optimal for a quadratic holding cost under some mild assumptions. Ye and Yao (2008) focus on a broad class of utility-maximizing resource allocation schemes (which includes the generalized $c\mu$-rule and max-weight scheme as special cases) in a SPN and prove it to also be asymptotically optimal. Extending the system considered in Van Mieghem (1995) to a parallel-server system, Mandelbaum and Stolyar (2004) establish the asymptotic optimality of the generalized $c\mu$-rule. Furthermore, by allowing for customer feedback, Mandelbaum and Stolyar (2004) conjecture that another copy of the generalized $c\mu$-rule with feedback customers could be asymptotically optimal, which is left as an open question (see page 852). In the current article, instead of addressing this open question, we consider a W-structured parallel system with customer feedback and propose a "local" policy that works reasonably well.

1.2. Contributions and organization

The contributions of this article are as follows:

(a) **New policy using local queue length information.** We propose a novel control policy using only local queue length information in a W model, and we establish that our local policy is asymptotically optimal with the objective of minimizing the system's congestion cost (a convex function of the queue lengths). The practical relevance of our local policy is that it can reduce the information retrieval cost. Despite the simplicity of the present model, our results serve as an initial attempt to investigate efficient control policies only using local information in more general SPNs.

(b) **New method to establish SSC in the presence of customer feedback.** A key step to proving SSC (Bramson, 1998) is to establish the uniform attraction results for the hydrodynamic limit. However, with customer feedback, establishing the required uniform attraction results does not seem easy. Even for the W-structured system, existing methods in the literature may not work (see Example 1). In this work, we construct a four-step procedure by heavily exploring the W structure to prove the uniform attraction, and then the global SSC, that is, all three queue lengths are determined by a one-dimensional workload process. This, along with the standard sandwich method (Chen and Shanthikumar, 1994), will establish the diffusion limit of the queue lengths under the proposed policy and in turn conclude its asymptotic optimality.

Organization of the current article

The rest of this article is organized as follows: We introduce the model and system dynamics in Section 2 and the heavy traffic framework in Section 3. The main results are presented in Section 4 and proved in Section 5. Numerical experiments are performed in Section 6. We conclude in Section 7 with some discussions on the possibility of extending the local policy to more general SPNs. There we provide a non-W example to explain how a local policy may be implemented and how it performs; we also illustrate how the current proof of SSC may fail for this non-W model. Additional technical proofs are presented in the "Supplemental Online Materials".

Notation

We conclude this section with some notations and conventions. All vectors are understood to be column vectors. A superscript "T" of a matrix (vector) is interpreted as transpose. We use the standard notation $\mathbb{R}$ for the set of real numbers and $\mathbb{R}_+^n = \{x = (x_1, x_2, ..., x_n), x_i \geq 0, i = 1, 2, ..., n\}$. Denote by $\mathbb{D}[0,\infty)$ the space of $\mathbb{R}$-valued functions that are right continuous on $[0,\infty)$ with left limit on $(0,\infty)$. We use the symbol $\Rightarrow$ for weak convergence of elements in the space $\mathbb{D}[0,\infty)$. For a sequence of functions $f^r(\cdot) \in \mathbb{D}[0,\infty), f^r(t) \to f(t)$ u.o.c. as $r \to \infty$ means that $f^r(t)$ uniformly converges to $f(t)$ on compact sets as $r \to \infty$. We use "$\equiv$" for equality by definition.

2. The model and system dynamics

Consider a queueing system with three customer classes, indexed by $i \in \mathcal{I} = \{1,2,3\}$, and two server stations, indexed by $j \in \mathcal{J} = \{1,2\}$, where there is a single server at each server station. Throughout this article, we use "server" and "server station" interchangeably. The set of customer classes that can be served by server j is denoted by $\mathcal{C}(j)$ (that is, $\mathcal{C}(1) = \{1,2\}$ and $\mathcal{C}(2) = \{2,3\}$), and the set of servers that can serve customer class i is denoted by $\mathcal{S}(i)$ (hence $\mathcal{S}(1) = \{1\}, \mathcal{S}(2) = \{1,2\}$ and $\mathcal{S}(3) = \{2\}$). Among class i customers, the service principle is *First-Come First-Served* (FCFS). After finishing the current service, a customer may immediately return for further service (with a possible change of class), or leave the system permanently. See Figure 1 for an illustration of the W model.

General external arrival processes

Let λ_i denote the *exogenous arrival rate* for customer class i, and the vector $\lambda = (\lambda_i)_{i\in\mathcal{I}}$ is the exogenous arrival rate vector. We assume that there is at least one $\lambda_i, i \in \mathcal{I}$, being positive, i.e., $\sum_{i\in\mathcal{I}} \lambda_i > 0$. To model the interarrival times of external customers, for each $i \in \mathcal{I}$, define a sequence of *independent and identically distributed* (i.i.d.) random variables $\{u_i(n)\}$ with $\mathbb{E}[u_i(1)] = 1$ and $\text{Var}[u_i(1)] = \alpha_i^2 < \infty$. The random variable $u_i(n)/\lambda_i$ represents the exogenous inter-arrival time between the $(i-1)$th and ith exogenous arrivals.

Introduce the partial summation

$$U_i(0) = 0 \quad \text{and} \quad U_i(n) = \sum_{l=1}^{n} u_i(l), n \geq 1.$$

The exogenous arrival process is $E = \{(E_i(\lambda_i t), i \in \mathcal{I}); t \geq 0\}$ with

$$E_i(t) = \max\{n : U_i(n) \leq t\}.$$

The term $E_i(\lambda_i t)$ denotes the numbers of exogenous arrivals of class-i customers until time t.

General service processes

For each pair (i, j) with $j \in \mathcal{S}(i)$, let m_{ij} and $\mu_{ij} = 1/m_{ij}$ be the mean and rate of service time for a class-i customer served by server j. Define a sequence of i.i.d. random variables $\{v_{ij}(n)\}$ such that $\mathbb{E}[v_{ij}(1)] = 1$ and $\text{Var}[v_{ij}(1)] = \beta_{ij}^2 < \infty$ for $i \in \mathcal{I}, j \in \mathcal{S}(i)$. The random variable $m_{ij}v_{ij}(k)$ denotes the service time of the kth (external or internal) class-i customers served by server j.

Introduce the partial summation

$$V_{ij}(0) = 0 \quad \text{and} \quad V_{ij}(n) = \sum_{l=1}^{n} v_{ij}(l), n \geq 1,$$

and the renewal process $S = \{(S_{ij}(\mu_{ij}t), i \in \mathcal{I}, j \in \mathcal{S}(i)); t \geq 0\}$ with

$$S_{ij}(t) = \max\{n : V_{ij}(n) \leq t\}.$$

Then $S_{ij}(\mu_{ij}t)$ denotes the number of class-i customers completing service by server j if server j has devoted t time units to class-i customers.

Markovian feedback structure

For class-i customers, consider a sequence of i.i.d. random variables $\{\phi_i(n)\}$ with $\phi_i(n)$ recording which class the nth class-i customer transfers to upon service completion. Let e_l be a three-dimensional vector with the lth entry being one and the others being zeros; We let $P_{il} \equiv P(\phi_i(1) = e_l)$ be the probability that an old class-i customer transitions to a new class-l customer, and $P_{i,0} \equiv 1 - \sum_{l=1}^{3} P_{il} = P(\phi_i(1) = 0)$ be the probability that the customer leaves the system permanently. We assume the 3×3 sub-stochastic matrix (*transition matrix*) $P = (P_{il})$ has a spectral radius *strictly* less than one, which ensures that with probability 1, all customers will eventually leave the system. Introduce the process $\Phi = \{\Phi_i(n); n \geq 0\}$ with

$$\Phi_i(n) := \sum_{l=1}^{n} \phi_i(l),$$

which records the feedback process of the first n class-i customers upon finishing service. Let $\Phi_{il}(n)$ be the lth entry of $\Phi_i(n)$. We assume $\{\phi_i(n)\}$ is independent of all other primitive data. The aforementioned processes, $\{E_i(t), t \geq 0\}, \{S_{ij}(t), t \geq 0\}$ and $\{\phi_i(n), n \geq 0\}, i \in \mathcal{I}, j \in \mathcal{S}(i)$, are all assumed to be mutually independent.

In summary, the model parameters include $(\lambda_i, m_{ij}, P_{il}, i \in \mathcal{I}, j \in \mathcal{S}_{(i)}, l \in \mathcal{J})$.

2.1. System dynamics and objective function

The system dynamics depend on the scheduling policy. A scheduling policy is defined by

$$\pi \equiv \{T_{ij}(t); i \in \mathcal{I}, j \in \mathcal{S}(i), t \geq 0\},$$

where $T_{ij}(t)$ denotes the time that server j has spent on class-i customers in time interval $[0, t]$. Under a scheduling policy π, the number of class-i customers departing (leaving or transferring) from server j is

$$D_{ij}(t) \equiv S_{ij}(\mu_{ij} T_{ij}(t)), \tag{2}$$

and in the interval $[0, t]$, the number of class-i customers transferring to class-l customers is

$$\Phi_{il}\left(\sum_{j \in \mathcal{S}(i)} D_{ij}(t)\right).$$

A scheduling policy π is called *admissible* if the following equations hold:

$$\begin{aligned} &T_{ij}(\cdot) \text{is non decreasing and } T_{ij}(0) = 0, \quad \text{for } i \in \mathcal{I}, j \in \mathcal{S}(i),\\ &\sum_{i \in \mathcal{C}(j)} \left[T_{ij}(t) - T_{ij}(s)\right] \leq t - s, \quad \text{for } s < t, j \in \mathcal{J}. \end{aligned}$$

Note that an admissible policy needs not to be work-conserving. Denote by Π the set of all admissible policies.

For $i \in \mathcal{I}$, denote by $Q_i^\pi(t)$ the number of class-i customers at time t under a scheduling policy $\pi \in \Pi$. Then, we have

$$\begin{aligned} Q_i^\pi(t) = Q_i^\pi(0) + E_i(\lambda_i t) &+ \sum_{l \in \mathcal{I}} \Phi_{li}\left(\sum_{j \in \mathcal{S}(l)} D_{lj}(t)\right) \\ &- \sum_{j \in \mathcal{S}(i)} D_{ij}(t), \end{aligned} \tag{3}$$

for $i \in \mathcal{I}$,

$$Q_i^\pi(t) \geq 0, \quad \text{for } i \in \mathcal{I}, t \geq 0. \tag{4}$$

Following from (2)–(4), we further have

$$\begin{aligned} Q_i^\pi(t) = X_i^\pi(t) + \lambda_i t &+ \sum_{l \in \mathcal{I}} P_{li} \sum_{j \in \mathcal{S}(l)} \mu_{lj} T_{lj}(t) \\ &- \sum_{j \in \mathcal{S}(i)} \mu_{ij} T_{ij}(t) \geq 0, \quad \text{for } i \in \mathcal{I}, \end{aligned} \tag{5}$$

where

$$\begin{aligned} X_i^\pi(t) = Q_i^\pi(0) &+ E_i(\lambda_i t) - \lambda_i t \\ &+ \sum_{l \in \mathcal{I}} \left(\Phi_{li}\left(\sum_{j \in \mathcal{S}(l)} S_{lj}(\mu_{lj} T_{lj}(t))\right) - P_{li} \sum_{j \in \mathcal{S}(l)} S_{lj}(\mu_{lj} T_{lj}(t))\right) \\ &+ \sum_{l \in \mathcal{I}} P_{li}\left(\sum_{j \in \mathcal{S}(l)} (S_{lj}(\mu_{lj} T_{lj}(t)) - \mu_{lj} T_{lj}(t))\right) \\ &- \sum_{j \in \mathcal{S}(i)} (S_{ij}(\mu_{ij} T_{ij}(t)) - \mu_{ij} T_{ij}(t)). \end{aligned}$$

Writing (5) in matrix notations yields

$$Q^\pi(t) = X^\pi(t) + \lambda t - (I - P^T) R(t). \tag{6}$$

Here $R(t) \equiv (R_i(t), i \in \mathcal{I})$ with $R_i(t) \equiv \sum_{j \in \mathcal{S}(i)} \mu_{ij} T_{ij}(t)$.

We assume that class-i customers incur a queueing cost at rate $C_i(Q_i^\pi(t))$ at time t with the cost rate function C_i satisfying the condition below.

Assumption 1 (Cost functions). *The function C_i satisfies $C_i(0) = 0$ and $C_i'(0) = 0$ and is strictly increasing and convex.*

Then the cumulative total cost by time t is

$$\mathcal{U}_\pi(t) = \int_0^t \sum_{i=1}^{3} C_i(Q_i^\pi(s)) ds.$$

Our objective is to minimize the cumulative congestion cost for each $T > 0$, that is,

$$\min_{\pi \in \Pi} \mathcal{U}_\pi(T). \tag{P-1}$$

2.2. The scheduling rule: Local information vs. global information

A global policy

For more general SPNs with multiple customer classes, multiple servers and customer feedback, Mandelbaum and Stolyar (2004) proposed a promising scheduling policy, hereby dubbed the *MS policy*, of which the asymptotic optimality still remains an open problem. See Mandelbaum and Stolyar (2004) for detailed definition of the MS policy for more general SPNs. The MS policy, when applied to our W queueing system, reduces to the policy that server j serves a customer from class

$$i \in \text{argmax}_i \left[C_i'(Q_i(t)) - \sum_{k=1}^{3} P_{ik} C_k'(Q_k(t))\right] \mu_{ij}, \tag{7}$$

unless the maximum is nonpositive, in which case the server remains idle.

To implement the MS policy in (7), one is required to have the global information of the SPN, that is, access to the queue lengths of all customer classes. In the following, we propose a generalized $c\mu$-type policy with μ appropriately modified. Comparing with the "global" MS policy, our new policy not only has a simpler structure, but also uses only "local" information, in the sense that each server needs only to have access to queue lengths of customer classes the server is eligible to serve, not to all customer classes. It is straightforward to see that the set of admissible policies using only local information is a strict subset of all admissible policies Π.

A "local" policy

Recall that μ_{ij} is the service rate of customer class $i \in \mathcal{I}$ by server $j \in \mathcal{S}(i)$. Define a constant vector ν^* such that

$$\nu^* = \left(1, \frac{\mu_{11}}{\mu_{21}}, \frac{\mu_{11}}{\mu_{21}} \times \frac{\mu_{22}}{\mu_{32}}\right)^T. \tag{8}$$

It is straightforward to see that, with ν^* defined in (8), we have

$$\nu_1^*\mu_{11} = \nu_2^*\mu_{21} \quad \text{and} \quad \nu_2^*\mu_{22} = \nu_3^*\mu_{32}. \tag{9}$$

As $\nu_1^*/\nu_2^* = m_{11}/m_{21}$ and $\nu_2^*/\nu_3^* = m_{22}/m_{32}$, $\nu_1^* : \nu_2^* : \nu_3^*$ can be viewed as the ratio of the workload, or expected service times by eliminating the role of different servers. Indeed, the average service time is determined by two factors: the customers' average workload and the server's speed (or capacity). For example, if class 1 customers have an average workload of 2 and server 1's speed is 1 workload per unit time, then $m_{11} = 2$; similarly, if class 3 customers have an average workload of 0.5 and server 2's speed is 0.5 workload per unit time, then $m_{32} = 1$. If class 2 customers have an average workload of 1, then $m_{21} = 1$ and $m_{22} = 2$. We can calculate that $\nu_1^* : \nu_2^* : \nu_3^* = 1 : 0.5 : 0.25 = 2 : 1 : 0.5$. Note that this is exactly the ratio of average workloads.

We can define $z_j^* = \nu_i^*\mu_{ij}$ for $j \in \mathcal{J}$ and any $i \in \mathcal{S}(j)$. Let

$$y^* = (I - P)^{-1}\nu^*. \tag{10}$$

Consequently, y^* can be understood as the *effective expected service times*, which have taken into account the potential future service times due to feedback. The queue lengths weighted by y^* represent "workloads" (see (21) for details).

We propose the following policy, denoted by π_*, which has a form similar to the generalized cμ-rule introduced by Van Mieghem (1995), except that we use $1/y_i^*$, instead of μ_i for $i \in \mathcal{I}$. Note that the transition probability matrix P is implicitly used via y^* in the policy. We specify π_* below:

(a) **Routing:** When a customer arrives (either external or internal), the customer joins any available server is eligible to serve him/her if there are any; otherwise the customer joins the waiting queue.

(b) **Scheduling:** When server j finishes serving a customer and there exist waiting customers that he/she is eligible to serve, the server chooses a customer from the class

$$i \in \operatorname{argmax}_{i\in\mathcal{C}(j)} \frac{C_i'(Q_i(t))}{y_i^*}. \tag{11}$$

As policy π_* is locally work-conserving, we have for $j \in \mathcal{J}$,

$$\int_0^\infty \mathbf{1}_{\left\{\sum_{i\in\mathcal{C}(j)} Q_i(s)>0\right\}} d\left(s - \sum_{i\in\mathcal{C}(j)} T_{ij}(s)\right) = 0, \tag{12}$$

where $\mathbf{1}_A$ is the indicator of event A. Also, it is easy to verify, based on the second part of π_*, that for $j \in \mathcal{J}$,

$$\int_0^\infty \left(\max_{l\in\mathcal{C}(j)} \frac{C_l'(Q_l(t))}{y_l^*} - \frac{C_i'(Q_i(t))}{y_i^*}\right) dT_{ij}(s) = 0. \tag{13}$$

We will prove that π_* is asymptotically optimal among all policies in Π.

3. Heavy traffic framework

The notion of asymptotic optimality requires the construction of a sequence of W queueing systems that have the same structure as in Section 2. The sequence of W models are indexed by $r \uparrow \infty$, where r measures the system's scale. For the rth system, we append a superscript "r" to all notations: The exogenous arrival process is $E^r = \{(E_i^r(t), i \in \mathcal{I}); t \geq 0\}$ with $E_i^r(t) = E_i(\lambda_i^r t)$; the service process is $S^r = \{(S_{ij}^r(t), i \in \mathcal{I}, j \in \mathcal{S}(i)); t \geq 0\}$ with $S_{ij}^r(t) = S_{ij}(\mu_{ij}^r t)$, and the transition sequence is $\Phi^r = \{(\Phi_{il}^r(n), i, l \in \mathcal{I}); n = 0, 1, 2, \ldots\}$. The arrival rate vector is λ^r. We assume the customer classes with positive arrival rates are invariant to r. Without loss of generality, we assume the transition sequence Φ^r and $\{\mu_{ij}^r, i \in \mathcal{I}, j \in \mathcal{S}(i)\}$ are invariant to r, hence, we can omit the superscript r for these two processes. Similarly, all other quantities associated to the rth system will be appended with a superscript r.

We assume the sequence of queueing systems satisfies the following assumption:

Assumption 2 (Heavy traffic assumption). *A sequence of queueing systems satisfies the heavy traffic assumption if*

(a) *There is a three-dimensional vector $\lambda = (\lambda_i) \geq 0$ and a four-dimensional vector $\mu = (\mu_{ij})$, such that the vector λ^e defined via*

$$\lambda^e = (I - P^T)^{-1}\lambda \tag{14}$$

satisfies

$$\lambda_1^e < \mu_{11}, \quad \lambda_3^e < \mu_{32},$$

$$\text{and} \quad \lambda_2^e = \left(1 - \frac{\lambda_1^e}{\mu_{11}}\right)\mu_{21} + \left(1 - \frac{\lambda_3^e}{\mu_{32}}\right)\mu_{22}.$$

(b) *When $r \to \infty, \alpha_i^r \to \alpha_i$ and $\beta_{ij}^r \to \beta_{ij}$ for some $\alpha_i \geq 0$ and $\beta_{ij} \geq 0$,*

$$\lambda_i^r \to \lambda_i, \quad i \in \mathcal{I}, \quad \text{and} \quad r\sum_{i=1}^3 y_i^*(\lambda_i^r - \lambda_i) \to \gamma,$$

for some $\gamma \in \mathbb{R}$.

Remark 1. *Under the aforementioned setting, the exogenous interarrival times are $\{u_i(n)/\lambda_i^r\}$ and the service times of class-i customers by server j are $\{v_{ij}(n)/\mu_{ij}^r\}$. Thus, it is easy to verify that, under the above heavy traffic assumption, there* *exists a function $g(a)$, such that $g(a) \to 0$ as $a \to \infty$ and*

$$\mathbb{E}\left[\frac{u_i(n)}{\lambda_i^r}\mathbf{1}_{\left\{\frac{u_i(n)}{\lambda_i^r}\geq a\right\}}\right] \leq g(a) \quad \text{and} \quad \mathbb{E}\left[\frac{v_{ij}(n)}{\mu_{ij}^r}\mathbf{1}_{\left\{\frac{v_{ij}(n)}{\mu_{ij}^r}\geq a\right\}}\right] \leq g(a). \tag{15}$$

Note that these two inequalities are needed for establishing the SSC, following the framework of Bramson (1998).

We define x_{ij}^* as the nominal capacity allocation of server j to customer class i, specifically,

$$x_{11}^* \equiv \frac{\lambda_1^e}{\mu_{11}}, \quad x_{32}^* \equiv \frac{\lambda_3^e}{\mu_{32}}, \quad x_{21}^* \equiv 1 - x_{11}^*, \quad \text{and} \quad x_{22}^* \equiv 1 - x_{32}^*. \tag{16}$$

Then $x_{11}^* < 1, x_{32}^* < 1$ and

$$\lambda_i^e = \sum_{j\in\mathcal{S}(i)} x_{ij}^*\mu_{ij}, \quad \text{for} \quad i \in \mathcal{I}. \tag{17}$$

Writing (17) in a matrix form and using (14), one can show that (17) is equivalent to

$$\lambda_i + \sum_{l\in\mathcal{I}} P_{li} \sum_{j\in\mathcal{S}(l)} x_{lj}^*\mu_{lj} = \sum_{j\in\mathcal{S}(i)} x_{ij}^*\mu_{ij}, \quad \text{for} \quad i \in \mathcal{I}. \tag{18}$$

The above two equations (17) and (18) exhibit the input–output relationship: the right-hand side is the output rate with λ_i^e in (17) interpreted as the effective input rate; it is equal to the left-hand side of (18).

Under a policy $\pi^r = \{T_{ij}^r(t); i \in \mathcal{I}, j \in \mathcal{S}(i), t \geq 0\}$, for $i \in \mathcal{I}, j \in \mathcal{S}(i)$, we define the centered busy-time process $I_{ij}^r \equiv \{I_{ij}^r(t); t \geq 0\}$ with $I_{ij}^r(t) \equiv x_{ij}^* t - T_{ij}^r(t)$. Then

$$\sum_{i\in\mathcal{C}(j)} I_{ij}^r(t) = t - \sum_{i\in\mathcal{C}(j)} T_{ij}^r(t). \tag{19}$$

Following from (5) and (18), we have

$$Q_i^{r,\pi^r}(t) = X_i^{r,\pi^r}(t) - \sum_{l\in\mathcal{I}} P_{li} \sum_{j\in\mathcal{S}(l)} \mu_{lj} I_{lj}^r(t) + \sum_{j\in\mathcal{S}(i)} \mu_{ij} I_{ij}^r(t),$$

where

$$\begin{aligned} X_i^{r,\pi^r}(t) = {} & Q_i^{r,\pi^r}(0) + \lambda_i^r t - \lambda_i t + E_i^r(\lambda_i^r t) - \lambda_i^r t \\ & + \sum_{l\in\mathcal{I}} \left(\Phi_{li}\left(\sum_{j\in\mathcal{S}(l)} S_{lj}^r(\mu_{lj} T_{lj}^r(t)) \right) - P_{li} \sum_{j\in\mathcal{S}(l)} S_{lj}^r(\mu_{lj} T_{lj}^r(t)) \right) \\ & + \sum_{l\in\mathcal{I}} P_{li} \left(\sum_{j\in\mathcal{S}(l)} (S_{lj}^r(\mu_{lj} T_{lj}^r(t)) - \mu_{lj} T_{lj}^r(t)) \right) \\ & - \sum_{j\in\mathcal{S}(i)} (S_{ij}^r(\mu_{ij} T_{ij}^r(t)) - \mu_{ij} T_{ij}^r(t)). \end{aligned}$$

Rewriting the above in matrix notation, we have

$$Q^{r,\pi^r}(t) = X^{r,\pi^r}(t) + (I - P^T) Z^r(t), \tag{20}$$

with the ith component of $Z^r(t)$ defined as $Z_i^r(t) \equiv \sum_{j\in\mathcal{S}(i)} \mu_{ij} I_{ij}^r(t)$. In the following, we remove superscript π^r for ease of notation; all processes are specified under policy π^r.

Introduce the *workload process* $W^r \equiv \{W^r(t), t \geq 0\}$ with $W^r(t)$ defined by

$$W^r(t) \equiv (y^*)^T Q^r(t). \tag{21}$$

With $y^* \equiv (I - P)^{-1}\nu^*$ and $Y^r(t) \equiv (\nu^*)^T Z^r(t)$, we write

$$W^r(t) = (y^*)^T X^r(t) + Y^r(t).$$

Following from (9) and (19), we get

$$\begin{aligned} Y^r(t) &= \sum_{i\in\mathcal{I}} \nu_i^* \sum_{j\in\mathcal{S}(i)} \mu_{ij} I_{ij}^r(t) = \sum_{j\in\mathcal{J}} z_j^* \sum_{i\in\mathcal{C}(j)} I_{ij}^r(t) \\ &= \sum_{j\in\mathcal{J}} z_j^* \left(t - \sum_{i\in\mathcal{C}(j)} T_{ij}^r(t) \right). \end{aligned}$$

Under any admissible policy, $Y^r(\cdot)$ is a nondecreasing process.

Diffusion-scaled processes

We define the diffusion-scaled processes $(\widehat{W}^r, \widehat{Y}^r, \widehat{Q}^r) \equiv \{(\widehat{W}^r(t), \widehat{Y}^r(t), \widehat{Q}_i^r(t),\ i \in \mathcal{I}), t \geq 0\}$, where

$$\begin{aligned} &\widehat{W}^r(t) \equiv \frac{1}{r} W^r(r^2 t), \quad \widehat{Y}^r(t) \equiv \frac{1}{r} Y^r(r^2 t), \\ &\text{and} \quad \widehat{Q}_i^r(t) \equiv \frac{1}{r} Q_i^r(r^2 t). \end{aligned} \tag{22}$$

Class-i customers incur a cost at rate $C_i(\widehat{Q}_i^r(t))$ at time t, with C_i satisfying Assumption 1. Then the cumulative cost incurred until time t is

$$\mathcal{U}_\pi^r(t) = \int_0^t \sum_{i=1}^3 C_i(\widehat{Q}_i^r(s)) ds.$$

The objective of the rth problem is to stochastically minimize $\mathcal{U}_\pi^r(t)$, for each $t > 0$, by choosing an appropriate policy π among Π^r, the set of all admissible policies for the rth system.

A sequence of local policies

We adopt the local policy in Section 2.2 in each of the sequence of W queueing systems. Let $\pi_* \equiv \{\pi_*^r\}$, with π_*^r defined below for the rth system:

(a) **Routing:** When a customer arrives (external or internal), the customer chooses any available server that is eligible.

(b) **Scheduling:** When server j finishes serving a customer and there remains waiting customers that server j is eligible to serve, it chooses a customer from the class

$$\arg\max_{i\in\mathcal{C}(j)} \frac{C_i'(\widehat{Q}_i^r(t))}{y_i^*}.$$

If there is a tie, the server can arbitrarily break the tie.

In the next section, we will show that this sequence of local policies can asymptotically minimize the cumulative cost (in a stochastic order sense), and hence, is asymptotically optimal.

4. Diffusion limit and asymptotic optimality

We next introduce a pair of mappings $(h(\cdot), q^*(\cdot)) : \mathbb{R}_+ \to \mathbb{R}_+^4$. Specifically, for $x \geq 0$, let

$$\begin{aligned} h(x) \equiv \min & \sum_{i\in\mathcal{I}} C_i(q_i) \\ \text{s.t.} & \sum_{i\in\mathcal{I}} y_i^* q_i = x, \text{ and } q_i \geq 0. \end{aligned} \tag{23}$$

From Assumption 1 and the Karush–Kuhn–Tucker condition, for each $x \geq 0$, h is well defined and there is a unique $q^* = (q_i^*, i \in \mathcal{I})$ satisfying $h(x) = \sum_{i\in\mathcal{I}} C_i(q_i^*)$ and

$$\frac{C'_{i_1}(q^*_{i_1})}{y^*_{i_1}} = \frac{C'_{i_2}(q^*_{i_2})}{y^*_{i_2}}, \quad \text{for} \quad i_1, i_2 \in \mathcal{I}.$$

We also denote by $q^* = q^*(x)$ because q^* depends on x. Denote the one-dimensional manifold $\mathcal{M}^* = \{q^*(x); x \geq 0\}$ and call a point in $\mathcal{M}^*$ a *fixed point*. We can see that $q_i^*(x)$ is increasing and nonzero for $x > 0$ and all $i \in \mathcal{I}$. Since $\sum_{i\in\mathcal{I}} y_i^*(q_i^*(x+\Delta x) - q_i^*(x)) = \Delta x$ and $y_i^* > 0$ for $i \in \mathcal{I}, q^*(x)$ is Lipschitz continuous (with the Lipschitz constant no greater than $1/\min_{i\in\mathcal{I}} y_i^*$).

Assumption 3 (Initial condition). *Assume that*

$$\widehat{Q}^r(0) \Rightarrow \widehat{Q}(0), \quad \textit{as} \quad r \to \infty,$$

for a random vector $\widehat{Q}(0)$ *with* $P(\widehat{Q}(0) \in \mathcal{M}^*) = 1$.

Remark 2. *Different from the initial queue length condition enforced in Mandelbaum and Stolyar (2004) which plays an critical role in asymptotic optimality, our Assumption 3 is imposed only to simplify the analysis; see discussions in the online appendix on how to relax the above condition.*

Let $\widehat{X}^*$ be a Brownian motion with initial value $(y^*)^T\widehat{Q}(0)$, drift rate γ, and variance $(y^*)^T\Gamma y^*$ with

$$\begin{aligned} \Gamma = {} & \text{diag}(\lambda_i\alpha_i^2) + \sum_{l\in\mathcal{I}} \sum_{j\in\mathcal{S}(l)} \beta_{lj}^2 \mu_{lj} x_{lj}^* \times (\Gamma_{ij}^l) \\ & + (I - P^T)\text{diag}\left(\sum_{j\in\mathcal{S}(i)} \beta_{ij}^2 \mu_{ij} x_{ij}^*\right)(I - P), \end{aligned}$$

in which

$$\Gamma_{ij}^l = \begin{cases} P_{li}(1 - P_{li}), & \text{if} \quad i = j, \\ -P_{li}P_{lj}, & \text{if} \quad i \neq j. \end{cases}$$

Define

$$(\widehat{W}^*, \widehat{Y}^*) = (\varphi, \psi)(\widehat{X}^*).$$

Here, (φ, ψ) is the standard (one-dimensional) reflection mapping (see Section 6.2 in Chen and Yao (2001)). We next present our main result.

Theorem 1 (Asymptotic optimality of the local policy). *Assume Assumptions 1–3 hold. Then,*

(a) *Under the sequence of policy $\pi_* = \{\pi_*^r\}$, the following holds when $r \to \infty$*

$$(\widehat{W}^r, \widehat{Y}^r, \widehat{Q}^r) \Rightarrow (\widehat{W}^*, \widehat{Y}^*, \widehat{Q}^*),$$

with $\widehat{Q}^* = q^*(\widehat{W}^*)$. As a result, for any $t > 0$, as $r \to \infty$

$$\mathcal{U}_{\pi_*}^r(t) \Rightarrow \int_0^t h(\widehat{W}^*(s))ds.$$

(b) *The sequence of local policies $\pi_* = \{\pi_*^r\}$ is asymptotically optimal in the sense that for any sequence of admissible policies $\pi = \{\pi^r\}$ and any $t > 0$, $x > 0$,*

$$\liminf_{r\to\infty} P\big(\mathcal{U}_\pi^r(t) > x\big) \geq \lim_{r\to\infty} P\big(\mathcal{U}_{\pi_*}^r(t) > x\big).$$

Remark 3 (From local SSC to global SSC). *First, it is not hard to understand that the local policy helps achieve the local SSC. Next, to establish the global SSC, we note that the relationship $\widehat{Q}^* = q^*(\widehat{W})$ guarantees that the behavior of $\widehat{Q}^r$ is close to the one-dimensional manifold process $\mathcal{M}^*$. Therefore, we can link one queue length (e.g., Q_1) to another (e.g., Q_3), even though they are not directly controlled by the same server. In this way we are able to achieve a global SSC among all queue length processes. (The connecting class 2 plays an important role to keep two servers collaboratively acting as one "super-server", even though each server only keeps track of its own customers and ignores those in other classes.)*

Remark 4 (Generality of the stochastic ordering result). *Part (b) of the Theorem claims that asymptotically, $\mathcal{U}_\pi^r$, under any other admissible policy, is stochastically larger than $\mathcal{U}_{\pi_*}^r$ on any finite horizon. We emphasize that this notion of asymptotic optimality is quite general. For example, under some additional regularity conditions (e.g., cost functions $C_i(\cdot)$ are required to be bounded), it is straightforward to obtain asymptotic optimality results in other forms, such as the most predominantly considered mean congestion cost, that is,*

$$\begin{aligned} \liminf_{r\to\infty} E\big[\mathcal{U}_\pi^r(t)\big] & \geq \lim_{r\to\infty} E\big[\mathcal{U}_{\pi_*}^r(t)\big] \\ & = \int_0^t E\Big[h(\widehat{W}^*(s))\Big] ds, \qquad \text{as } r \to \infty. \end{aligned}$$

5. Proof of theorem 1

5.1. Preliminaries

Using the Skorohod representation theorem (Theorem 5.1 in Chen and Yao (2001)), we will adopt the sample path approach. That is, we assume all primitive processes are defined in a probability space such that the weak convergence becomes the almost sure u.o.c. convergence. As accessory expressions for the diffusion-scaled processes, we first introduce the fluid-scaled processes $\bar{\bar{Q}}^r = \{(\bar{\bar{Q}}_i^r(t), i \in \mathcal{I}), t \geq 0\}, \bar{\bar{E}}^r \equiv \{(\bar{\bar{E}}_i^r(t), i \in \mathcal{I}), t \geq 0\}, \bar{\bar{S}}^r \equiv \{(\bar{\bar{S}}_{ij}^r(t), i \in \mathcal{I}, j \in \mathcal{S}(i)), t \geq 0\}, \bar{\bar{T}}^r \equiv \{(\bar{\bar{T}}_{ij}^r(t), i \in \mathcal{I}, j \in \mathcal{S}(i)), t \geq 0\}$ and $\bar{\bar{\Phi}}^r \equiv \{(\bar{\bar{\Phi}}_{il}^r(t), i, l \in \mathcal{I}), t \geq 0\}$ with

$$\begin{aligned} & (\bar{\bar{Q}}_i^r(t), \bar{\bar{Q}}_i^r(t), \bar{\bar{Q}}_{ij}^r(t), \bar{\bar{Q}}_{ij}^r(t), \bar{\bar{Q}}_{il}^r(t)) \\ & \quad \equiv \frac{1}{r^2}(Q^r(\lambda_i^r r^2 t), E(\lambda_i^r r^2 t), S_{ij}(\mu_{ij} r^2 t), T_{ij}^r(r^2 t), \Phi_{il}(\lfloor r^2 t \rfloor)), \end{aligned}$$

We define the diffusion-scaled processes $\widehat{E}^r \equiv \{(\widehat{E}_i^r(t), i \in \mathcal{I}), t \geq 0\}, \widehat{S}^r \equiv \{(\widehat{S}_{ij}^r(t), i \in \mathcal{I}, j \in \mathcal{S}(i)), t \geq 0\}$, $\widehat{\Phi}^r \equiv \{(\widehat{\Phi}_{il}^r(t), i, l \in \mathcal{I}), t \geq 0\}$ and $\widehat{X}^r \equiv \{(\widehat{X}_i^r(t), i \in \mathcal{I}), t \geq 0\}$, with

$$\widehat{E}_i^r(t) \equiv \frac{1}{r}(E(\lambda_i^r r^2 t) - \lambda_i^r r^2 t), \quad i \in \mathcal{I},$$
$$\widehat{S}_{ij}^r(t) \equiv \frac{1}{r}(S_{ij}(\mu_{ij} r^2 t) - \mu_{ij} r^2 t), \quad i \in \mathcal{I}, j \in \mathcal{S}(i),$$
$$\widehat{\Phi}_{il}^r(t) \equiv \frac{1}{r}(\Phi_{il}(\lfloor r^2 t \rfloor) - P_{il}\lfloor r^2 t \rfloor), \quad i, l \in \mathcal{I},$$
$$\widehat{X}_i^r(t) \equiv \frac{1}{r} X_i^r(r^2 t), \quad i \in \mathcal{I}.$$

More specifically, for the last expression

$$\begin{aligned}\widehat{X}_i^r(t) = {} & \widehat{Q}_i^r(0) + r(\lambda_i^r - \lambda_i)t + \widehat{E}_i^r(t) \\ & + \sum_{l \in \mathcal{I}} \widehat{\Phi}_{li}^r\Big(\sum_{j \in \mathcal{S}(l)} \bar{\bar{S}}_{lj}^r(\bar{\bar{T}}_{lj}^r(t))\Big) + \sum_{l \in \mathcal{I}} P_{li} \sum_{j \in \mathcal{S}(l)} \widehat{S}^r(\bar{\bar{T}}_{lj}^r(t)) \\ & - \sum_{j \in \mathcal{S}(i)} \widehat{S}_{ij}^r(\bar{\bar{T}}_{ij}^r(t)).\end{aligned} \tag{24}$$

Then, we assume that when $r \to \infty$, with probability 1, $\widehat{Q}^r(0) \to \widehat{Q}(0)$ and

$$(\widehat{E}^r, \widehat{S}^r, \widehat{\Phi}^r) \to (\widehat{E}, \widehat{S}, \widehat{\Phi}), \quad \text{u.o.c.}$$

In our analysis, we will fix any sample such that the above convergence holds. With the definition of $\widehat{W}^r(t), \widehat{Y}^r(t)$ in (22) we have

$$\widehat{W}^r(t) = (y^*)^T \widehat{X}^r(t) + \widehat{Y}^r(t). \tag{25}$$

5.2. Hydrodynamic model

Lemma 5.1 (Lemma 1 in Chen and Ye (2012)). *Let t^* and u^* be any given time lengths, and assume condition* (15). *Then the following convergence holds with probability one: as $r \to \infty$,*

$$\sup_{0 \le t \le rt^*} \sup_{0 \le u \le u^*} |(\widehat{E}_i^r((t+u)/r) - \widehat{E}_i^r(t/r))| \to 0, i \in \mathcal{I},$$
$$\sup_{0 \le t \le rt^*} \sup_{0 \le u \le u^*} |(\widehat{S}_{ij}^r((t+u)/r) - \widehat{S}_{ij}^r(t/r))| \to 0, i \in \mathcal{I}, j \in \mathcal{S}(i),$$
$$\sup_{0 \le t \le rt^*} \sup_{0 \le u \le u^*} |(\widehat{\Phi}_{il}^r((t+u)/r) - \widehat{\Phi}_{il}^r(t/r))| \to 0, i, l \in \mathcal{I}.$$

We consider a time interval $[\tau, \tau+\delta]$, where $\tau \ge 0$ 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 $\lceil r\delta/T \rceil$ segments with equal length T/r (except the last one). Next, we use the hydrodynamic model in Bramson (1998) (also refer to, e.g., Stolyar (2004) and Ye and Yao (2008)), introduced below, to investigate the diffusion-scaled processes:

$$\bar{W}^{r,\ell}(u) \equiv \widehat{W}^r\left(\tau + \frac{\ell T + u}{r}\right), \tag{26}$$

$$\begin{aligned}\bar{Q}^{r,\ell}(u) &\equiv \widehat{Q}^r\left(\tau + \frac{\ell T + u}{r}\right), \\ \bar{X}^{r,\ell}(u) &\equiv \widehat{X}^r\left(\tau + \frac{\ell T + u}{r}\right), \\ \bar{E}_i^{r,\ell}(u) &\equiv \widehat{E}_i^r\left(\tau + \frac{\ell T + u}{r}\right) - \widehat{E}_i^r\left(\tau + \frac{\ell T}{r}\right) + \lambda_i u, i \in \mathcal{I}, \\ \bar{S}_{ij}^{r,\ell}(u) &\equiv \widehat{S}_{ij}^r\left(\tau + \frac{\ell T + u}{r}\right) - \widehat{S}_{ij}^r\left(\tau + \frac{\ell T}{r}\right) + \mu_{ij} u, i \in \mathcal{I}, j \in \mathcal{S}(i), \\ \bar{\Phi}_{il}^{r,\ell}(u) &\equiv \widehat{\Phi}_{il}^r\left(\tau + \frac{\ell T + u}{r}\right) - \widehat{\Phi}_{il}^r\left(\tau + \frac{\ell T}{r}\right) + P_{il} u, i, l \in \mathcal{I}, \\ \bar{T}^{r,\ell}(u) &\equiv \widehat{T}^r\left(\tau + \frac{\ell T + u}{r}\right) - \widehat{T}^r\left(\tau + \frac{\ell T}{r}\right), \\ \bar{Y}^{r,\ell}(u) &\equiv \widehat{Y}^r\left(\tau + \frac{\ell T + u}{r}\right) - \widehat{Y}^r\left(\tau + \frac{\ell T}{r}\right),\end{aligned} \tag{27}$$

for $u \ge 0$ and $\ell = 0, 1, \ldots, \lfloor r\delta/T \rfloor$.

As in Chen and Ye (2012), Lemma 5.1 guarantees the following convergence:

$$\bar{E}_i^{r,\ell_r}(t) \to \lambda_i t, \text{u.o.c.} \tag{28}$$

$$\bar{S}_{ij}^{r,\ell_r}(t) \to \mu_{ij} t, i \in \mathcal{I}, j \in \mathcal{S}(i), \text{u.o.c.} \tag{29}$$

$$\bar{\Phi}_{il}^{r,\ell_r}(t) \to P_{il} t, i, l \in \mathcal{I}, \text{u.o.c.} \tag{30}$$

Given this convergence, we can prove the following lemma, which is simply a modification of Theorem 2 and Proposition 3 in Chen and Ye (2012) to the systems with customer feedback.

Lemma 5.2 (Hydrodynamic limits). *Let M be a given positive constant and ℓ_r be some integer in $[0, r\delta/T]$. Suppose $|\bar{Q}^{r,\ell_r}(0)| \le M$ for sufficiently large r. Then, under the proposed family of control policies, almost surely, for any sequence of $\{r\}$ there exists a further subsequence, denoted by $\mathcal{R}$, such that along $\mathcal{R}$, the hydrodynamic-scaled processes $(\bar{W}^{r,\ell_r}, \bar{Q}^{r,\ell_r}, \bar{Y}^{r,\ell_r}, \bar{T}^{r,\ell_r})$ converge uniformly on compact time intervals to limit process $(\bar{W}, \bar{Q}, \bar{Y}, \bar{T})$, which satisfies the following equations*:

$$\bar{Q}(t) = \bar{Q}(0) + \lambda t - (I - P^T)\bar{D}(t), \tag{31}$$

$$\sum_{i \in \mathcal{C}(j)} \bar{T}_{ij}(t) - [\bar{T}_{ij}(s)] \le t - s, \quad \text{for} \quad s < t, j \in \mathcal{J}, \tag{32}$$

$$\bar{W}(t) = (y^*)^T \bar{Q}(t) = \bar{W}(0) + \bar{Y}(t) \tag{33}$$

$$\bar{Y}(t) = \sum_{j \in \mathcal{J}} z_j^* \left(t - \sum_{i \in \mathcal{C}(j)} T_{ij}(t)\right), \tag{34}$$

$$\int_0^\infty \mathbf{1}_{\{\sum_{i \in \mathcal{C}(j)} \bar{Q}_i(s) > 0\}} d\left(s - \sum_{i \in \mathcal{C}(j)} \bar{T}_{ij}(s)\right) = 0, \quad j \in \mathcal{J}, \tag{35}$$

$$\int_0^\infty \left(\max\nolimits_{l \in \mathcal{C}(j)} \frac{C_l'(\bar{Q}_l(t))}{y_l^*} - \frac{C_i'(\bar{Q}_i(t))}{y_i^*}\right)^+ d\bar{T}_{ij}(s) = 0, \quad j \in \mathcal{J} \quad \text{and} \quad i \in \mathcal{C}(j). \tag{36}$$

In (31), $\bar{D}(t) = (\bar{D}_i(t), i \in \mathcal{I})$ with $\bar{D}_i(t) = \sum_{j \in \mathcal{S}(i)} \mu_{ij} \bar{T}_{ij}(t)$.

Remark 5. *We call any $(\bar{W}, \bar{Q}, \bar{Y}, \bar{T})$ satisfying (31)–(36) a hydrodynamic model solution. One can prove that any hydrodynamic model solution is Lipschitz, hence, absolutely continuous and differentiable almost everywhere. We refer to such time points t as regular points (and adopt the convention that $t = 0$ is not a regular point).*

5.3. Uniform attraction

Denote by

$$\mathcal{I}^*(t) \equiv \left\{ i \in \mathcal{I} \middle| \frac{C_i'(\bar{Q}_i(t))}{y_i^*} = \max_{i \in \mathcal{I}} \frac{C_i'(\bar{Q}_i(t))}{y_i^*} \right\},$$

and

$$\mathcal{I}_*(t) \equiv \left\{ i \in \mathcal{I} \middle| \frac{C_i'(\bar{Q}_i(t))}{y_i^*} = \min_{i \in \mathcal{I}} \frac{C_i'(\bar{Q}_i(t))}{y_i^*} \right\}.$$

Writing in a matrix form, (31) is equivalent to

$$\bar{L}(t) \equiv (I - P^T)^{-1}\bar{Q}(t) = (I - P^T)^{-1}\bar{Q}(0) + \lambda^e t - \bar{D}(t).$$

Lemma 5.3. *There exists a constant $\epsilon_1 > 0$, such that if $\bar{Q}(t)$ is not a fixed point, then*

$$\sum_{i\in\mathcal{I}^*(t)} \nu_i^* \bar{L}_i'(t) \le -\epsilon_1, \qquad \sum_{i\in\mathcal{I}_*(t)} \nu_i^* \bar{L}_i'(t) \ge \epsilon_1.$$

This is a generalization of Lemma 4 in Mandelbaum and Stolyar (2004) to systems with customer feedback. The proof idea is similar to the one there. We provide the proof of the second inequality in the "Supplemental Online Materials" as an illustration.

Introduce

$${}_*\bar{Q}_i(t) \equiv \{\xi \ge 0 | C_i'(\xi)/y_i^* = \min_{i\in\mathcal{I}} C_i'(\bar{Q}_i(t))/y_i^*\}$$

and

$${}^*\bar{Q}_i(t) \equiv \{\xi \ge 0 | C_i'(\xi)/y_i^* = \max_{i\in\mathcal{I}} C_i'(\bar{Q}_i(t))/y_i^*\}.$$

Then ${}_*\bar{Q}$ and ${}^*\bar{Q}$ are also Lipschitz, and hence, have derivatives almost everywhere (such times t are called strictly regular).

We note here that even with the above lemma, we cannot use an argument similar to that in Mandelbaum and Stolyar (2004) to establish the uniform attraction. In particular, one cannot establish the following inequality, which is similar to their (28):

$$\frac{d}{dt}[{}^*x(t)\ ] \le -\epsilon \text{ if } \bar{Q}(t) \text{ is not a fixed point,} \tag{37}$$

where ${}^*x(t) := \sum_{i=1}^3 \nu_i^* \cdot {}^*\bar{Q}_i(t)$ and $\epsilon > 0$. This is because ${}^*\bar{Q}$ may not be a positive combination of $\bar{L}$. Below is an example in which (37) does not hold.

Example 1. *Consider a system with $\lambda = (1,1,1)^T, \mu_{11} = 8/7, \mu_{21} = 4, \mu_{22} = 1$ and $\mu_{32} = 4$. Assume $P_{23} = 1$ and $P_{1i} = P_{3i} = 0, i \in \mathcal{I}$. Then $\lambda^e = (1,1,2)^T, x_{11}^* = 7/8, x_{21}^* = 1/8, x_{32}^* = x_{22}^* = 1/2$.*

Assume that at time t, $\mathcal{I}^*(t) = \{3\}$ and $\mathcal{I}_*(t) = \{1\}$. Then $\bar{T}_{32}'(t) = 1, \bar{T}_{21}'(t) = 1, {}^*\bar{Q}_3(t) = \bar{Q}_3(t)$ and ${}^*\bar{Q}_3'(t) = \bar{Q}_3'(t)$. Note that

$$\bar{Q}_3'(t) = \lambda_3 + \mu_{21}P_{23} - (1 - P_{33})\mu_{32} = \lambda_3 + \mu_{21} - \mu_{32} = 1.$$

Thus, ${}^*\bar{Q}_3'(t) = \bar{Q}_3'(t) > 0$. Because all ${}^*\bar{Q}_i'(t)$ have the same sign, then ${}^*x'(t) := \sum_{i=1}^3 \nu_i^* \cdot {}^*\bar{Q}_i'(t) > 0$. This means (37) may be not always true. □

Given the above example, our analysis will not be based on ${}^*\bar{Q}$. Instead, the analysis will be based on a linear combination of ${}_*\bar{Q}$.

Proposition 1. *For any hydrodynamic limit model under the proposed policy and any fixed time $t_0 \ge 0$,*

(a) *If $\bar{Q}(t_0) \neq 0$, there exists a fixed constant $T_1 \ge 0$ (depending on $\bar{Q}(t_0)$), such that $\bar{Q}$ reaches the fixed point $q^*((y^*)^T\bar{Q}(t_0))$ within the finite time T_1 and then stays there.*

(b) *If $\bar{Q}(t_0) = 0$, then for all $t \ge t_0$,*

$$\bar{Q}(t) = 0.$$

The detailed proof of Proposition 1 depends heavily on the "W" structure and is postponed to the "Supplemental Online Materials". Here we provide a sketch of the proof for Part (a), which consists of four steps:

Step 1: There exists a finite $t_1 \ge t_0$, such that for all $t \ge t_1, \mathcal{I}^*(t) \neq \{2\}$.
Step 2: For $t \ge t_1$, such that $\bar{Q}(t)$ is not a fixed point, $\sum_{i=1}^3 {}_*\bar{Q}_i'(t) \ge \epsilon_1$ for some $\epsilon_1 > 0$.
Step 3: For all $t \ge t_0, \bar{W}(t) = \bar{W}(t_0)$.
Step 4: Assume that $\bar{Q}(t_2)$ is a fixed point, then for all $t \ge t_2, \bar{Q}(t) = \bar{Q}(t_2) = q^*(\bar{W}(t_0))$.

The existence of the constant T_1 then follows easily from Steps 2 and 3.

5.4. Proof of theorem 1

With Proposition 1, we can establish the following lemma, which plays a key role in the proof of Theorem 1. The lemma is similar to Lemma 6 in Chen and Ye (2012), and its proof is postponed to the "Supplemental Online Materials".

Lemma 5.4. *Consider the time interval $[\tau, \tau + \delta]$, with $\tau \ge 0$ and $\delta > 0$; choose a constant $c > 0$, such that*

$$\sup_{\tau\le t_1 < t_2 \le \tau+\delta} |(y^*)^T\widehat{X}^*(t_1) - (y^*)^T\widehat{X}^*(t_2)| \le c.$$

Suppose that

$$\lim_{r\to\infty} \widehat{W}^r(\tau) = \chi \text{ and } \lim_{r\to\infty} \widehat{Q}^r(\tau) = q^*(\chi) \tag{38}$$

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

(a) *(SSC)*

$$|\bar{Q}^{r,\ell}(u) - q^*(\bar{W}^{r,\ell}(u))| \le \epsilon$$

for all $u \in [0, T]$;

(b) *(Boundedness)*

$$\bar{W}^{r,\ell}(u) \le \chi + c + 1$$

for all $u \in [0, T]$.

(c) *(Complementarity) If $\bar{W}^{r,\ell}(u) > \epsilon$ for all $u \in [0, T]$, then*

$$\bar{Y}^{r,\ell}(u) - \bar{Y}^{r,\ell}(0) = 0,$$

for all $u \in [0, T]$.

Now we are ready to prove Theorem 1 .

We first prove (a). First, from the properties of the reflection mapping (refer to Propositions 1 and 2 in Reiman (1984); or the least element characterization on page 163 of

Chen and Yao (2001)), we have

$$\widehat{W}^r(t) \geq \varphi((y^*)^T \widehat{X}^r)(t), \quad \text{and} \quad \widehat{Y}^r(t) \geq \psi((y^*)^T \widehat{X}^r)(t). \tag{39}$$

Let $\tau = 0$ and δ be an arbitrary positive number in Lemma 5.4. Then, we have the following

$$\widehat{Y}^r(\cdot) \text{does not increase at } t \text{ if } \widehat{W}^r(t) \geq \epsilon, t \in [0, \delta].$$

Thus, from Theorem 2.2 in Chen and Shanthikumar (1994),

$$\begin{aligned} \widehat{W}^r(t) \leq & \quad \varphi((y^*)^T \widehat{X}^r - \epsilon)(t), \\ \widehat{Y}^r(t) \leq & \quad \psi((y^*)^T \widehat{X}^r - \epsilon)(t). \end{aligned} \tag{40}$$

Note that the limits of fluid-scaled processes $(\bar{\bar{W}}^r, \bar{\bar{Q}}^r, \bar{\bar{Y}}^r, \bar{\bar{T}}^r)$, denoted by $(\bar{\bar{W}}, \bar{\bar{Q}}, \bar{\bar{Y}}, \bar{\bar{T}})$, will share equations (31)–(36) with the hydrodynamic limit model. That is, equations (31)–(36) still hold with $(\bar{W}, \bar{Q}, \bar{Y}, \bar{T})$ replaced by $(\bar{\bar{W}}, \bar{\bar{Q}}, \bar{\bar{Y}}, \bar{\bar{T}})$ accordingly. Under Assumption 3, $\bar{\bar{Q}}(0) = 0$, which means $\bar{\bar{Q}}(0) \in \mathcal{M}^*$. Thus, $\bar{\bar{Q}}(t) = \bar{\bar{Q}}(0) = 0$ by Proposition 1, which implies that $\bar{\bar{W}}(t) = \bar{\bar{W}}(0) = 0$. From equations (31)–(36), those facts imply the following linear system holds for $\bar{\bar{T}}_{ij}(t), i \in \mathcal{I}, j \in \mathcal{S}(i)$:

$$\bar{\bar{T}}_{11}(t) + \bar{\bar{T}}_{21}(t) = t \tag{41}$$

$$\bar{\bar{T}}_{22}(t) + \bar{\bar{T}}_{32}(t) = t \tag{42}$$

$$\mu_{11} \bar{\bar{T}}_{11}(t) = \lambda_1^e t \tag{43}$$

$$\mu_{21} \bar{\bar{T}}_{21}(t) + \mu_{22} \bar{\bar{T}}_{22}(t) = \lambda_2^e t \tag{44}$$

$$\mu_{32} \bar{\bar{T}}_{32}(t) = \lambda_3^e t. \tag{45}$$

Then, one can verify that the unique solution of the aforementioned linear system is $\bar{\bar{T}}_{ij}(t) = x_{ij}^* t$, and further $\widehat{X}^r \to \widehat{X}^*$ u.o.c. Combining (39) and (40), letting $r \to \infty$ and then $\epsilon \to 0$,

$$(\widehat{W}^r, \widehat{Y}^r) \to (\varphi(\widehat{X}^*), \psi(\widehat{X}^*)).$$

The convergence of $\widehat{Q}^r$ follows from Lemma 5.4(a) and the continuous mapping theorem.

Next, we prove (b). This part is similar to the proof of Proposition 2 in Ata and Kumar (2005), where a linear cost function is handled. For concreteness, we sketch the proof. It suffices to prove the following for any fixed sample path: for any subsequence $\mathcal{R}$ of r, there exists a further subsequence $\mathcal{R}' \subset \mathcal{R}$, such that along $\mathcal{R}' = \{r'\}$ and for $t \geq 0$,

$$\liminf_{r' \to \infty} \sum_{i=1}^{3} C_i(\widehat{Q}_i^{r'}(t)) \geq h(\varphi(\widehat{X}^*)(t)). \tag{46}$$

We now let $\mathcal{R}$ be fixed. Due to the functional central limit theorem for the renewal process $E^{r'}(t)$ and the Lipschitz continuity of $\bar{\bar{T}}^{r'}(t)$, there exists a subsequence $\mathcal{R}'$, such that, for $t \geq 0$ and along $\mathcal{R}'$

$$\bar{\bar{E}}^{r'}(t) \to \lambda t, \quad \bar{\bar{T}}^{r'}(t) \to \tilde{\bar{\bar{T}}}(t), \quad \text{u.o.c., as } r' \to +\infty.$$

Here, we put a “~” above the symbol to distinguish from the other fluid limit $\bar{\bar{T}}$. Consequently, the fluid-scaled processes $\bar{\bar{Q}}^{r'}(t), \bar{\bar{W}}^{r'}(t)$ and $\bar{\bar{Y}}^{r'}(t)$ also converge along $\mathcal{R}'$ to Lipschitz continuous processes $\tilde{\bar{\bar{Q}}}(t), \tilde{\bar{\bar{W}}}(t)$ and $\tilde{\bar{\bar{Y}}}(t)$, which satisfy the following equations (as in (31)–(36)):

$$\begin{aligned}
&\tilde{\bar{\bar{Q}}}_i(t) = \tilde{\bar{\bar{Q}}}_i(0) + \lambda_i t - \sum_{j \in \mathcal{S}(i)} \mu_{ij} \tilde{\bar{\bar{T}}}_{ij}(t) + \sum_{l=1}^{3} P_{li} \sum_{j \in \mathcal{S}(l)} \mu_{lj} \tilde{\bar{\bar{T}}}_{lj}(t), \\
&\sum_{i \in \mathcal{C}(j)} \left[\tilde{\bar{\bar{T}}}_{ij}(t) - \tilde{\bar{\bar{T}}}_{ij}(s) \right] \leq t - s, \text{ for } s < t, j \in \mathcal{J}, \\
&\tilde{\bar{\bar{W}}}(t) = (y^*)^T \tilde{\bar{\bar{Q}}}(t) = \tilde{\bar{\bar{W}}}(0) + \tilde{\bar{\bar{Y}}}(t), \\
&\tilde{\bar{\bar{Y}}}(t) = \sum_{j \in \mathcal{J}} z_j^* \left(t - \sum_{i \in \mathcal{C}(j)} \tilde{\bar{\bar{T}}}_{ij}(t) \right), \\
&\int_0^\infty \mathbf{1}_{\left\{ \sum_{i \in \mathcal{C}(j)} \tilde{\bar{\bar{Q}}}_i(s) > 0 \right\}} d\left(s - \sum_{i \in \mathcal{C}(j)} \tilde{\bar{\bar{T}}}_{ij}(s) \right) = 0, \ j \in \mathcal{J}, \\
&\int_0^\infty \left(\max_{l \in \mathcal{C}(j)} \frac{C_l'(\tilde{\bar{\bar{Q}}}_l(t))}{y_l^*} - \frac{C_i'(\tilde{\bar{\bar{Q}}}_i(t))}{y_i^*} \right)^+ d\tilde{\bar{\bar{T}}}_{ij}(s) = 0, \\
&j \in \mathcal{J} \text{ and } i \in \mathcal{C}(j).
\end{aligned}$$

We consider two cases:

1. $\tilde{\bar{\bar{Q}}}(t) \neq 0$: Then $(y^*)^T \tilde{\bar{\bar{Q}}}(t) > 0$. Since $\widehat{W}^{r'}(t) = (y^*)^T \widehat{Q}^{r'}(t) = r'(y^*)^T \bar{\bar{Q}}^{r'}(t)$ and $r'(y^*)^T \bar{\bar{Q}}^{r'}(t) \to \infty$ as r' tends to ∞, we have that $\lim_{r' \to \infty} \widehat{W}^{r'}(t) = \infty$. As a result,

$$\liminf_{r' \to \infty} \sum_{i=1}^{3} C_i(\widehat{Q}_i^{r'}(t)) \geq \liminf_{r' \to \infty} h(\widehat{W}^{r'}(t)) = \infty,$$

which implies (46).

2. $\tilde{\bar{\bar{Q}}}(t) = 0$: Then $\tilde{\bar{\bar{W}}}(t) = (y^*)^T \tilde{\bar{\bar{Q}}}(t) = 0$. Since $\tilde{\bar{\bar{W}}}(t)$ is non-decreasing under heavy traffic, $\tilde{\bar{\bar{W}}}(s) = 0$ for $s \in [0, t]$ and then $\tilde{\bar{\bar{Q}}}(s) = 0$ for $s \in [0, t]$. Under policy π_*, this fact implies that equations (41)–(45) still hold with $\bar{\bar{T}}_{ij}(t), i \in \mathcal{I}, j \in \mathcal{S}(i)$ replaced by $\tilde{\bar{\bar{T}}}_{ij}(t), i \in \mathcal{I}, j \in \mathcal{S}(i)$. Then $\tilde{\bar{\bar{T}}}_{ij}(t) = x_{ij}^* t$, and further $\widehat{X}^{r'} \to \widehat{X}^*$ u.o.c. From the definition of $h(\cdot)$ in (23) and the least element property of reflection mapping, we have

$$\sum_{i=1}^{3} C_i(\widehat{Q}_i^{r'}(t)) \geq h(\widehat{W}^{r'}(t)) \geq h(\varphi((y^*)^T \widehat{X}^{r'})(t)).$$

This inequality implies (46) by taking the limit on both sides.

6. Numerical experiments

To supplement our theoretical result of asymptotic optimality, we next provide some engineering confirmations by conducting computer simulation experiments. In Section 6.1 we first consider a simple W model operated under the local policy. To understand the potential of the local policy for the more general non-W models, in Section 6.2 we focus on a non-W example.

Table 1. Comparisons of simulated congestion costs under four policies: (i) local policy, (ii) MS policy, (iii) static priority policy, and (iv) global FCFS.

Policies	*Local Policy*	*MS Policy*	*Priority*	*FCFS*
Costs	**2.178**	2.290	3.957	5.113
($\times 10^5$)	($\pm$0.175)	($\pm$0.188)	($\pm$0.343)	($\pm$0.398)

Table 2. Comparisons of simulated congestion costs under four policies for the non-W model depicted in Figure 2: (i) local policy, (ii) MS policy, (iii) static priority policy, and (iv) global FCFS.

Policies	*Local Policy*	*MS Policy*	*Priority*	*FCFS*
Costs	**0.717**	0.740	1.542	2.546
($\times 10^5$)	($\pm$0.055)	($\pm$0.062)	($\pm$0.125)	($\pm$0.191)

6.1. Simulation results for a W model

We consider a W model with arrival rates $\lambda_i = 1, i \in \mathcal{I}$, service rates $\mu_{11} = 5, \mu_{32} = 7, \mu_{21} = 2$ and $\mu_{22} = 1$, and transition probabilities $P_{11} = 0.6, P_{12} = 0.2, P_{23} = P_{33} = 0.5$ and $P_{ij} = 0$. In this example we consider quadratic congestion costs $C_1(x) = 3x^2, C_2(x) = x^2, C_3(x) = 2x^2$. In this setting, we obtain that $\lambda^e = (5/2, 3/2, 7/2)^T$ and $x^*_{ij} = 0.5$ for $i \in \mathcal{I}$ and $j \in \mathcal{S}(i)$.

To evaluate the performance of our local policy, we provide simulation results for the W model under several policies. In particular, we will benchmark the simulated congestion costs under the local policy with those under the following three policies:

1. **MS policy**. The scheduling rule proposed by Mandelbaum and Stolyar (2004) (see (7) for details) that requires the “global” queue lengths information.
2. **Static priority policy.** The policy in that server 1 and server 2 always prioritize on serving classes 1 and 3 customers over class 2 customers (because classes 1 and 3 customers are more costly).
3. **Global FCFS.** Both servers serve customers in the order of their arrivals across all customers that they are eligible to serve.

Under each policy, we generate $N = 800$ independent sample paths, each of which has a length of $T = 24 \times 60 \times 100 = 144,000$ time units. Using the simulated data, we construct 95% confidence intervals using sample means and sample variances obtained by averaging results in N independent runs. The average cumulative costs under the corresponding policies are summarized in Table 1.

According to Table 1, it is evident that our local policy generates a much smaller costs than the FCFS and priority policies. This is quite intuitive because neither FCFS nor the priority policy is directly designed to minimize the congestion cost. The most promising observation in Table 1 is that our local policy even outperforms the MS policy, despite the point that MS utilizes the global queue length information whereas our policy does not. We attempt to explain this observation below. First, although the MS policy conjectured in Mandelbaum and Stolyar (2004) seems intuitive and appropriate, it may not be optimal or asymptotically optimal at all. (Theoretical proof of the optimality of the MS policy remains an open problem after all.) Second, the MS policy is not work-conserving: when the MS policy is being implemented, a server should always remain idling as long as the maximum in (7) is nonpositive, even though there are positive queues that the server is eligible to serve.

6.2. Local policy for more general SPNs

The present simple W model has shed lights on the possibility and potential effectiveness of control policies utilizing only local information. But eventually we hope to be able to treat more general and practical SPNs. At this point there remains two open questions for a general (non-W) SPN:

1. How to properly define a general local policy?
2. How to provide mathematical justification for the effectiveness of the general local policy?

To answer the first question, we will need to properly define the weight parameter y^* for a general SPN, and even this does not seem to be straightforward. The identification of y^* should be closely related to the equivalent workload formulation in general multi-class multi-server queues; closed-or expressions for y^* may not be available, but we envision that the framework in Harrison and van Mieghem (1997) and Harrison (2000) may help to derive a computational scheme to compute y^*.

The proof of asymptotic optimality for the general local policy is another challenge. We point out that the W structure plays an important role in our current proof; and new methodologies are needed to prove the uniform attraction for the more general (non-W). We leave this as future research. To give the readers a taste of how a general local policy may look like and how it performs comparing to other well-known policies, we next consider a simple non-W model.

Example 2 (A simple non-W example). *There are four customer classes and two servers (see Figure 2). Server 1 serves classes 1 and 2, whereas Server 2 serves classes 2, 3, and 4. Assume that $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (1, 0, 1, 0)$ and $(\mu_{11}, \mu_{21}, \mu_{22}, \mu_{32}, \mu_{42}) = (6, 1, 1/2, 8, 4)$. The feedback transition probabilities are $P_{14} = 1, P_{1i} = 0$ $(i = 1, 2, 3)$; $P_{21} = 1/2, P_{2i} = 0$ $(i = 2, 3, 4)$; $P_{32} = 1, P_{3i} = 0$ $(i = 1, 3, 4)$; and $P_{4i} = 0$ $(i = 1, 2, 3, 4)$. Then the arrival of class 2 can only be from the departure of class 3. Then $\lambda^e = (3/2, 1, 1, 3/2)^T$ and $x^*_{11} = 1/4, x^*_{21} = 3/4, x^*_{22} = 1/2, x^*_{32} = 1/8$ and $x^*_{42} = 3/8$.*

The local policy has the same expression as in (11)*, with cost functions $C_1(x) = 3x^2, C_2(x) = C_3(x) = C_4(x) = x^2$ and weight $y^* = (I - P)^{-1}\nu^*$, where*

$$\nu^* = \left(1, \frac{\mu_{11}}{\mu_{21}}, \frac{\mu_{11}}{\mu_{21}} \times \frac{\mu_{22}}{\mu_{32}}, \frac{\mu_{11}}{\mu_{21}} \times \frac{\mu_{22}}{\mu_{42}}\right)^T = \left(1, 6, \frac{3}{8}, \frac{3}{4}\right)^T.$$

The definition of ν^* follows the similar idea as in (9).

Next we conduct simulation experiments. Following the settings in Section 6.1, we report the average cumulative costs under the corresponding policies in Table 2. We remark that under the static priority policy, class 2 has the lowest priority, class 3 has lower priority than class 4.

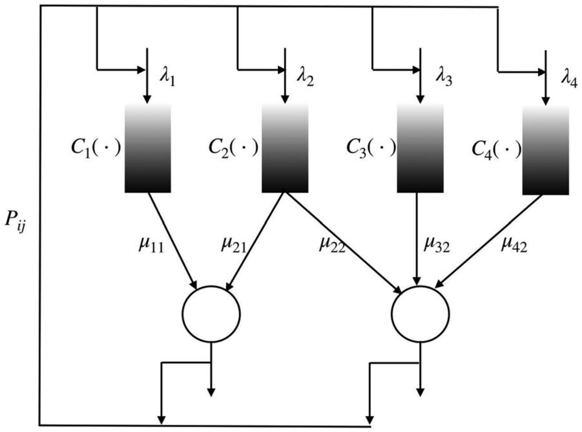

Figure 2. A queueing system with four customer classes and two servers.

Similar to results in Section 6.2, our local policy outperforms all other policies, including the MS policy that utilizes local information. Although Table 2 reveals the remarkable performance of our local policy, the proof of the asymptotical optimality remains an open question. The methods in Mandelbaum and Stolyar (2004) and ours show that at least one of $^*\bar{Q}$ *and* $_*\bar{Q}$ *is moving towards the fixed point in the proof of the uniform attraction. However, it is not true for this example. To see that, assume that at time t,* $\mathcal{I}^*(t) = \{4\}$ *and* $\mathcal{I}_*(t) = \{2\}$*. Then it is easy to verify that* $\bar{Q}_2'(t) = 0$*. For* $\bar{Q}_4(t)$*:*

$$^*\bar{Q}_4'(t) = \bar{Q}_4'(t) = \quad \mu_{11} - \mu_{42} = 2 > 0.$$

Hence, $\sum {}_*\bar{Q}_i'(t) = \bar{Q}_2'(t) = 0$ *but* $\sum {}^*\bar{Q}_i'(t) = \bar{Q}_4'(t) > 0$*, which means that Lemma 5.3 does not directly apply to this case, so that we cannot follow the four-step approach in the present article to prove the uniform attraction. It will be necessary to construct a new Lyapunov function L, which is left for the research.*

7. Conclusion

In this article, we study a routing and scheduling problem in a multi-class multi-server W queueing model with customer feedback. We propose a new control policy called the *local policy*. Different from previously studied policies in SPNs with customer feedback (e.g., Mandelbaum and Stolyar (2004)) which requires knowledge of the queue length information of the entire system, our local policy demands only local information (i.e., accesses to only queue lengths information of those customer classes that a server is eligible to serve). Despite of the simplicity of the present W model, our analysis and results serve as an initial attempt to gain insights into the scheduling of SPNs where information retrieval is costly, slow or inconvenient.

We substantiate the performance of our local policy by showing that it is asymptotically optimal as the traffic intensity approaches one, with the objective of minimizing the system's congestion cost (measured by a convex cost function of the queue lengths). To prove SSC (a key step to establish the asymptotic optimality), we follow the general framework in Bramson (1998), where one needs to verify the uniform attraction for the hydrodynamic limit. However, existing methods in the literature for uniform attraction do not apply to SPNs having customer feedback. In our proofs, we construct a four-step procedure by heavily exploring the W structure to prove the uniform attraction, and then the global SSC, which in turn guarantees that all queue length processes reduce asymptotically to a one-dimensional workload process. Supplementing our theoretical results, we also report convincing computer simulations experiments; these results confirm that our local policy performs well by benchmarking with other well-known policies.

One natural extension is to devise appropriate local policies for multi-class multi-server queues having a more general topological structure (e.g., a W model with n classes and $n-1$ servers with $n \geq 3$, and the more general non-W model configured by a general skill-based customer–server matching matrix). The first step is to carefully define the notion of local policies, which itself is not a trivial task; the second step is to develop new methodologies for the proof of the asymptotic optimality. Although numerical examples have revealed promising results, the required theoretical analysis in the aforementioned two directions remain open. We leave this extension to future research.

Acknowledgments

The authors would like to thank the associate editor and two anonymous referees for many constructive comments and suggestions. This article has been substantially improved due to their efforts.

Funding

This work was supported by the National Natural Science Foundation of China under Grants 11971074 and 11871116, and in part by Hong Kong Research Grants Council (Project 14504217).

Notes on contributors

Dr. Jiankui Yang is an associate professor in the School of Science, Beijing University of Posts and Telecommunications. He obtained his BE and PhD degrees from the Mathematics Department at Nanjing University. His current research interests include the stochastic models and quantitative analysis.

Dr. Junfei Huang is an associate professor in the Department of Decision Sciences and Managerial Economics at the Chinese University of Hong Kong. His research interests are in asymptotic analysis and optimal control of queueing systems and their applications in manufacturing and services.

Dr. Yunan Liu is an associate professor in Department of Industrial and Systems Engineering at North Carolina State University. He obtained his BE degree from the Electrical Engineering Department at Tsinghua University, MS and PhD degrees from the Industrial Engineering and Operations Research Department at Columbia University. His research interests include queueing theory, applied probability, simulations, optimal control, online learning, and their applications to call-center, health-care, transportation, and blockchain. His work was awarded first place in the INFORMS Junior Faculty Interest Group Paper Competition in 2016.

ORCID

Junfei Huang http://orcid.org/0000-0002-3764-354X
Yunan Liu http://orcid.org/0000-0001-9961-2610

References

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

Bell, S. and Williams, R. (2001) Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy. *Annals of Applied Probability*, **11**(3), 608–649.

Bramson, M. (1998) State-space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems*, **30**(1-2), 89–140.

Chen, H. and Shanthikumar, J.G. (1994) Fluid limits and diffusion approximations for networks of multi-server queues in heavy traffic. *Discrete Event Dynamic Systems*, **4**(3), 269–291.

Chen, H. and Yao, D.D. (2001) *Fundamentals of Queueing Networks: Performance*, *Asymptotics, and Optimization*, Springer-Verlag, New York, NY.

Chen, H. and Ye, H.Q. (2012) Asymptotic optimality of balanced routing. *Operations Research*, **60**(1), 163–179.

Dai, J.G. and Lin, W. (2008) Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Annals of Applied Probability*, **18**(6), 2239–2299.

Harrison, J.M. (1988) Brownian models of queueing networks with heterogeneous customer populations in *Stochastic Differential Systems, Stochastic Control Theory and Their Applications*, Springer, New York, NY, pp. 147–186.

Harrison, J.M. (1998) Heavy traffic analysis of a system with parallel servers: Asymptotic optimality of discrete review policies. *Annals of Applied Probability*, **8**(3), 822–848.

Harrison, J.M. (2000) Brownian models of open processing networks: Canonical representation of workload. *Annals of Applied Probability*, **10**(1), 75–103.

Harrison, J.M. and Lopez, M.J. (1999) Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, **33**(4), 339–368.

Harrison, J.M. and van Mieghem, J.A. (1997) Dynamic control of Brownian networks: State space collapse and equivalent workload formulations. *Annals of Applied Probability*, **7**(3), 747–771.

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

Liu, Y. and Whitt, W. (2017) Stabilizing performance in a service system with time-varying arrivals and customer feedback. *European Journal of Operational Research*, **256**(2), 473–486.

Lu, Y., Xie, Q., Kliot, G., Geller, A., Larus, J. and Greenberg, A.(2011) Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services. *Performance Evaluation*, **68**(11), 473–486.

Mandelbaum, A. and Stolyar, A.L. (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Research*, **52**(6), 836–855.

Reiman, M.I. (1984) Open queueing networks in heavy traffic. *Mathematics of Operations Research*, **9**, 441–458.

Stolyar, A. (2004) MaxWeight scheduling in a generalized switch: State space collapse and equivalent workload minimization under complete resource pooling. *Annals of Applied Probability*, **14**, 1–53.

Tran-Gia, P. and Mandjes, M. (1997) Modeling of customer retrial phenomenon in cellular mobile networks. *IEEE Journal on Selected Areas in Communications*, **15**(8), 1406–1414.

Van Mieghem, J.A. (1995) Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Annals of Applied Probability*, **5**(3), 809–833.

Vvedenskaya, N.D., Dobrushin, R.L. and Karpelevich, F.I. (1996) Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problems of Information Transmission*, **32**(1), 15–27.

Wang, J., Wang, Z. and Liu, Y. (2020) Reducing delays in retrial queues by simultaneously differentiating service rates and retrial rates. *Operations Research*, **68**(6), 1648–1667.

Ye, H.Q. and Yao, D.D. (2008) Heavy traffic optimality of stochastic networks under utility-maximizing resource control. *Operations Research*, **56**(2), 453–470.

Yom-Tov, G. and Mandelbaum, A. (2014) Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, **16**(2), 283–299.