J. Appl. Prob. **44,** 321–331 (2007)
Printed in England

A PARADOX FOR ADMISSION CONTROL OF MULTICLASS QUEUEING NETWORK WITH DIFFERENTIATED SERVICE

HENG-QING YE,* *Hong Kong Polytechnic University and National University of Singapore*

Abstract

In this paper we present counter-intuitive examples for the multiclass queueing network, where each station may serve more than one job class with differentiated service priority and each job may require service sequentially by more than one service station. In our examples, the network performance is improved even when more jobs are admitted for service.

Keywords: Multiclass queueing network; admission control; stability analysis; performance analysis; fluid approximation

2000 Mathematics Subject Classification: Primary 60K25
Secondary 60J25

1. Introduction

The queueing network model is an important tool for studying the service system, the manufacturing system, and the communication system. In many applications, the model is useful in identifying bottleneck resources so that better decisions can be made on designing and controlling the network. Such decisions may include, for example, selecting the system service capacity (e.g. the maximum service rates of work stations), adjusting system workload (e.g. the job arrival rate and pattern), and routeing jobs to service stations if jobs can be served by more than one station.

In practice, it is commonly believed that the performance for a queueing network system, say in terms of the average total number or the average delay of jobs in the system, would be improved if the service capacity (system workload, routeing alternatives, respectively) is increased (decreased, increased, respectively). Such an understanding is sound when studying the queueing system with single or parallel service stations and the product-form queueing network (cf. Chen and Yao (2001, Chapters 1–4) and references therein). However, one must be cautious in applying such intuition to complex queueing systems. In fact, from the study of the stability condition of a three-station multiclass queueing network in Dumas (1997), it is evident that with increased service capacity, the network (Dumas network) for certain work stations performs worse. The paradox concerning the (distributed) routeing in the queueing network is also discussed in Cohen and Kelly (1990), which is based on the well-known Braess paradox (Braess (1968)). In addition to these paradoxes on the service capacity and routeing, we provide paradoxical network examples of the admission control. These counter-intuitive examples show that the network performance could be degraded even when the arrival rate of jobs decreases.

Received 8 July 2005; revision received 14 February 2007.
* Postal address: Department of Logistics, Hong Kong Polytechnic University, Hong Kong, P. R. China.
Email address: lgtyehq@inet.polyu.edu.hk
Supported in part by a grant from National University of Singapore.

We describe the multiclass queueing network model and present counter-intuitive results in Section 2. In Section 3, we introduce the fluid model approach developed in recent years and then use this approach to prove our main results. We present our conclusions in Section 4.

2. Counter-examples and main results

The *multiclass queueing network* consists of J stations indexed by $j \in \mathcal{J} = \{1, \ldots, J\}$, and K job classes indexed by $k \in \mathcal{K} = \{1, \ldots, K\}$. Assume that the arrival process of class k jobs (or customers) is a Poisson process with arrival rate α_k (≥ 0), and the service time for each class k job is exponentially distributed with mean service time m_k (> 0). Denote $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)^\top$ and $\boldsymbol{m} = (m_1, \ldots, m_K)^\top$. We also assume that all the interarrival times and service times are independent. A class k job is served at station $\sigma(k)$ ($\sigma(\cdot)\colon \mathcal{K} \to \mathcal{J}$), and after its service completion it may become a class ℓ job with probability $p_{k\ell}$ and leave the network with probability $1 - \sum_{\ell=1}^{K} p_{k\ell}$. Let $\boldsymbol{P} = (p_{k\ell})$. Let $\boldsymbol{C} = (c_{jk})$ be a $J \times K$ matrix whose (j, k)th component

$$c_{jk} = \begin{cases} 1 & \text{if } j = \sigma(k), \\ 0 & \text{otherwise.} \end{cases}$$

While each station may serve more than one class of jobs, each job is served at one specific station (determined by the many-to-one mapping $\sigma(\cdot)$). We study the preemptive priority service discipline specified by a one-to-one mapping $\boldsymbol{\pi}\colon \mathcal{K} \to \mathcal{K}$. For any given ℓ and k, if $\pi(\ell) < \pi(k)$ and $\sigma(\ell) = \sigma(k)$, then a class k job cannot be served at station $\sigma(k)$ unless there is no class ℓ job. In short, we say that a class ℓ job has a higher priority than a class k job. For convenience, the mapping $\boldsymbol{\pi}$ is often expressed as a permutation of $\mathcal{K}$, which can be written as $\boldsymbol{\pi} = (i_1, \ldots, i_K)$ if $\boldsymbol{\pi}(k) = i_k$ and $k \in \mathcal{K}$. In addition, we only consider work-conserving (or nonidling) service disciplines, which specifies that a work station cannot be idle unless there is no job waiting for service in that station. For convenience, we denote the queueing network described above as $(\mathcal{J}, \mathcal{K}, \boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{C}, \boldsymbol{P}, \boldsymbol{\pi})$.

We study *open* multiclass queueing networks, assuming that $\boldsymbol{P}$ is transient, i.e. $\mathbf{1} + \boldsymbol{P} + \boldsymbol{P}^2 + \cdots$ is convergent. Let

$$\boldsymbol{\lambda} = (\mathbf{1} - \boldsymbol{P}')^{-1}\boldsymbol{\alpha}, \qquad \boldsymbol{\beta} = \boldsymbol{M}\boldsymbol{\lambda}, \quad \text{and} \quad \boldsymbol{\rho} = \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{M}\boldsymbol{\lambda}.$$

(Here, $\boldsymbol{M} = \operatorname{diag}(m)$ is a K-dimensional diagonal matrix whose kth diagonal element is m_k.) Call $\boldsymbol{\lambda}$ a *nominal total arrival rate* (vector), β_k (the kth component of $\boldsymbol{\beta}$) a *traffic intensity for class* k, $k \in \mathcal{K}$, and ρ_j (the jth component of $\boldsymbol{\rho}$) a *traffic intensity for station* j, $j \in \mathcal{J}$. Usually, the vector $\boldsymbol{\rho} = (\rho_j)$ is simply called the *traffic intensity of the queueing network*. In fact, $\boldsymbol{\lambda}$ is the unique solution to the *traffic equation*, $\boldsymbol{\lambda} = \boldsymbol{\alpha} + \boldsymbol{P}'\boldsymbol{\lambda}$ (where $\boldsymbol{P}'$ denotes the transpose of $\boldsymbol{P}$), and includes both external arrivals and internal transitions.

The dynamics of the network can be described using a K-dimensional queue length process $Q(t) = (Q_k(t),\ k \in \mathcal{K})$ ($t \geq 0$), where $Q_k(t)$ indicates the number of class k jobs in the network at time t. Assume for simplicity that queues are initially empty, i.e. $Q_k(0) = 0$ for all $k \in \mathcal{K}$. $Q(t)$ is a continuous time Markov chain under the Poisson arrival and exponential service assumptions. It is known that the Markov chain $Q(t)$ is positive recurrent, or stable, *only if* the traffic intensity for each station is less than one, that is if $\boldsymbol{\rho} < \boldsymbol{e}$. (Here, $\boldsymbol{e}$ is a J-dimensional column vector with all components being 1s.) The performance index of interest in this paper is given by

$$Q^* := \lim_{t\to\infty} \mathrm{E}\left[\sum_{k\in\mathcal{K}} Q_k(t)\right],$$

https://doi.org/10.1239/jap/1183667404 Published online by Cambridge University Press

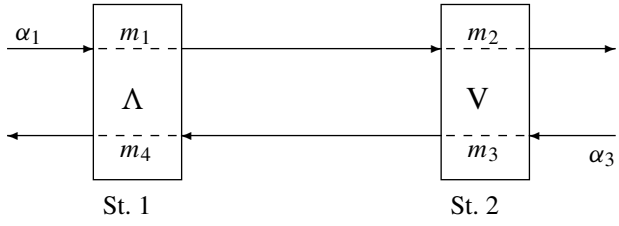

FIGURE 1: KRSS network.

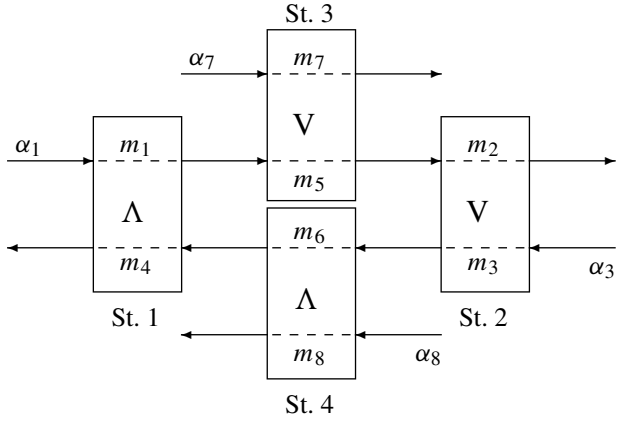

FIGURE 2: Modified KRSS network.

which is the expected stationary total queue length when it is finite. The expected queue length Q^* is finite if and only if the queue length process $Q(t)$ is positive recurrent.

As an example, the Kumar–Rybko–Seidman–Stolyar (KRSS) network is illustrated in Figure 1. This network, widely known as the Kumar–Seidman network or the Rybko–Stolyar network in queueing network literature, was first studied independently by Kumar and Seidman (1990) and Rybko and Stolyar (1992). The KRSS network consists of two stations and four job classes. Among the four job classes, only class 1 and 3 have external job arrivals, i.e. $\alpha_2 = \alpha_4 = 0$. A class 1 (class 3) job becomes a class 2 (class 4) job after its service completion at station 1 (station 2), while a class 2 (class 4) job leaves the system after its service completion at station 2 (station 1). A class 4 (class 2) job has a higher priority than a class 1 (class 3) job at station 1 (station 2). For this network, the parameters (matrices) $\boldsymbol{C}$ and $\boldsymbol{P}$ can be written down easily from Figure 1, and $\boldsymbol{\pi}$ is specified as $\boldsymbol{\pi} = (4, 1, 2, 3)$. With a little thought, it is straightforward to see that the traffic intensity is simply

$$\boldsymbol{\rho} = (\rho_1, \rho_2)^\top = (\alpha_1 m_1 + \alpha_3 m_4, \alpha_1 m_2 + \alpha_3 m_3)^\top.$$

It is well known (e.g. Chen and Zhang 2000) that the KRSS network is stable if and only if

$$\boldsymbol{\rho} < \boldsymbol{e} \quad \text{and} \quad \alpha_1 m_2 + \alpha_3 m_4 < 1.$$

The counterexample that presents a paradox in the admission control of open multiclass queueing networks is a variation of the KRSS network. This network is illustrated in Figure 2, and we will refer to it as the *modified* KRSS network throughout this paper. Compared with the original KRSS network, there are two additional stations, namely station 3 and 4, and four additional job classes, namely class 5, 6, 7, and 8. A class 7 (class 8) job has a higher priority

than a class 5 (class 6) at station 3 (station 4). The details of the specific network parameters $(\mathcal{J}, \mathcal{K}, \boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{C}, \boldsymbol{P}, \boldsymbol{\pi})$ for this network should be obvious from Figure 2. For the modified KRSS network, we have the following result.

Theorem 2.1. *Suppose in the modified KRSS network,* $\boldsymbol{\rho} < \boldsymbol{e}$ *and*

$$\alpha_1 m_2 + \alpha_3 m_4 > 1. \tag{2.1}$$

(1) *If* $m_5/(1 - \alpha_7 m_7) > m_2$ *and* $m_6/(1 - \alpha_8 m_8) > m_4$, *then the queue length process* $Q(t)$ *is positive recurrent, and* $Q^* < \infty$.

(2) *If* $m_5/(1 - \alpha_7 m_7) < m_1$ *and* $m_6/(1 - \alpha_8 m_8) < m_3$, *then the queue length process* $Q(t)$ *is transient, and* $Q^* = \infty$.

This theorem presents a phenomenon in which reducing the arrival rates of some job classes leads to worse performance of the queueing network. To see this, fix all the parameters of the modified KRSS network except α_7 and α_8. In statement (1) of Theorem 2.1, we have that

$$\alpha_7 > \frac{(1 - m_5/m_2)}{m_7} \quad \text{and} \quad \alpha_8 > \frac{(1 - m_6/m_4)}{m_8}, \tag{2.2}$$

and that the expected stationary total queue length Q^* is finite. However, when we reduce α_7 and α_8 to

$$\alpha_7 < \frac{(1 - m_5/m_1)}{m_7} \quad \text{and} \quad \alpha_8 < \frac{(1 - m_6/m_3)}{m_8}, \tag{2.3}$$

the queue length process $Q(t)$ becomes transient and thus Q^* becomes infinite. We will see in Section 3 that $\sum_{k \in \mathcal{K}} Q_k(t) \to \infty$ almost surely.

To gain better intuition of the paradoxical phenomenon, we examine the dynamics of the original KRSS network with no initial job (note that the initial condition has no impact on the long term network behavior). When a class 4 job is being served, class 1 jobs cannot move to class 2 for further service, and vice versa. From this observation, it is not difficult to infer that classes 2 and 4 will never be served at the same time and in effect form a virtual station (Dai and Vande Vate (1996)). Therefore, the total nominal traffic intensity for these two classes together, i.e. the virtual station, should not exceed one for the network to be stable. A similar argument establishes that the KRSS network is unstable when the nominal traffic intensity for the virtual station exceeds one, i.e. (2.1) holds. Now consider the modified KRSS network. The additional classes 5 and 6 act as *regulators* that regulate the traffic to classes 2 and 4 respectively so as to stabilize the network. (Readers are referred to Humes (1994) for futher discussion on the application of regulators to stabilize queueing networks.) When the workloads for classes 7 and 8 are light such that (2.3) holds, much of the service capacity for stations 3 and 4 is left to classes 5 and 6 respectively and hence, classes 5 and 6 do not hold back the traffic to avoid building up job queues at classes 2 and 4 respectively. Thus, the virtual station effect prevails and the network is still unstable under (2.1) (cf. Theorem 2.1(2)). However, when the workloads for classes 7 and 8 are heavy enough such that (2.2) holds, the service for lower priority classes 5 and 6 is in effect slowed down and the traffic to classes 2 and 4 is held back. Consequently, there would not be large buildup of queues at classes 2 and 4, and these two classes would not mutually block their services. Finally, the virtual station effect is avoided and the modified KRSS network is thus stabilized (cf. Theorem 2.1(1)). The above argument will be made rigorous in the proof of Theorem 2.1 in Section 3.

https://doi.org/10.1239/jap/1183667404 Published online by Cambridge University Press

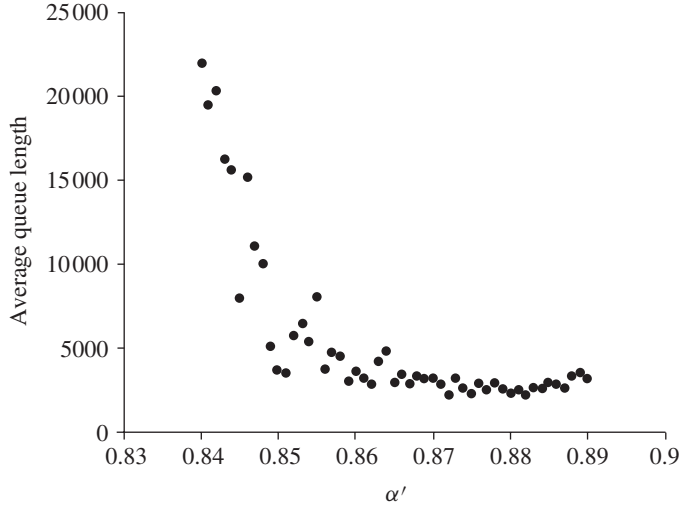

FIGURE 3: Simulation results for the modified KRSS network.

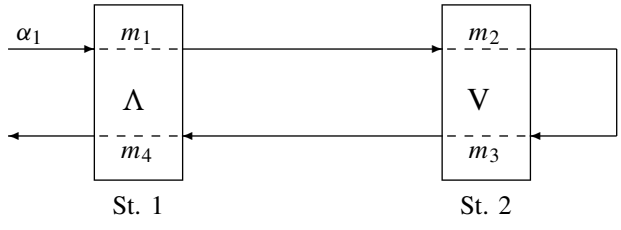

FIGURE 4: LK network.

Concerning the above paradoxical phenomenon, a subtle question to ask is whether this counter-intuitive phenomenon is due just to pathological jumps in the network performance. To address this question in more detail, we take, for the moment, that $\alpha_1 = \alpha_3 = 1, \alpha_7 = \alpha_8 = \alpha'$, $m_1 = m_3 = 0.2$, $m_2 = m_4 = 0.6$, $m_5 = m_6 = 0.1$, and $m_7 = m_8 = 1$. Then, let α' vary, say, from $\frac{8}{9}$ down to $\frac{1}{3}$, and thus $m_5/(1 - \alpha_7 m_7)$ and $m_6/(1 - \alpha_8 m_8)$ both vary from 0.9 (which is greater than m_3 and m_6) to 0.15 (which is less than m_3 and m_6). Based on Theorem 2.1, the expected stationary total queue length Q^* is finite when α' is $\frac{8}{9}$, but it becomes worse, i.e. $Q^* = \infty$, when α' is reduced to $\frac{1}{3}$. Now, the subtle questions are as follows. Is this performance degradation upon reducing arrival rate α' simply due to a jump from a stable to an unstable network at a critical point of α' when it varies from $\frac{8}{9}$ to $\frac{1}{3}$? Is the performance Q^* still an increasing function of the arrival rate α' within any interval of α' where the network is stable and its expected total queue length Q^* is finite? It is not obvious how to eliminate this possible pathological situation theoretically. However, our simulation results illustrated in Figure 2 indicate that the average total queue length Q^* is a decreasing function of α' within some intervals of α' (i.e. the interval [0.84, 0.89] in our simulation) where $Q(t)$ is stable. The network performance is improved *continuously* when more jobs are admitted to the system within a certain range of job arrival rates.

https://doi.org/10.1239/jap/1183667404 Published online by Cambridge University Press

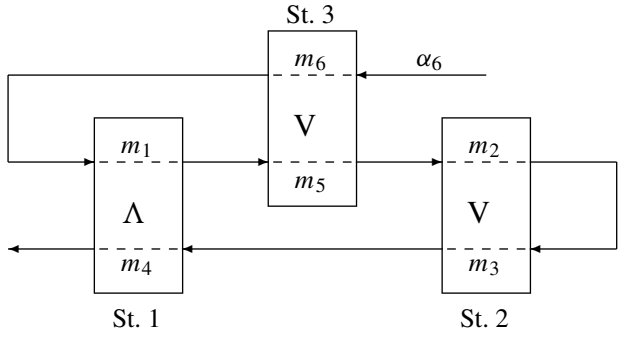

FIGURE 5: A modified LK network.

Another counterexample that provides a different perspective on the paradox in admission control is related to the Lu–Kumar (LK) network, which was first studied by Lu and Kumar (1991) and is illustrated in Figure 4. We omit a detailed description of this network, which should be clear from its comparison with the KRSS network. This counterexample is a variation of the LK network, called the *modified* LK network in this paper, and is illustrated in Figure 5. For the modified LK network, we study some special instances (for convenience) and summarize the counter-intuitive phenomenon in the following theorem.

Theorem 2.2. *Consider the modified LK network with* $\boldsymbol{m} = (0.1, 0.6, 0.1, 0.6, 0.7, 0.027)^\top$.

(1) *If* $\alpha_6 = 1.37$, *then the queue length process* $Q(t)$ *is positive recurrent, and* $Q^* < \infty$.

(2) *If* $\alpha_6 = 1$, *then the queue length process* $Q(t)$ *is transient, and* $Q^* = \infty$.

This theorem presents a situation in which, when the arrival rate α_6 drops from 1.37 to 1, the performance becomes worse. Similar to the simulation for the modified KRSS network, our simulation result also supports the conclusion that the average total queue length Q^* for the modified LK network would be a decreasing function of α_6 within some intervals of α_6 where $Q(t)$ is stable. In contrast to the modified KRSS network, a special feature of the modified LK network is that there is only one external arrival and this arrival is controllable. On the other hand, if we fix the rate α_6 of the unique external arrival and vary the service times m_k, $k = 1, \ldots, 6$, in proportion, then we recover an example for the paradox on service control. That is, increasing the service capacity may also worsen the system performance, since reducing the service times m_k in proportion (i.e. increasing the service capacity) is equivalent to reducing the external arrival α_6 in the modified LK network by changing the time scale suitably.

3. Multiclass fluid network model and proof of Theorem 2.1

In this section, we provide a proof of Theorem 2.1, but we omit a proof of Theorem 2.2 as it follows similarly to the proof of Theorem 2.1. We employ the fluid model approach in the proof. The development of this approach was inspired by the studies of some counterexamples in Kumar and Seidman (1990), Rybko and Stolyar (1992), Bramson (1994), etc., where the multiclass queueing networks are not stable even when the traffic intensity of each station in the network is less than one. An elegant result of the fluid model approach which states that a queueing network is stable if its corresponding fluid network model is stable, was first proposed in Rybko and Stolyar (1992) and then generalized and refined by Dai (1995), Chen (1995), Dai and Meyn (1995), Stolyar (1995) and Bramson (1998). Partial converse to this result is also

given in Meyn (1995), Dai (1996) and Puhalskii and Rybko (2000). In Subsection 3.1 we present a multiclass fluid network corresponding to the multiclass queueing network described in Section 2, and then quote the results mentioned above that will be used in the proof of Theorem 2.1.

3.1. A multiclass fluid network model

Parallel to the queueing network model $(\mathcal{J}, \mathcal{K}, \boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{C}, \boldsymbol{P}, \boldsymbol{\pi})$, a corresponding fluid network model which is also characterized by the same set of parameters, is obtained intuitively by replacing the discrete jobs in the queueing network with continuous fluids. Specifically, a class k fluid may flow exogenously into the network at the rate α_k, then be served at the station $\sigma(k)$, and after being served, a fraction $p_{k\ell}$ of fluid turns into a class ℓ fluid and the remaining fraction, $1 - \sum_{\ell=1}^{K} p_{k\ell}$, flows out of the network. When station $\sigma(k)$ devotes its full capacity to serving class k fluid (assuming that it is available to be served), it generates an outflow of class k fluid at rate μ_k. Among classes, fluid follows a priority service discipline, which is again described by the one-to-one mapping $\boldsymbol{\pi}$.

Define the fluid level process $\bar{Q}(t) = \{\bar{Q}_k(t),\ k \in \mathcal{K}\}$, where $\bar{Q}_k(t)$ denotes the fluid level of class k at time t; the time allocation process $\bar{T}(t) = \{\bar{T}_k(t),\ k \in \mathcal{K}\}$, where $\bar{T}_k(t)$ denotes the total amount of time that station $\sigma(k)$ has devoted to serving class k fluid during the time interval $[0, t]$; and the unused capacity process $\bar{Y}(t) = \{\bar{Y}_k(t),\ k \in \mathcal{K}\}$, where $\bar{Y}_k(t)$ denotes the (cumulative) unused capacity of station $\sigma(k)$ during the time interval $[0, t]$ after serving all classes at station $\sigma(k)$ which have a priority no less than class k (including class k). Let

$$H_k = \{\ell \colon \sigma(\ell) = \sigma(k),\ \pi(\ell) \le \pi(k)\},$$

be the set of indices for all classes that are served at the same station as class k and have a priority no less than that of class k. Note that $k \in H_k$ by definition. Thus, the dynamics of the fluid network model can be described as follows: for $k \in \mathcal{K}$ and $t \ge 0$,

$$Q_k(t) = Q_k(0) + \alpha_k t + \sum_{\ell=1}^{K} p_{\ell k}\mu_\ell T_\ell(t) - \mu_k T_k(t) \ge 0, \tag{3.1}$$

$$\bar{T}_k(\cdot) \text{ is nondecreasing with } \bar{T}_k(0) = 0, \tag{3.2}$$

$$\bar{Y}_k(t) = t - \sum_{\ell \in H_k} \bar{T}_\ell(t) \text{ is nondecreasing}, \tag{3.3}$$

$$\int_0^\infty \bar{Q}_k(t)\, \mathrm{d}\bar{Y}_k(t) = 0. \tag{3.4}$$

Equation (3.1) is the flow balance relation. Equation (3.3) describes the equivalent relation between the time allocation process $\bar{T}(t)$ and the unused capacity process $\bar{Y}(t)$. Equation (3.4) specifies both the work-conserving condition and the priority discipline, that is, for each k, (3.4) implies that at any time t there could exist some positive remaining capacity (rate) for serving those classes at station $\sigma(k)$ that have a strictly lower priority than class k, provided the fluid levels of all classes in H_k (having a priority no less than k) are zero. Particularly, for each lowest fluid class k at station $j = \sigma(k)$, (3.4) specifies the work-conserving condition for station j, which implies that station j cannot be idle if the total fluid level $(\sum_{\ell \colon \sigma(\ell)=j} \bar{Q}_\ell(t))$ in station j is positive at any time $t \ge 0$.

We shall refer to this network as the fluid network $(\mathcal{J}, \mathcal{K}, \boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{C}, \boldsymbol{P}, \boldsymbol{\pi})$. A pair $(\bar{Q}, \bar{T})$ (or equivalently $(\bar{Q}, \bar{Y})$) is said to be a *fluid solution* if they jointly satisfy (3.1)–(3.4). For

https://doi.org/10.1239/jap/1183667404 Published online by Cambridge University Press

convenience, we also call $\bar{Q}$ a fluid solution if there is a $\bar{T}$ such that the pair $(\bar{Q}, \bar{T})$ is a fluid solution. The fluid network is said to be *stable* if there is a time $\tau \geq 0$ such that $\bar{Q}(\tau + \cdot) \equiv 0$ for any fluid solution $\bar{Q}$ with $||\bar{Q}(0)|| = 1$, and it is said to be *weakly stable* if $\bar{Q}(\cdot) \equiv 0$ for any fluid solution $\bar{Q}$ with $\bar{Q}(0) = 0$. A well-known property we will use later in this paper is that the processes $\bar{Q}$, $\bar{Y}$, and $\bar{T}$ are Lipschitz continuous and hence, are differentiable almost everywhere on $[0, \infty)$. We summarize some known stability results on the relation between the queueing network model and its corresponding fluid network model, which are used in the proof of Theorem 2.1.

Theorem 3.1. *Consider the queueing network* $(\mathcal{J}, \mathcal{K}, \boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{C}, \boldsymbol{P}, \boldsymbol{\pi})$.

(1) *If the corresponding fluid network* $(\mathcal{J}, \mathcal{K}, \boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{C}, \boldsymbol{P}, \boldsymbol{\pi})$ *is stable, then the queue length process* Q *is positive recurrent.*

(2) *If the corresponding fluid network* $(\mathcal{J}, \mathcal{K}, \boldsymbol{\alpha}, \boldsymbol{m}, \boldsymbol{C}, \boldsymbol{P}, \boldsymbol{\pi})$ *is not weakly stable, then the queue length process* Q *is transient.*

Readers are referred to works of Chen and Yao (2001) and Dai (1996) for elementary proofs of the two conclusions, respectively.

3.2. Proof of Theorem 2.1

Proof of Theorem 2.1(1). According to Theorem 3.1(1), it is sufficient to show that the fluid network model corresponding to the modified KRSS queueing network, called the modified KRSS fluid network below, is stable. As an instance of the fluid network model described in (3.1)–(3.4), the dynamics of the modified KRSS fluid network can be detailed as follows.

$$\begin{gathered}
\bar{Q}_k(t) = \bar{Q}_k(0) + \alpha_k t - \mu_k \bar{T}_k(t) \geq 0, \qquad k = 1, 3, 7, 8, \\
\bar{Q}_k(t) = \bar{Q}_k(0) + \mu_\ell \bar{T}_\ell(t) - \mu_k \bar{T}_k(t) \geq 0, \qquad (k, \ell) = (5, 1), (2, 5), (6, 3), (4, 6), \\
\bar{T}_k(\cdot) \text{ is nondecreasing with } \bar{T}_k(0) = 0, \qquad k = 1, \ldots, 8, \\
\bar{Y}_k(t) = t - \bar{T}_k(t) \text{ is nondecreasing}, \qquad k = 4, 2, 7, 8, \\
\bar{Y}_k(t) = t - \bar{T}_\ell(t) - \bar{T}_k(t) \text{ is nondecreasing}, \qquad (k, \ell) = (1, 4), (3, 2), (5, 7), (6, 8), \\
\int_0^\infty \bar{Q}_k(t)\, \mathrm{d}\bar{Y}_k(t) = 0, \qquad k = 1, \ldots, 8.
\end{gathered} \tag{3.5}$$

First, we note that there exists a time $\tau_1 \geq 0$, such that

$$\bar{Q}_7(t) = \bar{Q}_8(t) = 0 \quad \text{for any } t \geq \tau_1, \tag{3.6}$$

as classes 7 and 8 have priority and hence, their fluids will drain within a finite time and then remain empty.

Next, we prove that there exists a time $\tau_2 \geq \tau_1$, such that

$$\bar{Q}_4(t) = \bar{Q}_2(t) = 0 \quad \text{for any } t \geq \tau_2. \tag{3.7}$$

Under (3.6), we have $\dot{\bar{Q}}_7(t) = \dot{\bar{Q}}_8(t) = 0$, and then $\dot{\bar{T}}_7(t) = \alpha_7 m_7$ and $\dot{\bar{T}}_8(t) = \alpha_8 m_8$ for all time $t \geq \tau_1$. (Here, we use a dot to denote the derivative of a process with respect to time t.) Combined with (3.5), this yields

$$\dot{\bar{Y}}_6(t) = 1 - \dot{\bar{T}}_6(t) - \dot{\bar{T}}_8(t) \geq 0 \quad \text{and} \quad \dot{\bar{T}}_6(t) \leq 1 - \dot{\bar{T}}_8(t) = 1 - \alpha_8 m_8 \quad \text{for } t \geq \tau_1.$$

https://doi.org/10.1239/jap/1183667404 Published online by Cambridge University Press

Then, we have

$$\dot{\bar{Q}}_4(t) = \mu_6 \dot{\bar{T}}_6(t) - \mu_4 \dot{\bar{T}}_4(t) \le \mu_6(1 - \alpha_8 m_8) - \mu_4 < 0,$$

for any $t \ge \tau_1$, where the last inequality is implied using the assumption that $m_6/(1-\alpha_8 m_8) > m_4$. Let $\tau_2' = \dot{\bar{Q}}_4(\tau_1)/(\mu_4 - \mu_6(1-\alpha_8 m_8))$. Then, we have

$$\bar{Q}_4(t) = 0, \quad \text{for any } t \ge \tau_2'. \tag{3.8}$$

Similarly, we have

$$\bar{Q}_2(t) = 0, \quad \text{for any } t \ge \tau_2'' = \frac{\dot{\bar{Q}}_2(\tau_1)}{\mu_2 - \mu_5(1-\alpha_7 m_7)}. \tag{3.9}$$

Let

$$\tau_2 = \max\left(\frac{1 + \Delta\tau_1}{\mu_4 - \mu_6(1-\alpha_8 m_8)}, \frac{1+\Delta\tau_1}{\mu_2 - \mu_5(1-\alpha_7 m_7)} \right),$$

with Δ being the Lipschitz constant for the fluid level process $\bar{Q}(t)$. Then, we have that $\tau_2 \ge \max(\tau_2', \tau_2'')$, noting that $||\bar{Q}(\tau_1)|| \le ||\bar{Q}(\tau_1)|| + M\tau_1 \le 1 + M\tau_1$. Now, (3.8) and (3.9) imply (3.7).

Finally, we prove that there exists a time $\tau \ge \tau_2 (\ge 0)$, such that

$$\bar{Q}_k(t) = 0, \quad \text{for } k = 1, 3, 5, 6 \text{ and } t \ge \tau, \tag{3.10}$$

which together with equations (3.6) and (3.7) implies that $\bar{Q}(t) = 0$ for $t \ge \tau$. Let

$$\begin{aligned}
\bar{W}_1(t) &:= m_1 \bar{Q}_1(t) + m_4(\bar{Q}_3(t) + \bar{Q}_6(t)) = (\alpha_1 m_1 + \alpha_3 m_4)t - (\bar{T}_1(t) + \bar{T}_4(t)), \\
\bar{W}_2(t) &:= m_3 \bar{Q}_3(t) + m_2(\bar{Q}_1(t) + \bar{Q}_5(t)) = (\alpha_1 m_2 + \alpha_3 m_3)t - (\bar{T}_2(t) + \bar{T}_3(t)), \\
\bar{W}_3(t) &:= m_5(\bar{Q}_1(t) + \bar{Q}_5(t)) = \alpha_1 m_5 t - \bar{T}_5(t), \\
\bar{W}_4(t) &:= m_6(\bar{Q}_3(t) + \bar{Q}_6(t)) = \alpha_3 m_6 t - \bar{T}_6(t),
\end{aligned}$$

for $t \ge \tau_2$. Here, $\bar{W}_i(t)$ $(i = 1, 2, 3, 4,)$ represents the immediate workload for station i implied in the system at time t. Define

$$\begin{aligned}
f_1(t) &:= m_6 \bar{W}_1(t), \qquad f_2(t) = m_5 \bar{W}_2(t), \\
f_3(t) &:= m_2 \bar{W}_3(t), \qquad f_4(t) = m_4 \bar{W}_4(t).
\end{aligned}$$

Then, it is straightforward to verify that, for $t \ge \tau_2$,

$$\dot{f}_i(t) < 0 \quad \text{if } \bar{Q}_i(t) > 0, \quad \text{for } i = 1, 2, 3, 4,$$

and

$$\begin{aligned}
f_1(t) \le f_4(t) \quad &\text{if } \bar{Q}_1(t) = 0, \qquad f_2(t) \le f_3(t) \quad \text{if } \bar{Q}_3(t) = 0, \\
f_3(t) \le f_2(t) \quad &\text{if } \bar{Q}_5(t) = 0, \qquad f_4(t) \le f_4(t) \quad \text{if } \bar{Q}_6(t) = 0.
\end{aligned}$$

Now applying the piecewise linear Lyapunov function approach for the multiclass fluid network model described in Theorem 3.1 of Chen and Ye (2002), we obtain (3.10).

https://doi.org/10.1239/jap/1183667404 Published online by Cambridge University Press

Proof of Theorem 2.1(2). According to Theorem 3.1(2), we need to show that the modified KRSS fluid network is not weakly stable. Similar to the above proof of Theorem 2.1(1), it is not difficult to show that there exists a time $\tau_1 \geq 0$, such that

$$\bar{Q}_7(t) = \bar{Q}_8(t) = 0, \quad \text{for any } t \geq \tau_1,$$

as classes 7 and 8 fluids have higher priorities at stations 7 and 8 respectively and thus, that there exists a time $\tau_2 \geq \tau_1$, such that

$$\bar{Q}_5(t) = \bar{Q}_6(t) = 0, \quad \text{for any } t \geq \tau_2,$$

as the remaining service capacity for classes 5 and 6 fluids is greater than that for class 1 and 3 fluids. Thus, the modified KRSS fluid network is reduced to the well-known KRSS fluid network, which is not weakly stable under the condition (2.1).

4. Discussion and concluding remark

In this paper we have presented a paradox for the admission control of a multiclass queueing network with differentiated service. This paradox is, to our knowledge, the first one of its kind, which is complementary to the existing ones on the service rate control and the routeing control.

The models, as well as the admission control and the differentiated service, studied in this paper are simplified and idealized models of practical systems. Take the semiconductor production as an example. The production line may consist of tens of processing stations (machines), and parts may require tens or even hundreds of stages of processing by the stations. The admission control may model the *central control* on whether to accept external orders, while the differentiated priority for jobs at each station could be due to the *local control* on scheduling jobs. It would not be surprising for the paradox to exist in such a complex system. It is known that there exist approaches like the max-weight type policy (e.g. Stolyar 2004) for stabilizing the network. However, these approaches may not be feasible to implement in all cases, and so simple stabilization techniques, such as artificially reducing service capacities and increasing workload, are still of interest. Therefore, the detection of and the remedy to such a paradoxical phenomenon are interesting future research topics.

References

[1] Braess, D. (1968). Über ein Paradoxon aus der Verkehrsplanung. *Unternehmenforschung* **12,** 258–268.

[2] Bramson, M. (1994). Instability of FIFO queueing networks. *Ann. Appl. Prob.* **4,** 414–431.

[3] Bramson, M. (1998). Stability of two families of queueing networks and a discussion of fluid limits. *Queueing Systems Theory Appl.* **23,** 7–31.

[4] Chen, H. (1995). Fluid approximations and stability of multiclass queueing networks: work-conserving discipline. *Ann. Appl. Prob.* **5,** 637–655.

[5] Chen, H. and Yao, D. D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization*. Springer, New York.

[6] Chen, H. and Ye, H. Q. (2002). Piecewise linear Lyapunov function for the stability of priority multiclass queueing networks. *IEEE Trans. Automatic Control* **47,** 564–575.

[7] Chen, H. and Zhang, H. (2000). Stability of multiclass queueing networks under priority service disciplines. *Operat. Res.* **48,** 26–37.

[8] Cohen J. E. and Kelly, F. P. (1990). A paradox of congestion in a queueing network. *J. App. Prob.* **27,** 730–734.

[9] Dai, J. G. (1995). On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid models. *Ann. Appl. Prob.* **5,** 49–77.

[10] Dai, J. G. (1996). A fluid-limit model criterion for instability of multiclass queueing networks. *Ann. Appl. Prob.* **6,** 751–757.

[11] Dai, J. G. and Meyn, S. P. (1995). Stability and convergence of moments for multiclass queueing networks via fluid models. *IEEE Trans. Automatic Control* **40,** 1899–1904.

[12] Dai, J. G. and Vande Vate, J. H. (1996). Global stability of two-station queueing networks. In *Proc. Workshop Stoch. Networks Stability Rare Events*, eds P. Glasserman, K. Sigman and D. Yao, Springer, New York, pp. 1–26.
[13] Dumas, V. (1997). A multiclass network with non-linear, non-convex, non-monotonic stability conditions. *Queueing Systems Theory Appl.* **25,** 1–43.
[14] Humes, C. (1994). A regulator stabilization technique: Kumar–Seidman revisited. *IEEE Trans. Automatic Control* **39,** 191–196.
[15] Kumar, P. R. and Seidman, T. I. (1990). Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Trans. Automatic Control* **35,** 289–298.
[16] Lu, S. H. and Kumar, P. R. (1991). Distributed scheduling based on due dates and buffer priorities. *IEEE Trans. Automatic Control* **36,** 1406–1416.
[17] Meyn, S. (1995). Transience of multiclass queueing networks via fluid limit models. *Ann. Appl. Prob.* **5,** 946–957.
[18] Puhalskii, A. and Rybko, A. N. (2000). Non-ergodicity of queueing networks under non-stability of their fluid models. *Prob. Inf. Transmission* **36,** 26–48.
[19] Rybko, A. N. and Stolyar, A. L. (1992). Ergodicity of stochastic processes describing the operations of open queueing networks. *Problemy Peredachi Informatsii* **28,** 2–26.
[20] Stolyar, A. L. (1995). On the stability of multiclass queueing network: a relaxed sufficient condition via limiting fluid processes. *Markov Process. Rel. Fields* **1,** 491–512.
[21] Stolyar, A. L. (2004). Max-weight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. *Ann. Appl. Prob.* **14,** 1–53.