Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

MAX-MIN OPTIMALITY OF SERVICE RATES IN QUEUEING SYSTEMS WITH CUSTOMER-AVERAGE PERFORMANCE CRITERION

Li Xia
Ming Xie
Wenjun Yin
Jin Dong

IBM China Research Laboratory
Diamond Building, Zhongguancun Software Park
Haidian District, Beijing 100193, P. R. CHINA

ABSTRACT

In this paper, we consider the optimization of service rates in queueing systems, especially in closed Jackson networks. The optimization criterion is the customer-average performance, which is another important performance metric compared with the traditional time-average performance. Based on the methodology of perturbation analysis, we can derive a performance difference equation when the service rates are changed. With this difference equation, we find the optimal service rates have a Max-Min property, i.e., the optimal service rates can be chosen from its maximal or minimal value. This property can reduce the complexity of this type of optimization problems. Moreover, we also prove the max-min optimality is valid for both state-dependent service rates and load-dependent service rates in queueing systems.

1 INTRODUCTION

Perturbation analysis (PA) is an important theory for the performance optimization of discrete event dynamic systems (DEDS). It gives an efficient way to estimate the performance gradient with respect to system parameters (e.g., the service rates of servers in queueing systems) based on a single sample path (Cao 1994, Cassandras and Lafortune 1999, Glasserman 1991, Gong and Ho 1987, Ho and Cao 1991, Ho, Cao, and Cassandras 1983). With the estimated performance gradients, the gradient-descent algorithms can be used to optimize the system parameters (Chong and Zak 2001). The traditional PA theory of queueing systems is extended to Markov systems in the past decade. With the PA theory of Markov systems (we also call it Markov performance potential theory), many new insights into Markov decision processes (MDP) are obtained from a new point of view (Cao and Chen 1997, Cao 2003). Although the perturbations in queueing systems are continuous in nature and those in Markov systems are discrete, Xia and Cao (2006b) establish the relationship between these two theories. It bridges the gap between PA theory in these two systems. Based on this relationship, we can develop the parallel results between these two theories, such as the performance difference equations and policy iteration algorithms to optimize the service rates under the customer-average performance criterion in queueing systems (Xia, Chen, and Cao 2008).

In most of the optimization algorithms, time-average performance is used as the performance metric of queueing systems. In fact, customer-average performance is also an important performance metric in queueing systems. It quantifies the system performance averaged by the number of served customers and reflects the idea of "customer-oriented", which is very popular in many service industries. In practice, many performance measurements belong to the customer-average performance, such as the average waiting time of each customer in a banking system, the probability of a packet successfully reaching the destination node in a wireless communication network, and so on. On the other hand, to the best of our knowledge, there seems not to exist any algorithm for the customer-average performance. Moreover, by a numerical experiment presented in this paper we find that the algorithm for time-average performance can not be directly used to optimize the customer-average performance. Therefore, it is desirable to study the optimization of service rates under the customer-average performance criterion.

In this paper we consider the optimization of service rates in a closed Jackson network. The objective is to change the service rates to achieve the optimal customer-average performance. We study the optimization of two types of service rates, state-dependent service rates and load-dependent service rates. With the performance difference equation in PA theory, we prove that the service

978-1-4244-2708-6/08/$25.00 ©2008 IEEE

rates have the Max-Min optimality when the performance function satisfies some properties. It means that we only need to choose the maximal or minimal values of service rates during the optimization of this type of problems. It can greatly reduce the complexity of optimization problems. The Max-Min optimality of service rates is similar to the results of Ma and Cao (1994), Yao and Schechner (1989). But in this paper the optimization criterion is customer-average performance, which is different from the time-average performance discussed by Ma and Cao (1994), Yao and Schechner (1989). Moreover, the previous articles (Ma and Cao 1994, Yao and Schechner 1989) only discuss the load-dependent service rates. Here we prove the Max-Min optimality for both state-dependent service rates and load-dependent service rates. Furthermore, in this paper we prove the Max-Min optimality based on the difference equation in PA theory with a very clear and concise manner. The proofs of Ma and Cao (1994), Yao and Schechner (1989) are much more difficult respectively from the performance gradients and linear programming. The performance difference equation and policy iteration used in this paper give another way to do the optimization of queueing systems.

The remainder of the paper is organized as follows. In Section 2 we give an introduction of PA theory in queueing systems related to this paper. Section 3 discusses the optimization problems in closed Jackson networks. We prove the Max-Min optimality of service rates based on the performance difference equation. The policy iteration algorithm is also introduced to do the optimization. In Section 4 we give the numerical experiments to demonstrate the optimality of Max-Min service rates. The difference between customer-average performance and time-average performance is also demonstrated by numerical examples. Finally, we conclude this paper in Section 5.

2 BACKGROUND ON PERTURBATION ANALYSIS

In this section, we give an introduction of PA theory in queueing systems, especially the concept of perturbation realization factors and performance sensitivity equations in queueing systems.

Consider a closed Jackson network consisting of M servers (Chen and Yao 2001, Gordon and Newell 1967). The number of total customers in the network is N. After the service of a customer has been completed at server i, this customer will depart from server i and enter server j with routing probabilities q_{ij}, $i,j=1,2,\cdots,M$. Without loss of generality, we assume $q_{ii}=0$. The service discipline in each server is FCFS (First-Come First-Served) and the buffer size is infinite. Let n_i be the number of customers at server i, and set $\mathbf{n}=(n_1,n_2,\cdots,n_i,\cdots,n_M)$. The service requirement of each customer at every server is assumed to be exponentially distributed with mean one, and the service rate of server i depends on $\mathbf{n}$ and is therefore denoted as $\mu_{i,\mathbf{n}}$, $i=1,2,\cdots,M$. It is easy to see that $\mathbf{n}$ is the system state, and the state space is $\mathscr{S}=\{all\ \mathbf{n}:\sum_{i=1}^{M}n_i=N\}$. We call this network a *state-dependent* closed Jackson network.

Let $\mathbf{n}(t)$ denote the system state at time t and $f(\mathbf{n}):\ \mathscr{S}\rightarrow\mathscr{R}=(-\infty,\infty)$ be the cost function. Let T_L be the Lth service completion time of the network (counting all the service completions of servers in the network). The *time-average performance* is defined as

$$\eta_T=\lim_{L\rightarrow\infty}\frac{\int_0^{T_L}f(\mathbf{n}(t))dt}{T_L}=\lim_{L\rightarrow\infty}\frac{F_L}{T_L},\tag{1}$$

where $F_L:=\int_0^{T_L}f(\mathbf{n}(t))dt$, and the *customer-average performance* is defined as

$$\eta^{(f)}=\lim_{L\rightarrow\infty}\frac{F_L}{L}.\tag{2}$$

We assume that the state process $\mathbf{n}(t)$ is ergodic so that the limits in (1) and (2) exist. The time-average performance is a traditional metric used in many systems, while the customer-average performance is also another important performance metric. These two performance metrics describe different aspects of the performance in queueing systems and both are widely used in practical applications. The difference between them will be discussed later.

We now review the perturbation realization factors in PA theory. In a closed network, if the service completion time of a server is delayed by a small amount of time Δ, we say that the server has a perturbation. This perturbation will affect the system performance $\eta^{(f)}$. The effect of a single perturbation Δ of server k when the system is at state $\mathbf{n}$ can be measured by the *perturbation realization factor*, which is defined as

$$\begin{aligned}c^{(f)}(\mathbf{n},k)&=\lim_{L\rightarrow\infty}\lim_{\Delta\rightarrow 0}E\left\{\frac{\Delta F_L}{\Delta}\right\}\\ =&\lim_{L\rightarrow\infty}\lim_{\Delta\rightarrow 0}E\left\{\frac{F'_L-F_L}{\Delta}\right\}\\ =&\lim_{L\rightarrow\infty}\lim_{\Delta\rightarrow 0}E\left\{\frac{1}{\Delta}\left(\int_0^{T'_L}f(\mathbf{n}'(t))dt-\int_0^{T_L}f(\mathbf{n}(t))dt\right)\right\},\end{aligned}\tag{3}$$

where $\mathbf{n}'(t)$ is the state of the perturbed sample path at time t, T'_L is the Lth service completion time on the perturbed sample path. From (3), we define the perturbation realization probability $c(\mathbf{n},k)$ as a special case of $c^{(f)}(\mathbf{n},k)$ where $f(\mathbf{n})=I(\mathbf{n})\equiv 1$ for all $\mathbf{n}\in\mathscr{S}$. $c(\mathbf{n},k)$ is the probability with which a perturbation of server k at state $\mathbf{n}$ will be realized at every server of the network ultimately.

With the perturbation realization factors $c^{(f)}(\mathbf{n},k)$, we may derive the system performance derivative with respect to the service rates as follows (Cao 1994).

$$\frac{d\eta^{(f)}}{d\mu_{k,\mathbf{n}}} = -\frac{\eta^{(I)}}{\mu_{k,\mathbf{n}}}\pi(\mathbf{n})c^{(f)}(\mathbf{n},k), \quad (4)$$

where $\eta^{(I)}$ is a special system performance corresponding to $f(\mathbf{n}) = I(\mathbf{n}) \equiv 1$ for all $\mathbf{n} \in \mathscr{S}$, and $\pi(\mathbf{n})$ is the steady-state probability of state $\mathbf{n}$. Since $\eta^{(I)}$ is a special case of $\eta^{(f)}$, from (2) we have

$$\eta^{(I)} = \lim_{L\to\infty} \frac{\int_0^{T_L} I(\mathbf{n}(t))dt}{L} = \lim_{L\to\infty} \frac{T_L}{L}. \quad (5)$$

It is easy to know that $\eta^{(I)}$ is the reciprocal of average throughput of the network, i.e.,

$$\eta_{th} = \lim_{L\to\infty} \frac{L}{T_L} = \frac{1}{\eta^{(I)}}, \quad (6)$$

where η_{th} is denoted as the throughput of queueing networks. With (1), (2), and (5), we have

$$\eta^{(f)} = \lim_{L\to\infty} \frac{F_L}{T_L}\frac{T_L}{L} = \eta_T \eta^{(I)}, \quad (7)$$

which shows the relationship between the customer-average performance $\eta^{(f)}$ and the time-average performance η_T. Since the service rates will affect the value of both η_T and $\eta^{(I)}$, the optimal values of service rates for $\eta^{(f)}$ and η_T are generally different. The numerical experiments in Section 4 demonstrate this point.

If we substitute $f(\mathbf{n}) = I(\mathbf{n})$ into (4), we get the derivative equation of $\eta^{(I)}$ as below.

$$\frac{d\eta^{(I)}}{d\mu_{k,\mathbf{n}}} = -\frac{\eta^{(I)}}{\mu_{k,\mathbf{n}}}\pi(\mathbf{n})c(\mathbf{n},k). \quad (8)$$

These formulas (4) and (8) are important for the gradient-based algorithm of PA theory in queueing systems. We can estimate the performance derivative from sample paths and use the gradient-descent algorithms to optimize the system performance through adjusting the service rates (Cao 1994, Chong and Zak 2001).

3 MAX-MIN OPTIMALITY OF SERVICE RATES

In this section, we consider the optimization problem of a state-dependent closed Jackson network, where the optimization parameters are the service rates $\mu_{k,\mathbf{n}}$ of each server k at each state $\mathbf{n}$, $k = 1, 2, \cdots, M$, $\mathbf{n} \in \mathscr{S}$. The objective is to minimize (or maximize) the customer-average performance $\eta^{(f)}$. With the notation of Markov decision processes (Puterman 1994), we can denote $\mathscr{U} = \{\mu_{k,\mathbf{n}}, k = 1, 2, \cdots, M, \mathbf{n} \in \mathscr{S}\}$ as a stationary policy based on parameter $\mu_{k,\mathbf{n}}$. The total policy space is $\Psi = \{all\ \mathscr{U}\}$, which is parameterized by $\mu_{k,\mathbf{n}}$. The cost function is $f(\mathbf{n}), \mathbf{n} \in \mathscr{S}$, which may be different for different actions. Furthermore, we denote the cost function as $f(\mathbf{n}, \vec{\mu}_{\mathbf{n}})$, where $\vec{\mu}_{\mathbf{n}} := (\mu_{1,\mathbf{n}}, \mu_{2,\mathbf{n}}, \cdots, \mu_{M,\mathbf{n}})$ is the action at state $\mathbf{n}$. With this notation, a policy can be denoted as $\mathscr{U} = \{\vec{\mu}_{\mathbf{n}}, \mathbf{n} \in \mathscr{S}\}$.

With PA of queueing systems, we can use the performance derivative equation (4) to optimize the service rates. But this type of gradient-descent algorithms has a slow convergence speed and may be trapped in a local optimum. In the middle of 90's of the last century, PA theory has been extended to Markov systems (Cao and Chen 1997, Cao 2003). The performance difference equation and policy iteration algorithm are derived along this direction. But this methodology can only optimize the traditional time-average performance. Recently, the performance difference equation for customer-average performance is also derived (Xia, Chen, and Cao 2008). It gives a new way to do optimization of queueing systems with customer-average performance criterion.

With the relationship formula of perturbation realization factors between queueing systems and Markov systems (Xia and Cao 2006b), researchers derive the customer-average performance difference equation for queueing systems. Here we just list the main results without explanation. The detailed proof can be found in Xia, Chen, and Cao (2008). When the service rates $\mu_{k,\mathbf{n}}$ of server k at state $\mathbf{n}$ is changed to $\mu'_{k,\mathbf{n}}$, the customer-average performance of closed Jackson networks will be changed from $\eta^{(f)}$ to $\eta'^{(f)}$. The performance difference equation is derived as below.

$$\eta'^{(f)} - \eta^{(f)} = \eta'^{(I)}\pi'(\mathbf{n})\left\{\frac{-\Delta\mu_{k,\mathbf{n}}}{\mu_{k,\mathbf{n}}}c^{(f)}(\mathbf{n},k) + h(\mathbf{n})\right\}, \quad (9)$$

where $\Delta\mu_{k,\mathbf{n}} = \mu'_{k,\mathbf{n}} - \mu_{k,\mathbf{n}}$, $h(\mathbf{n}) = f'(\mathbf{n}) - f(\mathbf{n})$, the superscript 'prime' represents the parameters of the perturbed system.

If the service rates of all the servers at all the states have changes, the performance difference equation is derived similarly as below.

$$\eta'^{(f)} - \eta^{(f)} =$$
$$\eta'^{(I)} \sum_{\mathbf{n}\in\mathscr{S}} \pi'(\mathbf{n})\left\{\sum_{k=1}^{M}\frac{-\Delta\mu_{k,\mathbf{n}}}{\mu_{k,\mathbf{n}}}c^{(f)}(\mathbf{n},k) + h(\mathbf{n})\right\}. \quad (10)$$

Furthermore, with the property of perturbation realization factors $c^{(f)}(\mathbf{n},k)$, equation (10) can be simplified as follows (Xia, Chen, and Cao 2008).

$$\eta'^{(f)} - \eta^{(f)} = \eta'^{(I)} \sum_{\mathbf{n}\in\mathscr{S}} \pi'(\mathbf{n})\left\{f'(\mathbf{n}) - \sum_{k=1}^{M}\frac{\mu'_{k,\mathbf{n}}}{\mu_{k,\mathbf{n}}}c^{(f)}(\mathbf{n},k)\right\}. \quad (11)$$

With these performance difference equations, we can find the optimal service rates of closed Jackson networks have a Max-Min optimality if the cost functions satisfy some requirements. The detailed Max-Min optimality is described as below.

Theorem 1 **(Max-Min Optimality for State-Dependent Service Rates)** *In a state-dependent closed Jackson network, if* $f(\mathbf{n}) = l_0(\mathbf{n}) + \sum_{k=1}^{M} l(\mathbf{n},k)\mu_{k,\mathbf{n}}$, $l_0(\mathbf{n})$ *and* $l(\mathbf{n},k)$ *are constants, i.e.,* $f(\mathbf{n})$ *changes linearly with service rates* $\mu_{k,\mathbf{n}}$*, then the optimal service rate of each server can be either maximal or minimal.*

Proof. In a state-dependent closed Jackson network, when the service rate of server k changes from $\mu_{k,\mathbf{n}}$ to $\mu_{k,\mathbf{n}} + \Delta\mu_{k,\mathbf{n}}$, the cost function f at state $\mathbf{n}$ also changes from $f(\mathbf{n})$ to $f'(\mathbf{n})$. Since $f(\mathbf{n}) = l_0(\mathbf{n}) + \sum_{k=1}^{M} l(\mathbf{n},k)\mu_{k,\mathbf{n}}$, we have $h(\mathbf{n}) = f'(\mathbf{n}) - f(\mathbf{n}) = \Delta\mu_{k,\mathbf{n}} l(\mathbf{n},k)$, where $l(\mathbf{n},k)$ is a constant which means the changed cost function per unit changed service rate $\Delta\mu_{k,\mathbf{n}}$. So, the performance difference equation (9) becomes

$$\eta'^{(f)} - \eta^{(f)} = \eta'^{(I)}\pi'(\mathbf{n})\left\{\frac{-\Delta\mu_{k,\mathbf{n}}}{\mu_{k,\mathbf{n}}}c^{(f)}(\mathbf{n},k) + h(\mathbf{n})\right\}$$

$$= \eta'^{(I)}\pi'(\mathbf{n})\Delta\mu_{k,\mathbf{n}}\left\{-\frac{c^{(f)}(\mathbf{n},k)}{\mu_{k,\mathbf{n}}} + l(\mathbf{n},k)\right\}. \qquad (12)$$

With (12), we can prove Theorem 1 easily. We first assume that the optimal service rate $\mu_{k,\mathbf{n}}$ is neither maximal nor minimal. Since $\eta'^{(I)}$ and $\pi'(\mathbf{n})$ are always positive, from (12) it is easy to know that $\left\{l(\mathbf{n},k) - \frac{c^{(f)}(\mathbf{n},k)}{\mu_{k,\mathbf{n}}}\right\}$ should be zero (otherwise we can change $\mu_{k,\mathbf{n}}$ to get a better $\eta'^{(f)}$). Since $\left\{l(\mathbf{n},k) - \frac{c^{(f)}(\mathbf{n},k)}{\mu_{k,\mathbf{n}}}\right\}$ is zero, it is obvious that the maximal or minimal value of $\mu_{k,\mathbf{n}}$ can also achieve the optimal performance. Therefore, it is proved that we can get the optimal performance from the maximal or minimal values of service rates. □

This theorem extends the similar results of Ma and Cao (1994), Yao and Schechner (1989), which discuss the optimization of the time-average performance of load-dependent networks, to the optimization of the customer-average performance of state-dependent closed networks. It implies that in the performance optimization of closed queueing networks, if f changes linearly with service rates, we only need to consider the maximal or minimal service rates for every state $\mathbf{n}$. So, the action space is reduced to 2^M. Using this property may speed up the optimization procedure greatly. Moreover, from (12) it is easy to know that $\eta^{(f)}$ is monotone with respect to $\mu_{k,\mathbf{n}}$ if we only change $\mu_{k,\mathbf{n}}$ and fix other parameters.

So far, we have discussed the optimization of state-dependent service rates in closed Jackson networks. The Max-Min optimality of service rates have been proved and it can be used to simplify the optimization procedure. With the performance difference equations (9) and (11), the policy iteration type of algorithms can be further derived. This type of algorithms can be implemented on-line based on the estimation of realization factors on a single sample path. The details of algorithm can be found in Xia, Chen, and Cao (2008) and we will use it to do numerical experiments in Section 4. Below, we further discuss the optimization problem for load-dependent service rates in closed Jackson networks. We use the similar procedure to prove the Max-Min optimality for load-dependent service rates.

In a *load-dependent* closed Jackson network, the service rate of a server k is dependent only on the queue length at server k. We denote the load-dependent service rates as μ_{k,n_k}, $k = 1,2,\cdots,M$, $n_k = 0,1,\cdots,N$. Obviously, it is known that $\mu_{k,n_k} \equiv 0$ when $n_k = 0$. As we know, the steady-state distribution of closed Jackson network has a product-form solution. With the product-form solution and customer-based aggregation of realization factors, the performance difference equation is derived as follows when the service rates are changed from μ_{k,n_k} to μ'_{k,n_k} for a particular server k and $n_k = 1,2,\cdots,N$ (Xia and Cao 2006a).

$$\eta'^{(f)} - \eta^{(f)} =$$

$$\eta'^{(I)}\sum_{n_k=1}^{N}\pi'(n_k)\left\{\frac{-\Delta\mu_{k,n_k}}{\mu_{k,n_k}}\tilde{c}^{(f)}(n_k,k) + \tilde{h}(n_k,k)\right\}, \qquad (13)$$

where $\tilde{c}^{(f)}(n_k,k) := \sum_{\mathbf{n}\in\mathscr{S}_{n_k}} \pi(\mathbf{n}|n_k)c^{(f)}(\mathbf{n},k)$, $\tilde{h}(n_k,k) := \sum_{\mathbf{n}\in\mathscr{S}_{n_k}} \pi(\mathbf{n}|n_k)h(\mathbf{n},k) = \sum_{\mathbf{n}\in\mathscr{S}_{n_k}} \pi(\mathbf{n}|n_k)[f'(\mathbf{n}) - f(\mathbf{n})]$, $\mathscr{S}_{n_k}$ is the set of states where the number of customers at server k is equal to n_k.

With the performance difference equation (13), we can also derive the Max-Min optimality of service rates in a load-dependent closed Jackson network. The proof is similar to that of Theorem 1 and it is neglected for the simplicity of paper.

Theorem 2 **(Max-Min Optimality for Load-Dependent Service Rates)** *In a load-dependent closed Jackson network, if* $f(\mathbf{n}) = l_0(\mathbf{n}) + \sum_{k=1}^{M} l(n_k,k)\mu_{k,n_k}$, $l_0(\mathbf{n})$ *and* $l(n_k,k)$ *are constants, i.e.,* $f(\mathbf{n})$ *changes linearly with service rates* μ_{k,n_k}*, then the optimal service rate of each server can be either maximal or minimal.*

With Theorem 2 we know if the cost function is linear with the service rates, we only need to choose the maximal or minimal service rates. It is similar to Theorem 1. These optimality properties can greatly simplify the optimization complexity of closed Jackson networks. With these theorems, we can only focus on the maximal or minimal values of service rates when we apply any optimization algorithm. This is also called the Bang-Bang control in the area of control science.

Table 1: Optimal service rates of a state-dependent closed Jackson network.

$\mathbf{n}$	$\mu^0_{1,\mathbf{n}}$	$\mu^0_{2,\mathbf{n}}$	$\mu^0_{3,\mathbf{n}}$	$D_{\mu_{1,\mathbf{n}}}$	$D_{\mu_{2,\mathbf{n}}}$	$D_{\mu_{3,\mathbf{n}}}$	$\mu^*_{1,\mathbf{n}}$	$\mu^*_{2,\mathbf{n}}$	$\mu^*_{3,\mathbf{n}}$	$\mu'^*_{1,\mathbf{n}}$	$\mu'^*_{2,\mathbf{n}}$	$\mu'^*_{3,\mathbf{n}}$
(0,0,5)	0.20	0.20	0.30	[0.01,2.00]	[0.03,1.50]	[0.09,3.00]	0.01	0.03	3.00	0.01	0.03	0.09
(0,1,4)	0.20	0.20	0.30	[0.02,2.50]	[0.02,1.50]	[0.03,3.50]	0.02	0.02	3.50	0.02	0.02	0.03
(0,2,3)	0.20	0.10	0.30	[0.01,1.50]	[0.05,2.50]	[0.05,4.50]	0.01	2.50	0.05	0.01	0.05	0.05
(0,3,2)	0.30	0.30	0.25	[0.03,1.50]	[0.06,3.00]	[0.01,1.50]	0.03	3.00	0.01	0.03	0.06	0.01
(0,4,1)	0.20	0.35	0.20	[0.02,2.00]	[0.09,2.00]	[0.04,2.00]	0.02	2.00	0.04	0.02	0.09	0.04
(0,5,0)	0.10	0.40	0.40	[0.05,1.50]	[0.04,2.50]	[0.01,3.00]	0.05	2.50	0.01	0.05	0.04	0.01
(1,0,4)	0.10	0.30	0.15	[0.04,2.00]	[0.05,3.00]	[0.06,2.00]	2.00	0.05	0.06	2.00	0.05	0.06
(1,1,3)	0.20	0.25	0.25	[0.05,2.50]	[0.02,3.50]	[0.08,1.50]	2.50	0.02	0.08	2.50	0.02	0.08
(1,2,2)	0.20	0.10	0.30	[0.02,2.00]	[0.06,1.50]	[0.04,2.50]	2.00	0.06	0.04	2.00	0.06	0.04
(1,3,1)	0.30	0.30	0.25	[0.06,1.50]	[0.03,2.00]	[0.02,2.50]	1.50	0.03	0.02	1.50	0.03	0.02
(1,4,0)	0.20	0.25	0.30	[0.04,2.50]	[0.05,3.50]	[0.07,3.50]	2.50	0.05	0.07	2.50	0.05	0.07
(2,0,3)	0.25	0.10	0.10	[0.06,1.00]	[0.03,3.50]	[0.01,3.00]	1.00	0.03	3.00	1.00	0.03	0.01
(2,1,2)	0.10	0.30	0.35	[0.07,1.50]	[0.04,4.00]	[0.06,2.00]	1.50	0.04	2.00	1.50	0.04	0.06
(2,2,1)	0.10	0.25	0.15	[0.08,2.50]	[0.01,2.00]	[0.05,3.50]	2.50	0.01	0.05	2.50	0.01	0.05
(2,3,0)	0.15	0.30	0.30	[0.07,2.00]	[0.02,1.50]	[0.01,4.00]	2.00	1.50	0.01	2.00	0.02	0.01
(3,0,2)	0.20	0.20	0.10	[0.03,2.00]	[0.01,1.00]	[0.04,2.00]	2.00	0.01	2.00	2.00	0.01	0.04
(3,1,1)	0.35	0.10	0.20	[0.05,2.50]	[0.01,2.00]	[0.08,4.00]	2.50	0.01	4.00	2.50	0.01	0.08
(3,2,0)	0.25	0.20	0.20	[0.09,4.50]	[0.06,3.50]	[0.06,3.50]	4.50	0.06	0.06	4.50	0.06	0.06
(4,0,1)	0.40	0.30	0.20	[0.01,3.50]	[0.08,1.50]	[0.09,2.00]	3.50	0.08	2.00	3.50	0.08	0.09
(4,1,0)	0.35	0.10	0.20	[0.06,4.00]	[0.04,3.00]	[0.04,2.50]	4.00	0.04	0.04	4.00	0.04	0.04
(5,0,0)	0.30	0.20	0.10	[0.04,4.50]	[0.06,3.50]	[0.05,2.00]	4.50	0.06	0.05	4.50	0.06	0.05

4 NUMERICAL EXPERIMENTS

In this section, we use the policy iteration algorithm to optimize the service rates of closed Jackson networks. The numerical results demonstrate the Max-Min optimality of service rates. The detailed implementation of policy iteration algorithm is neglected for the length limitation of paper. Readers who are interested can refer the literature Xia, Chen, and Cao (2008) and Xia and Cao (2006a), which are respectively for state-dependent and load-dependent closed Jackson networks.

First, we consider the optimization of state-dependent service rates in closed Jackson network. The number of servers is $M = 3$ and the number of customers is $N = 5$. The routing probability matrix is

$$Q = \begin{bmatrix} 0 & 0.4 & 0.6 \\ 0.7 & 0 & 0.3 \\ 0.2 & 0.8 & 0 \end{bmatrix}.$$

The initial service rates $\mu^0_{k,\mathbf{n}}$ are listed in Table 1. The cost function is defined as $f(\mathbf{n}) = n_1 + r \cdot \sum_{k=1}^{3} \mu_{k,\mathbf{n}}$, where r is the cost of unit service rate and we set $r = 1$ in this example. In fact, r can be a set of values which represent the different costs at different states. The objective is to minimize the customer-average performance $\eta^{(f)}$ through adjusting the service rates $\mu_{k,\mathbf{n}}$, $k = 1,2,3$ and $\mathbf{n} \in \mathscr{S}$. The value domains of service rates are listed as $D_{\mu_{k,\mathbf{n}}}$ in Table 1.

With the policy iteration algorithm for customer-average performance (Xia, Chen, and Cao 2008), we find the algorithm iterates some times and locates the optimal solution finally. The optimal service rates $\mu^*_{k,\mathbf{n}}$ are listed in Table 1. The minimal customer-average performance is $\eta^{*(f)} = 1.1601$. From Table 1 we can also see that the optimal service rates are either maximal or minimal. It demonstrates the correctness of Theorem 1.

As a comparison, we use the traditional algorithms in Markov decision processes to optimize the time-average performance of this problem. With this performance metric, we obtain the optimal service rates $\mu'^*_{k,\mathbf{n}}$ for time-average performance, which are listed in Table 1. The corresponding $\eta^{(f)} = 1.3845$, which is larger than the optimal value $\eta^{*(f)} = 1.1601$. The optimal service rates $\mu'^*_{k,\mathbf{n}}$ for time-average performance are also different from the optimal service rates $\mu^*_{k,\mathbf{n}}$ for customer-average performance. This experiment illustrates that the algorithm for time-average performance can not be directly used to optimize the customer-average performance. These two performance metrics characterize the different aspects of optimization problems.

Furthermore, in order to more clearly illustrate the difference between customer-average performance and time-average performance, we give another numerical example as follows. Consider a cyclic network with 2 servers. The number of customers is $N=3$. For simplicity, we can use n_1 to represent the system state $\mathbf{n}$. The service rates of server 2 is fixed at $\mu_{2,n_1}=1$, $n_1=0,1,2,3$. The service rates of server 1 is $\mu_{1,n_1}=1$ when $n_1=0,1,2$, and $\mu_{1,n_1}\in[0.5,2]$ when $n_1=3$. The cost function is $f(0)=f(1)=f(2)=1$, $f(3)=-\mu_{1,3}(1-\mu_{1,3})^2$. The objective is to choose an optimal $\mu_{1,3}$ to minimize the system performance $\eta^{(f)}$. With numerical computation, we obtain the performance curve as Figure 1. It is obvious that the optimal $\mu_{1,3}$ is 2 for $\eta^{(f)}$, while the optimal $\mu_{1,3}$ is 0.5 for η_T. This example clearly demonstrates the difference between these two performance metrics.

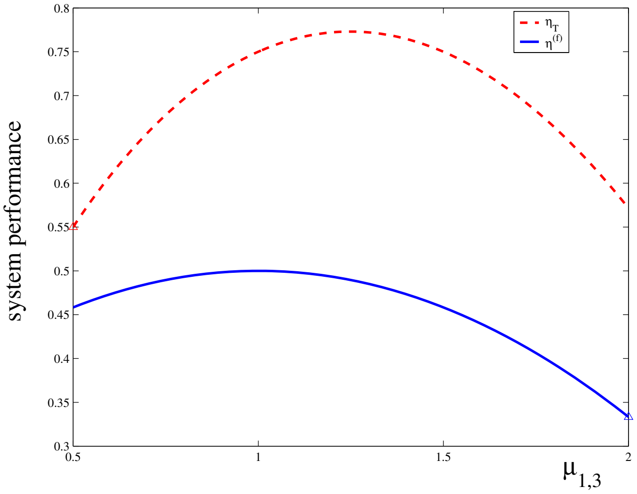

Figure 1: Difference between customer-average performance and time-average performance.

The above numerical experiment is about the state-dependent service rates. The situation for load-dependent service rates is similar and here is not presented for the paper concision. Interested readers can refer the numerical experiment by Xia and Cao (2006a), which can also demonstrate the Max-Min optimality of load-dependent service rates.

5 CONCLUSION

In this paper we discuss the service rates optimization problem in closed Jackson networks. With the difference equation of customer-average performance, we prove that the optimal service rates have a so-called Max-Min optimality. This optimality is valid for both state-dependent and load-dependent closed Jackson networks. With the Max-Min optimality, we only need to choose the maximal or minimal values for service rates in optimization procedure. It simplifies greatly the complexity of such type of optimization problems.

ACKNOWLEDGMENTS

The first author would like to thank Prof. Xi-Ren Cao for his instruction and help when the author did this research work in his Ph.D. pursuing period. Of course, the possible errors in this paper are solely the responsibility of the paper authors.

REFERENCES

Cao, X. R. 1994. *Realization probabilities - the dynamics of queueing systems*. New York: Springer Verlag.

Cao, X. R. 2003. From perturbation analysis to markov decision processes and reinforcement learning. *Discrete Event Dynamic Systems: Theory and Applications* 13:9–39.

Cao, X. R., and H. F. Chen. 1997. Potentials, perturbation realization, and sensitivity analysis of markov processes. *IEEE Transactions on Automatic Control* 42:1382–1393.

Cassandras, C., and S. Lafortune. 1999. *Introduction to discrete event systems*. Boston, MA: Kluwer Academic Publishers.

Chen, H., and D. Yao. 2001. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*. New York: Springer Verlag.

Chong, E. K. P., and S. H. Zak. 2001. *An introduction to optimization, 2^{nd} edition*. New York: John Wiley & Sons.

Glasserman, P. 1991. *Gradient estimation via perturbation analysis*. Norwell, MA: Kluwer Academic Publisher.

Gong, W. B., and Y. C. Ho. 1987. Smoothed (conditional) perturbation analysis of discrete event dynamical systems. *IEEE Transactions on Automatic Control* 32:858–866.

Gordon, W. J., and G. F. Newell. 1967. Closed queueing systems with exponential servers. *Operations Research* 15:252–265.

Ho, Y. C., and X. R. Cao. 1991. *Perturbation analysis of discrete-event dynamic systems*. Boston, MA: Kluwer Academic Publisher.

Ho, Y. C., X. R. Cao, and C. Cassandras. 1983. Infinitesimal and finite perturbation analysis for queueing networks. *Automatica* 19:439–445.

Ma, D. J., and X. R. Cao. 1994. A direct approach to decentralized control of service rates in a closed jackson network. *IEEE Transactions on Automatic Control* 39:1460–1463.

Puterman, M. L. 1994. *Markov decision processes: Discrete stochastic dynamic programming*. New York: John Wiley & Sons.

Xia, L., and X. R. Cao. 2006a. Aggregation of perturbation realization factors and service rate-based policy iteration for queueing systems. In *Proceedings of the 45th IEEE Conference on Decision and Control, San Diego, CA, USA*, 1063–1068.

Xia, L., and X. R. Cao. 2006b. Relationship between perturbation realization factors with queueing models and markov models. *IEEE Transactions on Automatic Control* 51:1699–1704.

Xia, L., X. Chen, and X. R. Cao. 2008. Policy iteration of customer-average performance in queueing systems. *Submitted to Automatica*.

Yao, D. D., and Z. Schechner. 1989. Decentralized control of service rates in a closed jackson network. *IEEE Transactions on Automatic Control* 34:236–240.

AUTHOR BIOGRAPHIES

LI XIA received the B.E. degree in automation in July 2002 and the Ph.D. degree in control science and engineering in July 2007, both from Tsinghua University, Beijing, China. He is currently a research staff at IBM China Research Laboratory, Beijing, China. His research interests include discrete event dynamic systems (DEDS) theory and applications, simulation optimization techniques, and supply chain management. His email address is <crlxiali@cn.ibm.com>.

MING XIE received his Ph.D. degree in Tsinghua University from China in 2006 and then joined IBM China Research Laboratory. Currently, his research interests include supply chain management, agent-based simulation, complex systems and business intelligence. His email address is <xieming@cn.ibm.com>.

WENJUN YIN received his Ph.D. degree in Tsinghua University from China in 2004 and then joined IBM China Research Laboratory. Currently, his research interests include supply chain management and logistics, business analytics and optimization, data mining and business intelligence. His email address is <yinwenj@cn.ibm.com>.

JIN DONG, Manager of Supply Chain Management and Logistics Research in IBM China Research Laboratory. He received his Ph.D. degree in Tsinghua University from China in 2001. Before joined IBM, he was the Research Assistant Professor in Industrial Engineering Department of Arizona State University in USA. His email address is <dongjin@cn.ibm.com>.