Queueing Systems 49, 335–361, 2005

Balancing Queues by Mean Field Interaction*

DONALD A. DAWSON
School of Mathematics and Statistics, Carleton University, Ottawa, Ontario Canada K1S 5B6

JIASHAN TANG
School of Mathematics and Statistics, Carleton University, Ottawa, Ontario Canada K1S 5B6; Department of Applied Mathematics and Physics, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210003, P. R. China

YIQIANG Q. ZHAO
School of Mathematics and Statistics, Carleton University, Ottawa, Ontario Canada K1S 5B6

Received 18 July 2003; Revised 6 November 2004

Abstract. Consider a queueing network with N nodes in which queue lengths are balanced through mean-field interaction. When N is large, we study the performance of such a network in terms of limiting results as N goes to infinity.

Keywords: queue, birth death process, mean-field, nonlinear master equation, law of large numbers

AMS subject classification: 60K25, 60K35, 60J27, 60F05

1. Introduction

The objective of this study is to investigate a mechanism of load balancing in a queueing network using ideas from the mean-field model of interacting systems. In stochastic service systems, especially in networks which consist of a number of parallel queues, load balancing is often used to improve the system performance by shortening the queue length, reducing the waiting time, and increasing the system throughput. The purpose of introducing such a balancing mechanism is essentially to change the input and output structures so as to improve the quality of service of the system. In practice, there are various balancing mechanisms which have been used in different cases, which can be roughly categorized into the following two groups.

1. Methods based on changing the service rate. These kinds of methods consist of two cases: In the first, the service rate in each queue will be changed according to its queue length while the total resources of the system remain unchanged, see for example [27]. This method has been adopted in many processor sharing systems. In the other case, as the service rate changes the total resources available to the system are also

*This project is supported by the Natural Sciences and Engineering Research Council of Canada through research grants.

changed. Service systems with a state-dependent service rate belong to this case, see for example [14].

2. Methods based on changing the arrival rate. Similarly, these kinds of methods can also be divided into two cases: In one case, the total arrival rate to the system is fixed. Methods related to the classical joining the shortest queue strategy, including the model with jockeying [18], often belongs to this category. For example, instead of joining the shortest queue, an arriving customer may join the queue with the shortest expected waiting time [36]. Recently, [13] considered a generalized model which includes joining the shortest queue as a special case. In the other case, the total arrival rate to the system is also changed according to the state of the system. Most of the work involves changing the arrival rate at the epoch at which there is a new arrival or a service completion [17,28]. The arrival rate can be also changed at any time according to the amount of instantaneous unfinished work in the system [21].

Other kinds of balancing ways also exist, for example the combination of the above two methods [2].

However, all the above work can be only used to deal with models consisting of a small number of queues, often only two queues. The exact performance analysis of a system consisting of a large number, say N, of queues is usually considered very challenging. The main reason is that after introducing a balancing strategy, the system becomes a high dimension interacting system. As N becomes large, the system becomes too complicated to be studied.

With the advances of the internet and other applications of stochastic service systems, a system consisting of a large number of queues is of considerable interest. So, we need to find methods to investigate such a system. In fact, if we regard such a system as an interacting system, then a natural candidate to investigate the behavior of such systems is the mean-field methodology. From the mean-field point of view, the performance analysis of a large system could be done by investigating the limiting behavior as $N \to \infty$. The major advantage of this method is that the limit can often be found by solving a deterministic system [5–7]. In this paper, we will use this method to investigate a queueing system consisting of a large number of queues.

Mean-field approximation is an important aspect of the limit theory of stochastic processes, which was first used in statistical physics [5,19]. The mean-field approach has been also proven to be useful in other applications including queueing theory. Dobrushin's mean-field model [33,34] or supermarket model [25] provides such an example. The system they considered consists of a large number, N, of identical infinite buffer size FCFS single servers, the arrival process is a single Poisson stream with rate λN, and the service times are i.i.d. Each arriving custom first chooses m, which is far smaller than N, out of N servers at random, and then joins a server out of the m chosen servers on the basis of some fixed dynamic routing policy, such as entering the sth shortest queue ($1 \leq s \leq m$). Some properties of such a system are investigated as $N \to \infty$. For more results and related methods, we refer the readers to [23,26,30,31,35] and references therein.

Recently, some new results on telecommunication networks with TCP connections by using mean-field approaching methodology are obtained in [1] and [3] (see also the survey in [32]). Similar ideas are also used in the fluid model in [16] (see also [22,24]).

Our study in this paper was motivated by balancing queues in stochastic networks. We use the mean-field method to investigate a system which is different from that in the papers mentioned above. The main objective of this research is to investigate some limiting behavior using the mean-field approach and to identify the stationary distribution of the limiting queueing system. As a consequence, comparison results are given to show the improvement in system performance after introducing mean-field interaction.

One may notice that the system with mean-field interaction studied in this paper is a system in continuous time and the main result obtained is valid only for large N. However, as shown in our study, such mean-field interaction is only needed at discrete time epochs and the error which occurs, when the main result is applied for a finite value of N around or larger than a few hundreds, is usually negligible for practical purpose. This means that our study and results can be practically applied in many interesting queueing networks.

2. Description of the model

We begin with some notation to be used in the subsequent sections. Let $E = \{0, 1, 2, \ldots\}$ be equipped with the discrete topology. For $T \geq 0$, $D_T(E)$ ($D_\infty(E)$) denotes the space of functions from $[0, T]$($[0, \infty)$) to E that are right-continuous and have a left limit. Let $\mathcal{F}_t = \sigma\{X(s), 0 \leq s \leq t\}$, be the smallest σ-field generated by $\{X(s), 0 \leq s \leq t\}$, where $X(s, w) = X_s(w)$. Also we denote $X_s(w) = w(s)$, where $w \in D_T(E)$. Let $\mathcal{F} = \sigma\{X_t, t \geq 0\}$ and let $\mathcal{P}(D_\infty(E), \mathcal{F})$ be the set of all probability measures on $(D_\infty(E), \mathcal{F})$. For integer $p > 0$, let $\mathcal{P}_p(E)$ be the set of all probability measures on E with finite pth moment. $\mathcal{P}(E)$ denotes the set of all probability measures on E.

Consider a single state-dependent $M/M/1$ queue with arrival rate λ_i, $i = 0, 1, \ldots,$ and service rate μ_i, $i = 1, 2, \ldots$. The queue length process $Y(t)$ is a birth-death process whose Q-matrix is given by

$$Q = (q_{ij}) = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & \cdots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & \cdots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{1}$$

For $N \geq 1$, we define a system consisting of N queues with mean-field interaction as described below as essentially a time inhomogeneous birth-death process. At $t = 0$, for $1 \leq j \leq N$, the arrival rate to the jth queue occurs according to $\{\lambda_i\}_{i=0,1,\ldots}$, and the service rate at queue j is $\{\mu_i\}_{i=1,2,\ldots}$. For $t > 0$, we introduce a function h, called the *interaction function*, to capture the mean-field interaction between queues. Let $h(x)$ be

a continuous nondecreasing function from $\mathbf{R}$, the set of all real numbers, to $\mathbf{R}$, whose first derivative is bounded above by $K_0 > 0$, such that:

$$h(0) = 0, \quad \lim_{x\to\infty} h(x) > 0, \quad h(k) \le \lambda_k \text{ for any } k \ge 0. \tag{2}$$

Let $X_j(t)$ denote the queue length of the jth queue, $1 \le j \le N$, at time t, and let $v_N(t) = \frac{1}{N}\sum_{j=1}^{N} X_j(t)$ denote the mean queue length of the N queues at time t. The arrival rate to the jth queue at time t is a modification of $\lambda_{X_j(t)}$, obtained by introducing a mean-field interaction, namely $\lambda_{X_j(t)} - h(X_j(t) - v_N(t))$, $j = 1, 2, \ldots, N$. This implies that: If $X_j(t) = v_N(t)$, i.e. the queue length of the jth queue is equal to the mean queue length of the whole system at time t, then the arrival rate to the jth queue is the same as $\lambda_{X_j(t)}$. If $X_j(t) > v_N(t)$, then the arrival rate to the jth queue decreases to $\lambda_{X_j(t)} - h(X_j(t) - v_N(t))$. If $X_j(t) < v_N(t)$, then the arrival rate increases to $\lambda_{X_j(t)} - h(X_j(t) - v_N(t))$. This system is referred to as a *mean-field interaction system*.

For the above system, let

$$U_N(t) = \frac{1}{N}\sum_{j=1}^{N} \delta_{X_j(t)}, \quad U_N = \frac{1}{N}\sum_{j=1}^{N} \delta_{X_j(\cdot)}, \tag{3}$$

where δ_x is the Dirac measure on a single point x. Then $U_N(t)$ is the empirical probability measure on E, the empirical distribution of queue length of the N queues at time t. The first moment of $U_N(t)$ is

$$\|U_N(t)\| = \langle U_N(t)(dx), x\rangle = \frac{1}{N}\sum_{j=1}^{N} X_j(t), \tag{4}$$

where for $u \in \mathcal{P}(E)$, $\langle u, f\rangle = \sum_{k\in E} f(k)u(\{k\})$. $\|U_N(t)\|$ is exactly the value of $v_N(t)$.

In this paper we will study the limiting behavior and the law of large numbers of U_N and/or $U_N(t)$ defined in (3).

The study of this paper was directly motivated by the concept "balancing", which has been often and widely employed in computer and telecommunication networks, in various aspects of modelling, controlling and scheduling. For example, [3] considers the problem of windows control under TCP connections combined with RED technology. In literature, definitions of the mean-field interaction are numerous, and the one adapted in this paper is close to the one in [5]. The limiting results obtained in this paper can help us understand the performance of a "typical" queue in the mean-field system. The following remarks suggest that this "typical" queue can be possibly used as an approximation to an individual queue in the mean-field system when the number of queues is large. However, a rigorous analysis on the speed of convergence of the finite-queue system to its corresponding infinite-queue system and the error estimation in approximating the finite-queue system using the "typical" queue should be carried out first to fully realize the potential value of this mean-field technology, which is beyond the scope of the current paper and will be addressed separately.

Remark 1. From the definition of the mean-field system provided above, the knowledge of the values of $X_j(t)$ is required for all j and t for the analysis. However, since $X_j(t)$ changes its value only at the instant of a new arrival or a service completion, the implementation of the mean-field interaction technology, or modification of the arrival rate, will be only necessary at these discrete time epochs.

Remark 2. In applications, N is usually in a range from several hundreds to several thousands. For such a value of N, the mean-field interaction technology introduced in this paper can be practically implemented without any difficulty. Our simulation results have shown that the error which occurs when using a mean-field interaction system with infinitely many queues to approximate the corresponding finite-queue system is usually negligible.

Remark 3. The method in this paper can also be used to deal with systems to which there are a bulk arrivals of customers. Although we have only modified the arrival rate here, modifications of the service rate can also be investigated in a similar way.

Remark 4. The formulation given above indicates that the usual mean-field method can only be applied to investigate networks in which all queues are symmetric. The idea of local mean-field interaction in [8] could be applied to deal with one kind of asymmetric network. For example, let l be a fixed positive integer. A network consisting of $N > l$ queues satisfies that: The first l queues are asymmetric and the remaining $N - l$ queues are symmetric. The limiting behavior of such a local mean-field interaction network, as N tends to infinity, could be considered using the ideas of [8].

In the detailed study provided in the following sections, analysis will be carried out for a more general interaction function h and a Q-matrix which satisfies certain conditions specified below.

Let $h(x, y)$ be a continuous function from $\mathbf{R}_+ \times \mathbf{R}_+$ to $\mathbf{R}$, where $\mathbf{R}_+$ is the set of all nonnegative numbers, which increases in x and decreases in y, and satisfies $\lim_{x\to\infty} h(x, y) > 0$ for any $y \in \mathbf{R}_+$. There exist positive constants K_1, K_2 and K_3 such that

$$\begin{aligned} &h(k, 0) \le \lambda_k, \quad k \ge 0, \\ &\text{if } y > K_1, \quad \text{then } h(0, y) \ge -K_2 y, \end{aligned} \tag{5}$$

and for any $(x_1, y_1), (x_2, y_2) \in \mathbf{R}_+ \times \mathbf{R}_+$,

$$(h(x_2, y_2) - h(x_1, y_1))^+ \le K_3((x_2 - x_1)^+ + (y_1 - y_2)^+), \tag{6}$$

where $x^+ = x \vee 0 = \max\{x, 0\}$. Also, throughout the paper $x \wedge y = \min\{x, y\}$. Evidently, if $\tilde{h}(x)$ is defined as in (2), then $h(x, y) = \tilde{h}(x - y)$ satisfies (5) and (6).

If in the mean-field interaction system defined earlier we replace the arrival rate to the jth queue at time t by $\lambda_{X_j(t)} - h(X_j(t), v_N(t))$, $j = 1, \ldots, N$, then the resulting

interaction system is more general. For the rest of this paper, we will consider this more general system.

In order to prove limiting properties, we need to impose some conditions on the Q-matrix defined by (1). Assume that λ_i and μ_i are dominated, as specified in (7),(8) and (9), by a polynomial $\alpha(x)$ with positive coefficients, of degree $m \geq 0$ such that

$$\sup_j \left\{ \frac{\lambda_j}{\alpha(j)}, \frac{\mu_j}{\alpha(j)} \right\} < 1. \tag{7}$$

For any $n \geq 1$, there exist two n-dependent constants $K_4(n) \geq 0$ and $K_5(n) \geq 1$ such that

$$\lambda_i((i+1)^n - i^n) + \mu_i((i-1)^n - i^n) \leq K_4(n) + K_5(n)i^n, \quad \text{for } i \geq 0. \tag{8}$$

In addition,

$$K_6 \stackrel{\Delta}{=} \sup_{0 \leq i \leq j} \left\{ \frac{(\lambda_j - \mu_j) - (\lambda_i - \mu_i)}{j - i} \right\} < \infty. \tag{9}$$

(7) simply implies that as a function of j, both λ_j and μ_j are controlled by a polynomial. Roughly speaking, (8) often implies that the arrival rate is not too much larger than the service rate.

Obviously, (8) holds if the arrival rate λ_i and the service μ_i are either constants or linear functions of i. The condition in (9) guarantees the uniqueness of some involved Q-processes.

The rest of paper is organized as follows. In Section 3, we investigate the solution of a class of master equations. In Section 4, we prove the convergence of U_N to the solution of the master equation. In Section 5, we discuss the stationary distribution of the solution of the master equation.

3. Solution to the nonlinear master equation

In this section, we will investigate a class of nonlinear master equations whose solution is just the limit of $U_N(\cdot)$ as $N \to \infty$.

Recall that for the birth-death process $\{Y(t), t \geq 0\}$ with its Q-matrix given by (1), there exists an infinitesimal generator Ω defined by

$$\Omega f(i) = \sum_{j \in E} q_{ij} f(j) = \sum_{j \neq i} q_{ij}(f(j) - f(i)), \tag{10}$$

where $f \in C_b(E)$ is an arbitrary bounded continuous functions in E, such that

$$f(Y(t)) - \int_0^t \Omega f(Y(s))ds \tag{11}$$

is a martingale.

For our interaction system, for any nonnegative function $g(t)$, we define the following operator:

$$\Omega_{h,g(t)} f(i) = \sum_{j \neq i} q_{ij}(f(j) - f(i)) - h(i, g(t))(f(i+1) - f(i)). \tag{12}$$

The so called nonlinear master equation has the following form

$$\frac{\mathrm{d}\langle u(t), f\rangle}{\mathrm{d}t} = \left\langle u(t), \Omega_{h,\|u(t)\|} f \right\rangle, \quad f \in C_b(E), \tag{13}$$

where $u(\cdot)$ is a measure-valued function from $[0, +\infty]$ to $\mathcal{P}(E)$.

The nonlinear master equations proposed by [29] are used to describe a class of nonlinear pure jump Markov processes in chemistry, physics and biology. Later [37] re-established them using probabilistic methods. The nonlinear system of the master equations are analogous to nonlinear diffusion equations and are not only interesting in themselves but also play an important role in the formulation of the mean-field interaction system. The reason is that the solution of the nonlinear equations are often the limit of some class of the empirical probability measure of the corresponding mean-field interaction system.

In this section we will investigate the solution of (13).

Definition 1. Let $u \in \mathcal{P}(E)$. $P \in \mathcal{P}(D_\infty(E), \mathcal{F})$ is called a solution of the nonlinear master equation with initial distribution u if its marginal distribution $u_t(\cdot) = P \circ X_t^{-1}(\cdot)$ satisfies (13) and $u_0 = u$. Moreover, P is called a q-solution if, in addition, it is Markovian in the sense of McKean [15], i.e. for any $j \in E$,

$$P(X_{t+s} = j | \mathcal{F}_t) = p(t, X_t, t+s, j), \quad P - a.s., \tag{14}$$

where transition function $p(s, i, t, j)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}s} p(t, i, t+s, j) = \sum_{k \in E} p(t, i, t+s, k) \Omega_{h,\|u_{t+s}\|} I_{\{j\}}(k), \quad t \geq 0. \tag{15}$$

Remark 5. Equation (15) is essentially the forward equation. However, the interacting measure u_t is dependent on the initial value $u_0 = u$. If the interacting measure is fixed then the Markovian property (in the sense of McKean) will degenerate to the usual Markovian property. Moreover, the unique solution of (15) is also the minimal solution [9].

We have the following result.

Theorem 1. The q-solution of (13) exists and is unique.

In order to prove this theorem, motivated by (10) and (11), it is convenient to introduce the following martingale problem.

Definition 2. Let $u \in \mathcal{P}(E)$. $P \in \mathcal{P}(D_\infty(E), \mathcal{F})$ is called a solution of the martingale problem $[u, \Omega_{h,\|u_t\|}]$, if

(1) $P \circ X_0^{-1} = u$;
(2) $P \circ X_t^{-1} = u_t$;
(3) For any $j \in E$,

$$\left(I_{\{j\}}(X_t) - \int_0^t \Omega_{h,\|u_s\|} I_{\{j\}}(X_s) ds, \mathcal{F}_t, P\right) \tag{16}$$

is a martingale.

We will prove Theorem 1 by establishing the following.

Theorem 2.

(i) For $u \in \mathcal{P}_p(E)$, where $p > 1$, any q-solution of (13) is a solution of the martingale problem $[u, \Omega_{h,\|u_t\|}]$;
(ii) Solutions of the martingale problem exist;
(iii) The martingale problem has only one solution;
(iv) The unique solution of the martingale problem is a q-solution of (13).

If "$h(i, g(t))$" in (12) is replaced by "$-g(t)$", then Theorem 2 has been proved in [10] (see also [15] and [12]), where $Q = (q_{ij})$ is not necessarily triagular. Although the infinitesimal generator defined by (12) is different from that in [10], the proof is similar.

Before progressing to the proofs, let us introduce some topologies which will be used in the sequel.

Let $\mathcal{P}(D_\infty(E), \mathcal{F})$ be equipped with the usual weak topology and for $u, v \in \mathcal{P}_p(E)$, let $\rho(u, v)$ be the minimum L^p-analogue metric (see [15]) defined by

$$\rho(u, v) = \rho_p(u, v) = \inf_{F \in \mathcal{P}(u,v)} \left(\int \int |x - y|^p F(\mathrm{d}x, \mathrm{d}y) \right)^{\frac{1}{p}}, \tag{17}$$

where $\mathcal{P}(u, v)$ is the set of all probability measures F on $E \times E$ with marginals u and v. The space $\mathcal{P}_p(E)$ is equipped with a topology determined by $\rho(\cdot, \cdot)$.

For $u, v \in \mathcal{P}(E)$, we define an incomplete metric $\bar{\rho}$ as the following, which gives a vague topology on the space $\mathcal{P}(E)$,

$$\bar{\rho}(u, v) = \sum_{j=0}^{\infty} \frac{|u(\{j\}) - v(\{j\})|}{2^j}. \tag{18}$$

Note that the vague topology is weaker than the weak topology.

We prove Theorem 2 by a series of lemmas. Keep in mind that we will use $c_1, c_2, \ldots$ to denote different constants and $c_k(x)$ to denote the constant depending on parameter x.

First of all, similar to the proof of Lemma 3.1 in [10] we have that

Lemma 3. For any $u \in \mathcal{P}(E)$, every q-solution of (13) is a solution of the martingale problem defined by Definition 2.

The following Lemma is used to prove the existence of the solution to the martingale problem.

Lemma 4. For $p \geq 1$, let $f(x) = x^p$ and for $t \geq 0$, let $u(t) \in \mathcal{P}(E)$. If for $T > 0$ we have $c_1(T) = \sup_{0 \leq t \leq T} \|u(t)\| < \infty$, then there exists a constant $c_2(K_1, K_2, K_4, K_5, T, p) > 0$ such that for $0 \leq t \leq T$,

$$\Omega_{h,\|u_t\|} f(x) \leq c_2(K_1, K_2, K_4, K_5, T, p)(x^p + 1).$$

Proof. According to (12) and (8), we have

$$\begin{aligned}\Omega_{h,\|u_t\|} f(x) &= \mu_x(f(x-1) - f(x)) + \lambda_x(f(x+1) - f(x)) - h(x, \|u_t\|)(f(x+1) - f(x)) \\ &\leq K_4(p) + K_5(p)x^p - h(x, \|u_t\|) \cdot ((x+1)^p - x^p).\end{aligned}$$

Noting that $0 < (x+1)^p - x^p \leq p2^p x^{p-1} + 1$ and from hypothesis (5) we get

$$\begin{aligned}\Omega_{h,\|u_t\|} f(x) &\leq K_4(p) + K_5(p)x^p + K_1 \vee K_2\|u_t\|(p2^p x^p + 1) \\ &\leq (K_5(p) + K_2 c_1(T)p2^p)x^p + (K_4(p) + K_1 + K_2 c_1(T)) \\ &\leq c_2(K_1, K_2, K_4, K_5, T, p)(x^p + 1).\end{aligned}$$

□

Lemma 5. For every $u \in \mathcal{P}_p(E)$, there exists at least one solution to the martingale problem $[u, \Omega_{h,\|u_t\|}]$. Moreover, $u_t = P_u \circ X_t^{-1}$ is a continuous function from $[0, \infty)$ to $(\mathcal{P}(E), \rho)$.

Proof. The idea of the proof is to construct a solution to the martingale problem in three steps. The first step is to prove that for any positive integer n, there exists a $P_u^n \in \mathcal{P}(D_\infty(E), \mathcal{F})$ such that $P_u^n \circ X_0^{-1} = u$ and for every $j \in E$,

$$\left(I_j(X_t) - \int_0^t \Omega_{h,\|u_s^n\|} I_j(X_s) ds, \mathcal{F}_t, P_u^n\right) \tag{19}$$

is a martingale, where

$$u_s^n = \tilde{u}_{[ns]/n}^n, \quad \tilde{u}_s^n = P_u^n \circ X_s^{-1}. \tag{20}$$

In fact, intuitively P_u^n can be constructed as follows. For $t \in [0, \frac{1}{n})$, the process starts from $P_u^n \circ X_0^{-1} = u$ and the interacting measure is $P_u^n \circ X_0^{-1} = u = u_0$, therefore the process runs according to the infinitesimal generator $\Omega_{h,\|u_0\|}$. For $t \in [\frac{1}{n}, \frac{2}{n})$, the process starts from $P_u^n \circ X_{\frac{1}{n}}^{-1} = u_{\frac{1}{n}}$ and takes the interacting measure to be $u_{\frac{1}{n}}$, therefore the process runs according to the infinitesimal generator $\Omega_{h,\|u_{\frac{1}{n}}\|}$. Continue the construction in this fashion. The sample path of the stochastic process corresponding to P_u^n is obtained by gluing together these segments.

The second step is to show that the probability law P_u^n obtained above is convergent as n tends to infinity. The third step is to prove that the limit of P_u^n is exactly as we wanted it. The detailed procedure of the proof is similar to that of Theorem 1.1 and its Corollary in [10]. The key point is that for every $T > 0$,

$$\sup_n \sup_{0\le t\le T} \left\| u_t^n \right\| < \infty, \quad \sup_n \sup_{0\le t\le T} \left\| \tilde{u}_t^n \right\|_p < \infty. \tag{21}$$

Based on the estimates in Lemma 4, we can prove (21) by a method similar to that used to prove Lemma 2.3 in [10]. □

The following lemma is used to prove the uniqueness of the solution to the martingale problem.

Lemma 6. For every $u \in \mathcal{P}_p(E)$ and $v_t \in \mathcal{P}(E), t > 0$, if for any $0 < T < \infty$, $\sup_{0\le t\le T} \|v_t\| < \infty$, then there exists only one $P_{u,v.} \in \mathcal{P}(D_\infty(E), \mathcal{F})$ such that $P_{u,v.} \circ X_0^{-1} = u$ and for any $j \in E$,

$$\left(I_j(X_t) - \int_0^t \Omega_{h,\|v_s\|} I_j(X_s) ds, \mathcal{F}_t, P_{u,v.} \right)$$

is a martingale. Furthermore, $\phi(t, j) = P_{u,v.}(X_t = j)$ is the minimal nonnegative solution of the equation

$$\phi(t, j) = u(\{j\}) e^{\int_0^t q_{jj}(\tau) d\tau} + \int_0^t \sum_{k\ne j} \phi(s, k) q_{kj}(s) e^{\int_s^t q_{jj}(\tau) d\tau} ds, \tag{22}$$

where $Q(t) = (q_{ij}(t))$ is the inhomogeneous Q-matrix corresponding to $\Omega_{h,\|v_t\|}$, i.e.

$$q_{ij}(t) = \begin{cases} -\lambda_0 + h(0, \|v_t\|), & i = 0, j = 0, \\ \mu_i, & i \ge 1, j = i - 1, \\ \lambda_i - h(i, \|v_t\|), & j = i + 1, \\ -(\lambda_i + \mu_i) + h(i, \|v_t\|), & j = i \ge 1, \\ 0, & \text{others.} \end{cases}$$

Proof. The proof is similar to the proof of Lemma 3.2 in [10]. □

Lemma 7. For every $u \in \mathcal{P}_p(E)$, there exists at most one solution to the martingale problem $[u, \Omega_{h,\|u_t\|}]$.

Proof. For $u \in \mathcal{P}(E)$, let $P_u^1, P_u^2 \in \mathcal{P}(D_\infty(E), \mathcal{F})$ be any two solutions to the martingale problem $[u, \Omega_{h,\|u_t\|}]$ and let $u_t^l = P_u^l \circ X_t^{-1}$, $u_0^l = u$, $l = 1, 2$. From (21) we know that the condition in Lemma 6 holds for u_t^1 and u_t^2 and thus from Lemma 6 we know that in order to prove $P_u^1 = P_u^2$, it is sufficient to prove that $u_t^1 = u_t^2$ for $t \geq 0$. For this purpose, for $f \in C_b(E \times E)$, we define the coupling operator A_t as □

$$
\begin{aligned}
&A_t f(j_1, j_2)\\
&\quad = \big(\mu_{j_1} - \mu_{j_2}\big)^+(f(j_1{-}1, j_2) - f(j_1, j_2)) + \big(\lambda_{j_1} - \lambda_{j_2}\big)^+(f(j_1 + 1, j_2) - f(j_1, j_2))\\
&\qquad + \big(\mu_{j_2}{-}\mu_{j_1}\big)^+(f(j_1, j_2{-}1) - f(j_1, j_2)) + (\lambda_{j_2} - \lambda_{j_1})^+(f(j_1, j_2{+}1) - f(j_1, j_2))\\
&\qquad + \big(\mu_{j_1} \wedge \mu_{j_2}\big)(f(j_1 - 1, j_2 - 1) - f(j_1, j_2)) + (\lambda_{j_1} \wedge \lambda_{j_2})(f(j_1 + 1, j_2 + 1)\\
&\qquad - f(j_1, j_2)) + \big(\big(-h(j_1, \|u_t^1\|)\big){-}\big(-h\big(j_2, \|u_t^2\|\big)\big)\big)^+(f(j_1 + 1, j_2) - f(j_1, j_2))\\
&\qquad + \big(\big(-h(j_2, \|u_t^2\|\big)\big) - \big(-h\big(j_1, \|u_t^1\|\big)\big)\big)^+(f(j_1, j_2 + 1) - f(j_1, j_2))\\
&\qquad + \big(-h\big(j_1, \|u_t^1\|\big)\big) \wedge \big(-h\big(j_2, \|u_t^2\|\big)\big)(f(j_1 + 1, j_2 + 1) - f(j_1, j_2)).
\end{aligned}
\tag{23}
$$

From [4], there exists the minimal A_t process whose transition function is denoted by $p(t, (i_1, i_2), (j_1, j_2))_{(i_1,i_2),(j_1,j_2)\in E\times E}$. Suppose that F is a probability measure on $E \times E$ with marginals u_0^1 and u_0^2. We define

$$
F(t) = \sum_{(i_1,i_2)\in E\times E} F((i_1, i_2)) \sum_{(j_1,j_2)\in E\times E} p(t, (i_1, i_2), (j_1, j_2))|j_2 - j_1|. \tag{24}
$$

Recalling definition (17), we know that

$$
\rho_1\big(u_t^1, u_t^2\big) \leq F(t). \tag{25}
$$

The following lemma is needed to complete the proof.

Lemma 8. For any $(i_1, i_2) \in E \times E$, then

$$
\sum_{(j_1,j_2)\in E\times E} q_{(i_1,i_2),(j_1,j_2)}(t)|j_2 - j_1| \leq (K_6 + K_3)|i_2 - i_1| + K_3\rho_1\big(u_t^1, u_t^2\big). \tag{26}
$$

Proof. From (23) we know that

$$
\begin{aligned}
&\sum_{(j_1,j_2)\in E\times E} q_{(i_1,i_2)(j_1,j_2)}(t)|j_2 - j_1|\\
&\quad = \big\{\big(\mu_{i_1} - \mu_{i_2}\big)^+(|i_2 - (i_1 - 1)| - |i_2 - i_1|) + \big(\lambda_{i_1} - \lambda_{i_2}\big)^+(|i_2 - (i_1 + 1)| - |i_2{-}i_1|)\\
&\qquad + \big\{\big(\mu_{i_2}{-}\mu_{i_1}\big)^+(|(i_2 - 1) - i_1| - |i_2 - i_1|) + \big(\lambda_{i_2}{-}\lambda_{i_1}\big)^+(|(i_2 + 1) - i_1|{-}|i_2 - i_1|)\big\}\\
&\qquad + \big\{\big(h(i_2 - \|u_t^2\|\big) - h\big(i_1 - \|u_t^1\|\big)\big)^+(|i_2 - i_1 - 1| - |i_2 - i_1|)
\end{aligned}
$$

$$+\big(h\big(i_1-\|u_t^1\|\big)-h\big(i_2-\|u_t^2\|\big)\big)^+(|i_2-i_1-1|-|i_2-i_1|)\big\}$$
$$\triangleq \mathcal{K}_1+\mathcal{K}_2.$$

On one hand, if $i_1 < i_2$, then

$$\mathcal{K}_1=\big(\mu_{i_1}-\mu_{i_2}\big)^+-\big(\lambda_{i_1}-\lambda_{i_2}\big)^+-\big(\mu_{i_2}-\mu_{i_1}\big)^+ +\big(\lambda_{i_2}-\lambda_{i_1}\big)^+=\mu_{i_1}-\mu_{i_2}+\lambda_{i_2}-\lambda_{i_1}.$$

It follows from (9) that $\mathcal{K}_1 \le K_6(i_2-i_1)=K_6|i_2-i_1|$.

If $i_1 > i_2$, then $\mathcal{K}_1=0$.

If $i_1=i_2$, then

$$\begin{aligned}\mathcal{K}_1&=-\big(\mu_{i_1}-\mu_{i_2}\big)^+ +\big(\lambda_{i_1}-\lambda_{i_2}\big)^+ +\big(\mu_{i_2}-\mu_{i_1}\big)^+ -\big(\lambda_{i_2}-\lambda_{i_1}\big)^+\\ &=\lambda_{i_1}-\lambda_{i_2}+\mu_{i_2}-\mu_{i_1}\\ &\le K_6(i_1-i_2)=K_6|i_2-i_1|.\end{aligned}$$

Thus $\mathcal{K}_1 \le K_6|i_2-i_1|$.

On the other hand, from (6) we have

$$\begin{aligned}\mathcal{K}_2&\le \big(h(i_2,\|u_t^2\|)-h\big(i_1,\|u_t^1\|\big)\big)^+ +\big(h\big(i_1,\|u_t^1\|\big)-h\big(i_2,\|u_t^2\|\big)\big)^+\\ &\le K_3((i_2-i_1)^+ +\big(\|u_t^1\|-\|u_t^2\|\big)^+ +(i_1-i_2)^+ +\big(\|u_t^2\|-\|u_t^1\|\big)^+)\\ &=K_3\big(|i_2-i_1|+|\|u_t^1\|-\|u_t^2\||\big).\end{aligned}$$

Therefore,

$$\sum_{(j_1,j_2)\in E\times E} q_{(i_1,i_2)(j_1,j_2)}(t)|j_2-j_1| \le (K_6+K_3)|i_2-i_1|+K_3|\|u_t^1\|-\|u_t^2\||.$$

If $u, v \in \mathcal{P}(E)$ and $F(dx,dy)$ is a probability measure on $E\times E$ with marginals u and v, then

$$\begin{aligned}|\|u\|-\|v\||&=\left|\int\int_{E\times E} xF(dx,dy)-\int\int_{E\times E} yF(dx,dy)\right|\\ &\le \int\int_{E\times E}|x-y|F(dx,dy)\end{aligned}$$

and therefore by definition (17) we have

$$|\|u\|-\|v\|| \le \rho_1(u,v).$$

This completes the proof of Lemma 8.

We now return to the proof of Lemma 7. For $F(t)$ defined by (24), along the lines of (3.12) and (3.13) in [10] and Lemma 8, we get

$$\begin{aligned}\frac{\mathrm{d}F}{\mathrm{d}t} &= \sum_{(i_1,i_2)\in E\times E} F((i_1,i_2)) \sum_{(l_1,l_2)\in E\times E} p(t,(i_1,i_2),(l_1,l_2)) \sum_{(j_1,j_2)\in E\times E} q_{(l_1,l_2)(j_1,j_2)}(t)|j_2-j_1| \\ &\le \sum_{(i_1,i_2)\in E\times E} F((i_1,i_2)) \sum_{(l_1,l_2)\in E\times E} p(t,(i_1,i_2),(l_1,l_2))((K_6+K_3)|l_2-l_1| \\ &\quad + K_3\rho_1\big(u_t^1,u_t^2\big)\big) \\ &= (K_6+K_3)F(t) + K_3\rho_1\big(u_t^1,u_t^2\big).\end{aligned}$$

Multiplying the two sides by $e^{-(K_6+K_3)t}$, changing the variable from t to s and integrating it over $(0,t)$ we have

$$F(t) \le e^{(K_6+K_3)t}F(0) + K_3\int_0^t e^{(K_6+K_3)(t-s)}\rho_1\big(u_s^1,u_s^2\big)ds.$$

From (25) we get

$$\rho_1\big(u_t^1,u_t^2\big) \le e^{(K_6+K_3)t}F(0) + K_3\int_0^t e^{(K_6+K_3)(t-s)}\rho_1\big(u_s^1,u_s^2\big)ds.$$

Choosing

$$F((i,j)) = \begin{cases} 0, & i\neq j, \\ u(\{i\}), & i=j, \end{cases}$$

thus $F(0)=0$ and

$$e^{-(K_6+K_3)t}\rho_1\big(u_t^1,u_t^2\big) \le K_3\int_0^t e^{-(K_6+K_3)s}\rho_1\big(u_s^1,u_s^2\big)ds$$

Using Gronwall's Lemma we get $u_t^1 = u_t^2$ for all $t\ge 0$. □

Lemma 9. The unique solution of the martingale problem $[u,\Omega_{h,\|u_t\|}]$ is a q-solution of (13).

Proof. Similar to the proof of Theorem 1.2 in [10]. □

4. Law of large numbers

In this section, we will prove that the empirical probability measure U_N defined by (3) converges to the q-solution of the nonlinear master equation of (13). First of all, we need to give a mathematical description of the stochastic processes of the interacting queueing system introduced in Section 2.

Recalling that $X^{(N)}(t) = (X_1^{(N)}(t), \ldots, X_N^{(N)}(t))$ denotes the number of customers in the queueing system at time t, for $x = (x_1, \ldots, x_N) \in E^{\otimes N}$, we define

$$q_{x,y}^{(N)} = \begin{cases} \mu_{x_k}, & y = (x_1, \ldots, x_{k-1}, x_k - 1, x_{k+1}, \ldots, x_N), \\ & x_k \geq 1, k = 1, \ldots, N, \\ \lambda_{x_k} - h\left(x_k, \dfrac{x_1 + \cdots + x_N}{N}\right), & y = (x_1, \ldots, x_{k-1}, x_k + 1, x_{k+1}, \ldots, x_N), \\ & k = 1, \ldots, N, \\ -\displaystyle\sum_{y \neq x} q_{x,y}^{(N)}, & y = x, \\ 0, & \text{otherwise} \end{cases} \tag{27}$$

For $f \in C_b(E^{\otimes N})$, define

$$\begin{aligned} \Omega^{(N)} f(x) &= \sum_{y \neq x} q_{x,y}^{(N)} (f(y) - f(x)) \\ &= \sum_{k=1}^{N} \Bigg(\mu_{x_k} (f(x - e_k) - f(x)) + \lambda_{x_k} (f(x + e_k) - f(x)) \\ &\qquad - h\left(x_k, \frac{1}{N} \sum_{j=1}^{N} x_j\right) (f(x + e_k) - f(x)) \Bigg). \end{aligned} \tag{28}$$

Then $\Omega^{(N)}$ can be regarded as a Q-matrix which is conservative and totally stable, i.e.

$$\sum_{y \in E^{\otimes N}} q_{x,y}^{(N)} = 0, \quad \text{for all} \quad x \in E^{\otimes N}, \quad \left|q_{x,x}^{(N)}\right| < \infty \quad \text{for all} \quad x \in E^{\otimes N}. \tag{29}$$

If for every $x = (x_1, \ldots, x_N) \in E^{\otimes N}$ we define $|x| = \sum_{k=1}^{N} x_k$, then from (5), (8) and (27), for any fixed N and $x = (x_1, \ldots, x_N)$,

$$\begin{aligned} \sum_{y \neq x} q_{x,y}^{(N)} (|y| - |x|) &= \sum_{k=1}^{N} \left(\lambda_{x_k} - \mu_{x_k} - h\left(x_k, \frac{|x|}{N}\right)\right) \\ &\leq \sum_{k=1}^{N} \left(K_4(1) + K_5(1) x_k + K_1 + K_2 \frac{|x|}{N}\right) \\ &= N(K_4(1) + K_1) + (K_5(1) + K_2)|x| \\ &\leq (N(K_4(1) + K_1) \vee (K_5(1) + K_2))(1 + |x|). \end{aligned} \tag{30}$$

Then from Theorem 2.2 and Corollary 2.3 of [11] we have

Lemma 10. Supposing that (5) and (8) hold, then for each $N \geq 1$, the Q-process $X^{(N)} = \{X_1^{(N)}(t, w), \ldots, X_N^{(N)}(t, w), t \geq 0, w \in D_\infty(E^{\otimes N})\}$ for $\Omega^{(N)}$ exists and is

unique. More-over, for any $u \in \mathcal{P}(E)$, the distribution $P^{(N)} \in \mathcal{P}(D_\infty(E^{\otimes N}))$ with initial value $P^{(N)} \circ X_0^{(N)^{-1}} = u^{\otimes N}$ is the unique solution to the martingale problem associated with $\Omega^{(N)}$.

The main result of this section is stated in Theorem 12, the proof of which is similar to that in [7]. We first prove a lemma.

Lemma 11. If for $k = 1, 2, \ldots, m$ and $M > 0$, we define $f_k^{(M)}(x) = \frac{1}{N}\sum_{i=1}^N x_i^k \wedge M$ and $f_k(x) = \lim_{M\to\infty} f_k^{(M)}(x)$, then

$$\begin{aligned}
\Omega^{(N)} f_1^{(M)}(x) &\le (K_4(1) + K_1) + (K_5(1) + K_2) f_1(x), \\
\Omega^{(N)} f_k^{(M)}(x) &\le K_4(k) + 2^k(K_1 + K_2 f_1(x)) + (K_5(k) + k2^k(K_1 + K_2)) f_k(x), \qquad (31) \\
&\qquad\qquad k = 2, \ldots, m.
\end{aligned}$$

Proof. If $k = 1$, then from (5), (7), (8) and (9) we know that

$$\begin{aligned}
\Omega^{(N)} f_1^{(M)}(x) = \sum_{i=1}^N \Bigg(& \mu_{x_i}\big(f_1^{(M)}(x - e_i) - f_1^{(M)}(x)\big) + \lambda_{x_i}\big(f_1^{(M)}(x + e_i) - f_1^{(M)}(x)\big) \\
& - h\left(x_i, \frac{|x|}{N}\right)\big(f_1^{(M)}(x + e_i) - f_1^{(M)}(x)\big)\Bigg) \\
\le \frac{1}{N}\sum_{i=1}^N & \left(K_4(1) + K_5(1)x_i - h\left(x_i, \frac{|x|}{N}\right)((x_i + 1) \wedge M - x_i \wedge M)\right) \\
\le \frac{1}{N}\sum_{i=1}^N & \left(K_4(1) + K_5(1)x_i + \left(K_1 + K_2\frac{|x|}{N}\right)((x_i + 1) \wedge M - x_i \wedge M)\right) \\
\le & (K_4(1) + K_1) + (K_5(1) + K_2) f_1(x).
\end{aligned}$$

□

For $k = 2, \ldots, m$, we note that $(x+1)^k - x^k \le k2^k x^{k-1} + 1$, and for positive x and y we have $xy^{k-1} + x^{k-1}y \le x^k + y^k$. This implies that $(x_1 + \cdots + x_N)(x_1^{k-1} + \cdots + x_N^{k-1}) \le N(x_1^k + \cdots + x_N^k)$, i.e $f_1(x) \cdot f_{k-1}(x) \le f_k(x)$. Therefore,

$$\begin{aligned}
\Omega^{(N)} f_k^{(M)}(x) &\le \frac{1}{N}\sum_{i=1}^N \left(K_4(k) + K_5(k)x_i^k - h\left(x_i, \frac{|x|}{N}\right)\big((x_i + 1)^k \wedge M - x_i^k \wedge M\big)\right) \\
&\le K_4(k) + K_5(k) f_k(x) + (K_1 + K_2 f_1(x))(k2^k f_{k-1}(x) + 1) \\
&\le K_4(k) + 2^k(K_1 + K_2 f_1(x)) + (K_5(k) + k2^k(K_1 + K_2)) f_k(x).
\end{aligned}$$

Theorem 12. Let $U_N(t)$ be defined by (3) which satisfies

$$\sup_N E^{(N)}\langle U_N(0)(\mathrm{d}x), x^m\rangle < \infty, \tag{32}$$

$$U_N(0) \overset{weakly}{\longrightarrow} U(0) \quad and \quad \langle U(0)(\mathrm{d}x), x^2\rangle < \infty. \tag{33}$$

Then the sequence $\{U_N\}_{N=1}^{\infty}$ converges in the sense of weak convergence of measure-valued stochastic processes to the unique q-solution of the nonlinear master equation of (13).

Proof. If we define $\mathcal{P}_N(E) = \{\{\frac{n_0}{N}, \frac{n_1}{N}, \ldots\}, n_i \in E, \sum_{i\in E} n_i = N\}$, then from the Markov property and the definition in (3) we know that $U_N(t)$ is a measure-valued Markovian process in $D_\infty(\mathcal{P}_N(E))$ and its generator is

$$\begin{aligned} G_N F(\mu) \ = & \sum_{i\in E} n_i \Bigg\{ \mu_i \left(f\left(\langle \mu, \phi\rangle + \frac{\phi(i-1)-\phi(i)}{N} \right) - f(\langle \mu, \phi\rangle) \right) \\ & + \lambda_i \left(f\left(\langle \mu, \phi\rangle + \frac{\phi(i+1)-\phi(i)}{N} \right) - f(\langle \mu, \phi\rangle) \right) \Bigg\} \\ & - \sum_{i\in E} n_i h\left(i, \sum_{j\in E} \frac{jn_j}{N} \right) \left(f\left(\langle \mu, \phi\rangle + \frac{\phi(i+1)-\phi(i)}{N} \right) - f(\langle \mu, \phi\rangle) \right), \end{aligned} \tag{34}$$

where $\mu \in \mathcal{P}_N(E), \phi \in C_b(E), f \in C_b^2(\mathbf{R}) = \{f : \mathbf{R} \to \mathbf{R}, f'' \text{ is bounded}\}$, and $F(\mu) = f(\langle \mu, \phi\rangle)$.

Under the estimation in Lemma 11, together with (5), (7), (8), (9), (32) and (33), a similar argument to that in [7] implies that

$$\sup_N \sup_{t\le T} \mathbf{E}^{\mathbf{P}^{(N)}}(\langle U_N(t), |x|^m\rangle) < \infty. \tag{35}$$

Moreover $\{U_N : N \ge 1\}$ is weakly compact in $D_\infty(\mathcal{P}(E))$. Therefore, for each $\{N'\} \subset \{N\}$, there exists $\{N''\} \subset \{N'\}$ and a measure-valued process $\{U(t), t \ge 0\}$ such that $U_{N''}$ converges weakly to U as $N'' \to \infty$. (34) and (35) imply that the distribution P^∞ of U is a solution to the martingale problem with respect to the following operator G,

$$GF(\mu) = f'(\langle \mu, \phi\rangle)\big\langle \mu, \Omega_{h,\|\mu\|}\phi \big\rangle. \tag{36}$$

Taking two special cases, $f(x) = x$ and $f(x) = x^2$, some calculations imply that $U(t)$ satisfies

$$\langle U(t), \phi\rangle - \langle U(0), \phi\rangle = \int_0^t \big\langle U(s), \Omega_{h,\|U(s)\|}\phi \big\rangle ds, \quad P^\infty - \text{a.s.} \tag{37}$$

This means that $U(t)$ is a solution of the nonlinear master equation of (13). Theorem 1 implies that $U(t)$ is the unique solution of (13). This completes the proof. □

5. Stationary distribution

In this section, we investigate the stationary distribution of the solution of the master equation.

For $y \in [0, \infty)$, we define

$$Q_y = \begin{pmatrix} -\lambda_0 + h(0, y) & \lambda_0 - h(0, y) & 0 & \cdots \\ \mu_1 & -(\lambda_1 + \mu_1) + h(1, y) & \lambda_1 - h(1, y) & \cdots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) + h(2, y) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \quad (38)$$

Definition 3. $\pi \in \mathcal{P}_p(E)$ is called a stationary distribution of the q-solution of (13) if $P \circ X_0^{-1} = \pi$ implies that for all $t \geq 0$, $P \circ X_t^{-1} = \pi$.

Let Π_p denote the set of all stationary distributions of $\pi \in \mathcal{P}_p(E)$ and $\Pi = \cup_{p \geq 1} \Pi_p$.

Lemma 13.

(1) For each $\pi \in \Pi$, if we let $\pi_j = \pi(\{j\})$, then

$$(\pi_0, \pi_1, \ldots) Q_{\|\pi\|} = 0. \quad (39)$$

(2) If $\pi \in \mathcal{P}_p(E)$ satisfies (39) then $\pi \in \Pi_p$.

Theorem 14. Assume

$$L_y \stackrel{\triangle}{=} 1 + \sum_{k=1}^{\infty} \frac{(\lambda_0 - h(0, y)) \cdots (\lambda_{k-1} - h(k-1, y))}{\mu_1 \cdots \mu_k}. \quad (40)$$

If $L_y < \infty$ for each y, then

(1) If $|\Pi|$, the cardinal number of Π, is greater than null, then the equation

$$f(y) \stackrel{\triangle}{=} y + \sum_{k=1}^{\infty} (y - k) \frac{(\lambda_0 - h(0, y)) \cdots (\lambda_{k-1} - h(k-1, y))}{\mu_1 \cdots \mu_k} = 0, \quad (41)$$

has at least one nonnegative solution.

(2) Assuming $0 \le y_0 < \infty$ is a nonnegative solution of (41), we take

$$\pi_{y_0} = \frac{1}{L_{y_0}} \left(1, \ldots, \frac{(\lambda_0 - h(0, y_0)) \cdots (\lambda_{k-1} - h(k-1, y_0))}{\mu_1 \cdots \mu_k}, \ldots \right). \tag{42}$$

If $\pi_{y_0} \in \mathcal{P}_p(E)$, then $\pi_{y_0} \in \Pi_p$.

(3) $|\Pi|$ is equal to the number of nonnegative finite solutions of (41).

The proofs to the above two conclusions are similar to that of [10].

In the following, we provide some examples to illustrate how the system performance can be improved by introducing mean-field interaction. Consider a queuing network consisting of N *M/M/*1 queues as N goes to infinity. If all queues run independently without mean-field interaction with the generator defined by (1), then the stationary distribution of each queue or a "typical queue" simply corresponds to the distribution corresponding to generator (1). For the network with mean-field interaction, the stationary distribution of each queue or a "typical queue" corresponds to that of generator (38) where y is the mean queue length. We will make a comparison between the "typical queue" with and without interaction.

It is easy to have a light-tailed stationary distribution for a birth-death process, for example, when all $\lambda_j = \lambda$ for $j \ge 0$ and $\mu_j = \mu$ for $j \ge 1$. The following example illustrates the possible distribution of heavy-tailed behavior for a birth-death process.

Example 1. Let $\lambda_j = 1 \vee j$ for $j \ge 0$ and $\mu_j = j + 3$ for $j \ge 1$. It follows from the equation which the stationary distribution satisfies that $\pi_k = \frac{6}{k(k+1)(k+2)(k+3)} \pi_0$, $k \ge 1$. It is easy to see that

$$\pi_0 = \frac{3}{4}, \quad \pi_k = \frac{9}{2k(k+1)(k+2)(k+3)}, \quad k \ge 1.$$

Since for any $\varepsilon > 0$, there exists an integer $c_3(\varepsilon)$ such that for $k \ge c_3(\varepsilon)$ we have $\pi_k e^{\varepsilon k} = \frac{9e^{\varepsilon k}}{2k(k+1)(k+2)(k+3)} > 1$, which implies that

$$\sum_{k=0}^{\infty} \pi_k e^{\varepsilon k} = \infty.$$

Therefore, the stationary distributions $\{\pi_k, k \ge 0\}$ is heavy-tailed.

In fact, for the birth-death processes generated by (1), if we define

$$\theta_0 = 1, \quad \theta_j = \frac{\lambda_0 \cdots \lambda_{j-1}}{\mu_1 \cdots \mu_j}, \quad j \ge 1, \tag{43}$$

then we have

Theorem 15. Let $\sum_{j=0}^{\infty} \theta_j < \infty$ and $\sum_{j=0}^{\infty} \frac{1}{\lambda_j \theta_j} = \infty$. Then:

(a) The stationary distribution is light-tailed iff $\limsup_{n\to\infty} \sqrt[n]{\theta_n} < 1$;

(b) The stationary distribution is heavy-tailed iff $\limsup_{n\to\infty} \sqrt[n]{\theta_n} = 1$.

Proof. From the assumption we know that the birth-death process is positive recurrent and there exists an n_0 such that if $n > n_0$, then $\theta_n < 1$, which implies that $\limsup_{n\to\infty} \sqrt[n]{\theta_n} \leq 1$. We define

$$g(z) = \sum_{n=0}^{\infty} \theta_n z^n.$$

Then the radius of convergence is $\gamma = (\limsup_{n\to\infty} \sqrt[n]{\theta_n})^{-1}$.

(a) If $\limsup_{n\to\infty} \sqrt[n]{\theta_n} < 1$, then $\gamma > 1$. For a small $\varepsilon > 0$ such that $e^{\varepsilon} = \gamma_1 < \gamma$ we have

$$\sum_{n=0}^{\infty} \theta_n e^{\varepsilon n} = \sum_{n=0}^{\infty} \theta_n \gamma_1^n < \infty.$$

This implies that the stationary distribution is light-tailed.

If the stationary distribution is light-tailed, then there exists an $\varepsilon_0 > 0$ such that for $\varepsilon < \varepsilon_0$, we have $\sum_{n=0}^{\infty} \theta_n e^{\varepsilon n} < \infty$. Then the radius of convergence is $\gamma \geq e^{\varepsilon_0} > 1$, therefore $\limsup_{n\to\infty} \sqrt[n]{\theta_n} = \frac{1}{\gamma} < 1$.

The same method is applied to the heavy-tailed case (b). □

In the next example, we show how the balancing model considered in this paper can improve performance.

Example 2. Consider a network consisting of N queues, where N is large. If all the queues are identical to that given in Example 1 and run independently, then the stationary distribution of each queue is heavy-tailed. The mean arrival rate to each queue is $\sum_{k=0}^{\infty} \lambda_k \pi_k = \frac{9}{8} = 1.125$. If we introduce the mean-field interaction considered in this paper, and for two numbers $0 < \beta < 1$ and $M \geq 0$, let the interaction function be $h(x, y) = \beta(1 \vee x) - y \wedge M$. Then the Q-matrix of a "typical queue" is given by $\tilde{\lambda}_j = (1 - \beta)(1 \vee j) + y \wedge M$, $j \geq 0$ and $\tilde{\mu}_j = j + 3$, $j \geq 1$. In this case, if we let $\beta = 0.2$ and $M = 10$, then the equation in (41) has three solutions, which are:

Solution	π_0	Mean arrival rate
$y_1 = 1.046478$	0.6916	1.3899
$y_2 = 2.23825$	0.2972	4.4467
$y_3 = 35.00039$	0.00002	38.002

For $y_1, \pi_0 = 1/1.4459 = 0.6916$ and the mean arrival rate is 1.3899 which is bigger than that for the system without interaction. Moreover, the stationary distribution for the system with interaction is light-tailed. In fact, For $x_1 = 1 - \beta = 0.8 < 1$,

$$\pi_0 = 0.6916, \pi_1 = \frac{x_1 + y_1 \wedge 10}{4}\pi_0,$$
$$\pi_k = \frac{(x_1 + y_1 \wedge 10)(x_1 + y_1 \wedge 10)(2x_1 + y_1 \wedge 10)\cdots((k-1)x_1 + y_1 \wedge 10)}{4\cdots(k+3)}\pi_0,$$
$$k \geq 2.$$

There must exist an integer $c_4 > 5$ such that $\frac{y_1 \wedge 10}{C_4} + x_1 \leq x_0 < 1$, and for $k > c_4$,

$$\begin{aligned}
\pi_k &= \pi_0 \cdot \frac{(x_1 + y_1 \wedge 10)(x_1 + y_1 \wedge 10)\cdots((c_4-1)x_1 + y_1 \wedge 10)}{4\cdots(c_4-1)} \\
&\quad \cdot \frac{(c_4 x_1 + y_1 \wedge 10)\cdots((k-1)x_1 + y_1 \wedge 10)}{c_4 \cdots (k+3)} \\
&= \pi_0 \cdot const \cdot \frac{(c_4 x_1 + y_1 \wedge 10)\cdots((k-1)x_1 + y_1 \wedge 10)}{c_4 \cdots (k+3)} \\
&\leq \pi_0 \cdot const \cdot \frac{c_4 x_0 \cdot (c_4+1)x_0 \cdots (k-1)x_0}{c_4 \cdots (k+3)} \\
&= \pi_0 \cdot const \cdot \frac{x_0^{k-c_4}}{k(k+1)(k+2)(k+3)}.
\end{aligned}$$

If ε satisfies $x_0 e^{\varepsilon} < 1$, then

$$\sum_{k=c_4}^{\infty} \pi_k e^{\varepsilon * k} \leq \sum_{k=c_4}^{\infty} \pi_0 \cdot const \cdot \frac{(x_0 e^{\varepsilon})^k}{x_0^{c_4} k(k+1)(k+2)(k+3)} < \infty.$$

From above we know that the stationary distribution is light-tailed. The same conclusion can be made for the other two solutions.

For the general case, we define

$$\tilde{\theta}_0 = 1, \quad \tilde{\theta}_j = \frac{(\lambda_0 - h(0, y)) \cdots (\lambda_{j-1} - h(j-1, y))}{\mu_1 \cdots \mu_j}, \quad j \geq 1. \tag{44}$$

We have the following conclusion.

Theorem 16. Let λ_k and μ_k satisfy (7), (8) and (9), $\sum_{j=0}^{\infty} \theta_j < \infty$, $\lambda_0 > 1$, and λ_k increase in k. Assume that $0 < \beta < 1$, and $0 \leq M < \infty$ are two parameters. Define the interaction function as

(A) If $\sup \lambda_k < \infty$, define $h(k, y) = \beta\lambda_k - \frac{y \wedge M}{k \vee 1}$;

(B) If $\sup \lambda_k = \infty$, $(\lambda_k - \lambda_l)^+ \le K_7(k-l)^+$ and $\sum_{j=0}^{\infty} \frac{1}{\lambda_j \theta_j} = \infty$, define $h(k, y) = \beta\lambda_k - y \wedge M$, where K_7 is a constant.

Then for either of the interaction functions defined in (A) and (B), we have

(1) For each (β, M), the stationary distribution of the interaction queue may not be unique, but all $\{\tilde{\theta}_j\}$ will be light-tailed.
(2) If the average arrival rate and the average queue length of the non-interaction queue are finite, then there exists a domain L (or G) in $(0, 1) \times [0, \infty)$ such that for each $(\beta_0, M_0) \in L$ (or G), all the average arrival rates of the interaction queue are not greater than (or not less than) that of the non-interaction one.
(3) We can also choose a (β_1, M_1) such that the stationary distribution of the interaction queue is unique and the average arrival rate of the interaction queue is less than (equal to or greater than) that of the non-interaction queue.

Proof. For $h(k, y)$ defined in (A). Assumption $\sum_{j=0}^{\infty} \theta_j < \infty$ implies that there exists an integer $c_5 > 0$ such that if $j > c_5$ then $\theta_j < 1$. If λ_k increases in k and $\sup \lambda_k$ is bounded from above, then we can assume that $\lambda_k < c_6 < \infty$ for $k \ge 0$, which implies that

$$\begin{aligned}\sum_{j=0}^{\infty} \frac{1}{\lambda_j \theta_j} &\ge \sum_{j=0}^{c_5} \frac{1}{\lambda_j \theta_j} + \sum_{j=c_5+1}^{\infty} \frac{1}{\lambda_j} \\ &\ge \sum_{j=0}^{c_5} \frac{1}{\lambda_j \theta_j} + \sum_{j=c_5+1}^{\infty} \frac{1}{c_6} \\ &= \infty\end{aligned}$$

This together with $\sum_{j=0}^{\infty} \theta_j < \infty$ implies that the non-interaction birth-death process is positive recurrent. Since in this case $h(k, y) = \beta\lambda_k - \frac{y \wedge M}{k \vee 1}$, it is easy to show that $h(k, y)$ defined here satisfies (5) and (6). Let β and M be fixed. There exists an integer $c_7 > 1$ such that $0 < 1 - \beta + \frac{M}{k\lambda_k} = c_8 < 1$ for $k \ge c_7$, which implies that

$$\tilde{\lambda}_k = \lambda_k - \beta\lambda_k + \frac{y \wedge M}{k} = \lambda_k \left(1 - \beta + \frac{y \wedge M}{k\lambda_k}\right) \le c_8 \lambda_k \quad \text{for} \quad k \ge c_7, y \ge 0. \tag{45}$$

This indicates that L_y defined by (40) is finite for any $y \ge 0$, and $\lim_{y\to\infty} f(y) = \infty$. Moreover, from (45) and the assumption of the parameters of the non-interaction birth death process, we know that the finite solution of (41) does exist, and therefore the stationary distribution of the interaction system exists. Letting y_0 be any solution of (41), let us prove that $\{\tilde{\theta}_j\}$ is light-tailed. In fact, for this y_0, from above we

know that

$$\begin{aligned}\tilde{\theta}_k &= \frac{(\lambda_0 - h(0, y_0)) \cdots (\lambda_{c_7-1} - h(c_7 - 1, y_0))}{\mu_1 \cdots \mu_{c_7}} \\ &\quad \cdot \frac{\lambda_{c_7}\left(1 - \beta + \frac{y_0 \wedge M}{c_7 \lambda_{c_7}}\right) \cdots \lambda_{k-1}\left(1 - \beta + \frac{y_0 \wedge M}{(k-1)\lambda_{k-1}}\right)}{\mu_{c_7+1} \cdots \mu_k} \\ &= \text{const} \cdot \frac{\lambda_{c_7}\left(1 - \beta + \frac{y_0 \wedge M}{c_7 \lambda_{c_7}}\right) \cdots \lambda_{k-1}\left(1 - \beta + \frac{y_0 \wedge M}{(k-1)\lambda_{k-1}}\right)}{\mu_{c_7+1} \cdots \mu_k} \\ &\le \text{const.} \frac{\lambda_{c_7} \cdots \lambda_{k-1}}{\mu_{c_7+1} \cdots \mu_k} (c_8)^{k-c_7}.\end{aligned}$$

If ε satisfies $c_8 e^{\varepsilon} < 1$, then $\sum_{k=c_7}^{\infty} \tilde{\theta}_k e^{\varepsilon k} \le \sum_{k=c_7}^{\infty} \frac{\text{const}}{c_8^{c_7}} \cdot \frac{\lambda_{c_7} \cdots \lambda_{k-1}}{\mu_{c_7+1} \cdots \mu_k} (c_8 e^{\varepsilon})^k < \infty$, which means that $\{\tilde{\theta}_j\}$ is light-tailed. We get (1).

Let us prove (2). Note that for any nonnegative y, the average queue length of the stationary distribution corresponding to the generator (38) is

$$f_1(\beta, M, y) = \frac{\sum_{k=1}^{\infty} k \frac{\lambda_0 \cdots \lambda_{k-1}}{\mu_1 \cdots \mu_k} \left(1 - \beta + \frac{y \wedge M}{\lambda_0}\right)\left(1 - \beta + \frac{y \wedge M}{\lambda_1}\right) \cdots \left(1 - \beta + \frac{y \wedge M}{(k-1)\lambda_{k-1}}\right)}{1 + \sum_{k=1}^{\infty} k \frac{\lambda_0 \cdots \lambda_{k-1}}{\mu_1 \cdots \mu_k} \left(1 - \beta + \frac{y \wedge M}{\lambda_0}\right)\left(1 - \beta + \frac{y \wedge M}{\lambda_1}\right) \cdots \left(1 - \beta + \frac{y \wedge M}{(k-1)\lambda_{k-1}}\right)}. \tag{46}$$

Apparently, $0 < f_1(0, 0, 0) < \infty$ is the average queue length of the non-interaction queue. It is easy to see that the function $f_1(\beta, M, y)$ increases in M and y and decreases in β. Moreover, for each (β, M), equation (41) is equivalent to the following equation:

$$f_1(\beta, M, y) = y. \tag{47}$$

The average arrival rate to the interaction queue is given by

$$f_2(\beta, M) = \frac{\lambda_0 \left(1 - \beta + \frac{y_0 \wedge M}{\lambda_0}\right) + \lambda_0 \left(1 - \beta + \frac{y_0 \wedge M}{\lambda_0}\right) \sum_{k=1}^{\infty} \frac{\lambda_1 \cdots \lambda_k}{\mu_1 \cdots \mu_k} \left(1 - \beta + \frac{y_0 \wedge M}{\lambda_1}\right) \cdots \left(1 - \beta + \frac{y_0 \wedge M}{k\lambda_k}\right)}{1 + \sum_{k=1}^{\infty} \frac{\lambda_0 \cdots \lambda_{k-1}}{\mu_1 \cdots \mu_k} \left(1 - \beta + \frac{y_0 \wedge M}{\lambda_0}\right) \cdots \left(1 - \beta + \frac{y_0 \wedge M}{(k-1)\lambda_{k-1}}\right)},$$

where y_0 is a solution of (41). $f_2(0, 0)$ is the average arrival rate of the non-interaction queue which satisfies $0 < f_2(0, 0) < \infty$.

Since $\lambda_0 > 1$, $f_2(\beta, M)$ increases in M and decreases in β if $0 < \beta < \beta_2 = 1 - \frac{1}{\lambda_0}$. On one hand, for any $0 < \varepsilon_0 < f_1(0, 0, 0)$, from the monotonicity of $f_1(\beta, 0, 0)$ we know that there exists a $0 < \beta_3(\varepsilon_0) \le 1$ such that for $0 < \beta < \beta_3(\varepsilon_0)$ and any nonnegative M, any solution of (47) will be not less than ε_0. On the other hand, noting that $f_2(\beta, \varepsilon_0) >$

$f_2(\beta, 0)$ and $\lim_{\beta\to 0} f_2(\beta, 0) = f_2(0, 0)$ (also from the monotone of $f_2(\beta, \varepsilon_0)$), we know that there exists a $0 < \beta_4(\varepsilon_0) \le \beta_2$ such that if $0 < \beta < \beta_4(\varepsilon_0)$ then $f_2(\beta, \varepsilon_0) \ge f_2(0, 0)$. Therefore, if we take $(\beta_0, M_0) \in G = \cup_{0<\varepsilon_0<f_1(0,0,0)}(0, \beta_3(\varepsilon_0) \wedge \beta_4(\varepsilon_0)) \times [\varepsilon_0, \infty)$ then all the average arrival rates of the interaction queue will not be less than that of the non-interaction queue.

In order to prove the case of "not greater than", define

$$f_3(\beta, M) = \frac{\lambda_0\big(1-\beta+\frac{M}{\lambda_0}\big) + \lambda_0\big(1-\beta+\frac{M}{\lambda_0}\big)\sum_{k=1}^{\infty}\frac{\lambda_1\cdots\lambda_k}{\mu_1\cdots\mu_k}\big(1-\beta+\frac{M}{\lambda_1}\big)\cdots\big(1-\beta+\frac{M}{k\lambda_k}\big)}{1+\sum_{k=1}^{\infty}\frac{\lambda_0\cdots\lambda_{k-1}}{\mu_1\cdots\mu_k}\big(1-\beta+\frac{M}{\lambda_0}\big)\cdots\big(1-\beta+\frac{M}{(k-1)\lambda_{k-1}}\big)}.$$

It is easy of see that if $0 < \beta \le \beta_2$ then $f_2(\beta, M) \le f_3(\beta, M)$ for any $M \ge 0$, $f_3(\beta, M)$ increases in M and decreases in β. since for each $\beta_0 \in (0, \beta_2)$, $f_3(\beta_0, M)$ is a continuous function of M and $\lim_{M\to 0} f_3(\beta_0, M) = f_3(\beta_0, 0) = f_2(\beta_0, 0) < f_2(0, 0)$, there exists an $M_2(\beta_0)$ such that for each $0 \le M_0 \le M_2(\beta_0)$, $f_3(\beta_0, M_0) \le f_2(0, 0)$. Therefore, for each $(\beta_0, M_0) \in L = \{(\beta, M), 0 < \beta \le \beta_2, 0 \le M \le M_2(\beta)\}$, all the average arrival rates are not greater than that of the non-interaction queue.

For (3), from above we know that for each $\beta_1 \in (0, \sup_{0<\varepsilon_0<f_1(0,0,0)} \beta_3(\varepsilon_0) \wedge \beta_4(\varepsilon_0))$, $\lim_{M\to 0} f_3(\beta_1, M) = f_2(\beta_1, 0) < f_2(0, 0)$, and there exists an $M_3 > 0$ such that $f_3(\beta_1, M_3) = f_2(0, 0)$. In this case the solution of (47) is unique, which is $y = f_1(\beta_1, M_3, M_3) \ge M_3$. Now we have that the average arrival rate of the interaction queue is less than (equal to or greater than) that of non-interaction queue for (β_1, M_1) if we choose $M_1 \in [0, M_3)(M_1 = M_3$, or M_1 is greater than but very close to $M_3)$.

A similar argument can be applied to prove the other interaction function defined in (B). This completes the proof. □

Remark 6. The role of the parameter M in the interaction function of Theorem 16 is not only to balance the queue, but also to guarantee the stability of the interaction system.

Remark 7. Theorem 16 provides properties of the stationary distribution of one special kind of the mean-field interaction system discussed in this paper. In fact, for different birth and death rates satisfying (7), (8) and (9), different interaction functions satisfying (5) and (6) can be chosen, and similar results to Theorem 16 can also be established.

Theorem 17. Let $\{\theta_j\}$ and $\{\tilde{\theta}_j\}$ be defined as in (43) and (44) respectively, if $\{\theta_j\}$ is light-tailed, then $\{\tilde{\theta}_j\}$ cannot be heavy-tailed.

Proof. If $\{\theta_j\}$ is light-tailed, then $\lim_{n\to\infty} \sup \sqrt[n]{\theta_n} < 1$. From the assumption of the interaction function, we know that for any $y \ge 0$, there exists a $k_0 \in \mathbf{N}$ such that $h(k_0, y) \ge 0$. Therefore,

$$\tilde{\theta}_n = \frac{\tilde{\lambda}_0\cdots\tilde{\lambda}_{k_0}\tilde{\lambda}_{k_0+1}\cdots\tilde{\lambda}_{n-1}}{\mu_1\cdots\mu_n} \le \frac{\tilde{\lambda}_0\cdots\tilde{\lambda}_{k_0}\lambda_{k_0+1}\cdots\lambda_{n-1}}{\mu_1\cdots\mu_n},$$

which implies that

$$\lim_{n\to\infty}\sup\sqrt[n]{\tilde{\theta}_n} \le \lim_{n\to\infty}\sup\sqrt[n]{\frac{\tilde{\lambda}_0\cdots\tilde{\lambda}_{k_0}\lambda_{k_0+1}\cdots\lambda_{n-1}}{\mu_1\cdots\mu_n}} = \lim_{n\to\infty}\sup\sqrt[n]{\theta_n} < 1.$$

From Theorem 15 we know that $\{\tilde{\theta}_j\}$ is light-tailed. □

In addition to the improvement of the tail properties described in the last two theorems, the following example demonstrates that "better" throughputs and waiting times can be obtained by introducing the mean-field interaction.

Example 3. A linear birth death process with birth rate $\lambda_k = a + b * k$ and death rate $\mu_k = c + d * k$ has many industrial applications (see for example [20]). Instead of using the interaction function defined in Theorem 16, we define a new one by

$$h(x, y) = \begin{cases} \beta b(x-y), & x \ge y, \\ (x-y) - \dfrac{M-1}{2}(x-y)^2, & y-1 < x < y, \\ -\dfrac{M+1}{2} - M(1-e^{x-y+1}), & x \le y-1. \end{cases}$$

Note that when $h(x, x) = 0$, the arrival rate to the queue will not be changed. In the following numerical example, let $a = 1.1$, $b = 1$, $c = 3.2$ and $d = 1$. We have that:

$a = 1.1, b = 1, c = 3.2, d = 1$		Mean queue length	Mean arrival rate	Mean waiting time
Non-interaction queue		1. 0000	2.1000	0.4762
Mean field	$\beta = 0.8000$	1. 0000	3.0065	0.3326
interaction queue	$M = 5.4609$		(*Raised*: 43.17%)	(*Lowered*: 30.15%)
	$\beta = 1.0000$	0.6469	2.1000	0.3080
	$M = 4.0922$	(*Lowered*: 35.31%)		(*Lowered*: 35.31%)
	$\beta = 0.1000$	2.2713	4.7696	0.4762
	$M = 2.6116$	(*Raised*: 127.13%)	(*Raised*: 127.13%)	

Here the mean waiting time is calculated from the mean queue length and the mean arrival rate by Little's Formula. This table provides a measure of the improvement of the network after introducing the mean-field interaction.

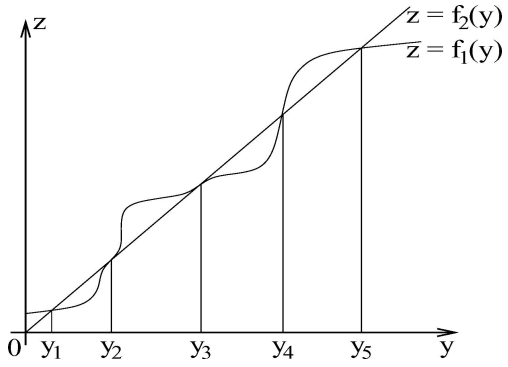
Let us give a remark to end this paper.

Remark 8. Generally, the stationary distribution of a birth-death process is unique. However, the situation is different in the case of the interaction system. From Theorem 14 we know that the number of stationary distributions is determined by the number of solutions of equation (41). Equation (41) could have several different solutions. Each

solution is a mean queue length, to which there corresponds a stationary distribution. In order to find solutions, let us revisit the procedure used to establish equation (41). For any nonnegative y, the mean queue length of the stationary distribution corresponding to the generator (38) is

$$f_1(y) = \frac{\sum_{k=1}^{\infty} k \frac{(\lambda_0 - h(0,y)) \cdots (\lambda_{k-1} - h(k-1,y))}{\mu_1 \cdots \mu_k}}{1 + \sum_{k=1}^{\infty} \frac{(\lambda_0 - h(0,y)) \cdots (\lambda_{k-1} - h(k-1,y))}{\mu_1 \cdots \mu_k}}.$$

If y_0 is the mean queue length of a stationary distribution, we should have that $y_0 = f_1(y_0)$. This implies that y_0 is a solution of $y = f_1(y)$, which can be solved by iteration. Note that this equation is also equivalent to equation (41). Solutions of this equation are also the abscissa of those intersection points of two curves $z = f_1(y)$ and $z = f_2(y) := y$. Sketches of $z = f_1(y)$ and $z = f_2(y)$ are as follows:



In this figure, it is easy to see that the above equation has five solutions: $y_1, \ldots, y_5$. These five solutions are very different. y_1, y_3 and y_5 are stable fixed points, while y_2 and y_4 are unstable fixed points. At the stable fixed points, namely, y_1 and y_5, the fluctuations (as $N \to \infty$) are described by a Gaussian central limit theorem, whereas the fluctuation around y_3 need not be Gaussian.

For a concrete system, if the arrival rate, service rate and the interaction function are fixed, then the stationary distributions are also fixed. The particular stationary distribution arising in the limit as time goes to infinity depends only on the initial value of the distribution at time $t = 0$.

References

[1] C. Adjih, P. Jacquet and N. Vvedenskaya, Performance evalution of a single queue under multi-user TCP/IP connection, INRIA Research report, #4141 (2001).

[2] G.L. Arsenišvili, Single-channel queuing systems with varying intensities, Trudy Tbiliss. Univ. 189, (1977) 65–79.

[3] F. Baccelli, R.D. McDonald, and J. Reynier, A mean field model for multiple TCP connections through a buffer implementing RED. INRIA Research report #4449 (2002).
[4] M.F. Chen, Existence theorems for interacting particle systems with non-compact state space. Sci Sinica. A 30 (1987) 148–156.
[5] D.A. Dawson, Critical dynamics and fluctuations for a mean-field model of cooperative behavior. Journal of Statistical Physics, 31(1) (1983) 29–85.
[6] D.A. Dawson and J. Gärtner, Long-time behavior of interaction diffusions. Stochastic calculus in application, in *Proceedings of the Cambridge Symposium*, eds. J.R. Norris, Pitman Research Notes in Mathematics Series 197, (John Wiley & Sons, New York 1988).
[7] D.A. Dawson and X. Zheng, Law of large numbers and a central limit theorem for unbounded jump mean field models, Adv. Appl. Math. 12(3) (1991) 293–326.
[8] N. Duffield, Local mean-field Markov processes: An application to message-switching networks. Probab. Theory Related Fields 93(4) (1992) 485–505.
[9] W. Feller, On the integro-differential equations of purely discontinuous Markoff processes, Trans. Amer. Math. Soc. 48 (1940) 488–515.
[10] S. Feng and X. Zheng, Solutions of a class of nonlinear master equations, Stochastic Processes and their Applications 43 (1992) 65–84.
[11] S. Feng, Large deviations for unbounded jump type Markov processes with mean field interaction, Technical Report Series of the Laboratory for Research in Statistics and Probability, Carleton University. No. 182, (1991).
[12] S. Feng, Nonlinear master equation of multitype particle systems, Stochastic Processes and their Applications 57 (1995) 247–271.
[13] R.D. Foley and D.R. McDonald, Join the shortest queue: Stability and exact asymptotics, Ann. Appl. Probab. 11(3) (2001) 569–607.
[14] T. Fujisawa, On a queuing process with queue-length dependent service, Yokohama Math. J. 10 (1962) 53–72.
[15] T. Funaki, A certain class of diffusion processes associated with nonlinear parabolic equations, Zeitschrift fur Wahr. 67 (1984) 331–348.
[16] H.C. Gromoll, A.L. Puha and R.J. Williams. The fluid limit of a heavily loaded processor sharing queue. Preprint (2001).
[17] N. Hadidi, A queueing model with variable arrival rates, Period. Math. Hungar. 6 (1975) 39–47.
[18] F.A. Haight, Two queues in parallel, Biometrika 45 (1958) 401–410.
[19] K. Hepp and E.H. Lieb, On the superradiant phase transition for molecules in a quantized radiation field: The Dicke Maser model, Ann. Physics 76 (1973) 360–404.
[20] L.L. Hoffman, Characterizing linear birth and death processes, J. Amer. Statist. Assoc. 87 (420) (1992) 1183–1187.
[21] C. Knessl, B.J. Matkowsky, Z. Schuss and C. Tier, Distribution of the maximum buffer content during a busy period for state-dependent M/G/l queues, Comm. Statist. Stochastic Models 3(2) (1987) 191–226.
[22] A. Mandelbaum and G. Pats, State-dependent queues: Approximations and applications. Stochastic Networks, IMA Vol. Math. Appl. 71, 239–282, (Springer, New York) (1995).
[23] J.B. Martin and Y.M. Suhov, Fast Jackson networks, Annals of Applied Probability 9(3) (1999) 854–870.
[24] S.P. Meyn, Sequencing and routing in multiclass queueing networks, I. Feedback regulation, SIAM J. Control Optim. 40(3) (2001) 741–776 (electronic).
[25] M. Mitzenmather, The power of two choices in randomized load balancing. Ph.D. thesis, University of Carlifornia, Berkeley (1996).
[26] M. Mitzenmacher and B. Voecking, Selecting the shortest of two, improved. Analytic Methods in applied probability, Amer. Math. Soc. Transl. Ser. 2 207, (2002). 165–176, (Amer. Math. Soc., Providence, R1).

[27] J.A. Morrison, Diffusion approximation for head-of-the-line processor sharing for two parallel queues. SIAM J. Appl. Math. 53(2) (1993) 471–490.
[28] B. Natvig, On a queuing model where potential customers are discouraged by queue length, Scand. J. Statist. 2(1) (1975) 34–42.
[29] G. Nicolis and I. Prigogine, *Self-Organization in Nonequilibrium System* (John Wiley & Sons, New York, 1977).
[30] V.I. Oseledets and D.V. Khmelev, Global stability of infinite systems of nonlinear differential equations, and nonhomogeneous countable Markov chains, Problemy Peredachi Informatsii 36(1) (2000) 60–76; translation in Probl. Inf. Transm. 36(1) 54–70.
[31] V.I. Oseledets and D.V. Khmelev, Stochastic Transportation Networks and Stability of Dynamical Systems, Theory of Probability and Its Applications 46(1) (2002) 154–161.
[32] P. Tinnakornsrisuphap and A. M. Makowski, TCP traffic modeling via limit theorems. Technical research Report, http://www.isr.umd.edu/TechReports/ISR/2002/TR_2002-23/TR_2002-23.phtml (2002).
[33] N.D. Vvedenskaya, R.L. Dobrushin and F.I. Karpelevich, Queueing system with selection of the shortest of two queues: An asymptotic approach, Problems of Information Transmission 32(1) (1996) 15–27.
[34] N. D. Vvedenskaya and Yu. M. Suhov, Dobrushin's mean-field approximation for a queue with dynamic routing. Markov Processes and Related Fields 3 (1997) 493–526.
[35] N.D. Vvedenskaya and M.Y. Suhov, Fast Jackson networks with dynamic routing, Problems of Information Transmission 38(2) (2002) 136–153.
[36] W. Winston, Optimality of the shortest line discipline, J. Appl. Probability 14(1) (1977) 181–189.
[37] S. Yan and Z. Li, The stochastic models for non-equilibrium system and formulation of master equations, Acta Phys. Sinica 29 (1980).