Max-Min Optimality of Service Rate Control in Closed Queueing Networks

Li Xia *Member, IEEE,* and Basem Shihada

***Abstract*—In this paper, we discuss the optimality properties of service rate control in closed Jackson networks. We prove that when the cost function is linear to a particular service rate, the system performance is monotonic w.r.t. (with respect to) that service rate and the optimal value of that service rate can be either maximum or minimum (we call it Max-Min optimality); When the second-order derivative of the cost function w.r.t. a particular service rate is always positive (negative), which makes the cost function strictly convex (concave), the optimal value of such service rate for the performance maximization (minimization) problem can be either maximum or minimum. To the best of our knowledge, this is the most general result for the optimality of service rates in closed Jackson networks and all the previous works only involve the first conclusion. Moreover, our result is also valid for both the state-dependent and load-dependent service rates, under both the time-average and customer-average performance criteria.**

Keywords: service rate control, perturbation analysis, closed Jackson network, discrete event dynamic systems

I. Introduction

Queueing model is an important modeling technique in stochastic processes and operations research. In this paper, we study the optimal control of service rates in a closed Jackson network. The service rates of each server are adjustable according to the queue length of that server or the system state, which are called *load-dependent* service rates or *state-dependent* service rates, respectively. The cost function is determined by the system state and service rates. Each service rate has a certain value domain. Our objective is to identify a set of service rates which makes the system performance optimal under the criteria of *time-average* or *customer-average*.

The problem of service rate control of queueing networks has been richly studied in the literature. [7] studied the threshold optimality of service rates for a two-node cyclic network. [12] discussed a network where a number of queues are connected in a circle. Similar studies were also conducted on other simple queueing models, such as tandem queues and cyclic queues [11]. However, most of these studies are based on the optimality equation of Markov decision process (MDP) and use an inductive analysis to study the structure of optimal policies. Such method is difficult to be applied to a general queueing network (such as Jackson networks) because the associated optimality equation is too complex to analyze. For a closed Jackson network, [15] used linear programming to prove that the optimal service rates can be threshold-type when the cost function includes a holding cost determined by system state and an operating cost which is a linear summation of all the service rates. [8] further used the performance derivative to directly prove that when the cost function is linear to a service rate, the time-average cost is monotonic w.r.t. that service rate. The result obtained by [8] is considered more general compared to its previous studies. [14] proved that the system performance is monotonic to a service rate when the cost function is linear to that service rate under the criterion of customer-average performance.

In this paper, we use the performance difference equation to study the service rate control problem. Difference equation is an important progress of perturbation analysis (PA) in the past decade [3], [4], [13]. It can obtain the performance difference of a Markov system under any two policies. This method provides a much clearer perspective to study the structure of optimal policies. Utilizing the property of the product-form solution of Jackson networks, we obtain a concise form of difference equation for the service rate control problem. Based on the difference equation, we further derive the first-order and second-order performance derivatives w.r.t. service rates. With these derivatives, we fold our new results in the following two aspects. First, when the cost function is linear to a service rate, the system performance will be monotonic w.r.t. that service rate. Therefore, that service rate has the Max-Min optimality, without considering any middle values. This result is equivalent to the most general results in the literature [8]. Second, when the second-order derivative of the cost function w.r.t. a service rate is always positive (negative), the optimal value of that service rate for the performance maximization (minimization) problem can be either maximum or minimum. The second condition requires the cost function be strictly convex (concave) w.r.t. that service rate. This second conclusion is a totally new progress compared with all of the previous works in the literature. Compared with the linear function, strictly convex (concave) function is also very common in the practice.

The contributions of this paper are folded in the following three aspects. First, we extend the Max-Min optimality of service rates to a more general form of cost functions. That is, we extend the form of cost functions from linearity to strictly convexity (concavity). Second, we prove that our results are valid for both the load-dependent and state-dependent service rates, under both the time-average and customer-average criteria. Most of the previous studies focused on the load-dependent

The work was supported in part by the National Natural Science Foundation of China (60736027), the National 111 International Collaboration Project (B06002), and the TNList Cross-Discipline Foundation.

Li Xia is with the Center for Intelligent and Networked Systems (CFINS), Department of Automation, TNList, Tsinghua University, Beijing 100084, China (e-mail: xial@tsinghua.edu.cn).

Basem Shihada is with the Division of Mathematical and Computer Science & Engineering, King Abdullah University of Science and Technology, Thuwal 21534, Saudi Arabia (e-mail: basem.shihada@kaust.edu.sa).

service rates with a time-average criterion. Finally, compared with the complicated proofs in the previous studies, our paper presents straightforward proofs. This conciseness benefits from the performance difference equation which directly describes the change of system performance when the service rates change. This is a new and efficient approach to optimize the performance of queueing systems.

II. Background of Perturbation Analysis

PA is an analytical methodology to optimize the performance of discrete event dynamic systems (DEDS) [1]. It aims to efficiently extract the performance sensitivity information from a single sample path by utilizing the special structure of systems [2], [5], [6]. In the past decade, PA has been extended from the original gradient-based optimization to the recent difference-based optimization [3], [4], [13], which can provide more sensitivity information. In this section, we give a brief overview on PA theory, focusing on the recent progress on the difference-based optimization.

Consider a closed Jackson network with M servers and N customers. The service time of each server obeys an exponential distribution. When a customer completes its service at server i, it will be transferred to server j with a routing probability q_{ij}, $i,j = 1,2,\cdots,M$, and $\sum_j q_{ij} = 1$ for all i. The service discipline is FCFS (first come first serve) and the buffer of every server is adequate. The queue length of server i is n_i and the system state is $\boldsymbol{n} = (n_1, n_2, \cdots, n_M)$, $n_i = 0, 1, \cdots, N$, $i = 1, 2, \cdots, M$, and $\sum_i n_i = N$. The service rate of server i is denoted as μ_i, and it can be further denoted as μ_{i,n_i} for load-dependent service rate or $\mu_{i,\boldsymbol{n}}$ for state-dependent service rate. The cost function of the system is denoted as $f(\boldsymbol{n})$, $\boldsymbol{n} \in \mathcal{S}$, where $\mathcal{S} = \{\text{all } \boldsymbol{n}\}$ is called the system state space. Note that for the service rate control problem in Section III, $f(\boldsymbol{n})$ is related to service rates and can be further notated as $f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}})$.

For simplicity, we assume that the network is strongly connected (any customer may visit every server in the network), which is true for most situations. Therefore, the stochastic process of such queueing network is ergodic. The *customer-average performance* η_C of the system is defined as follows.

$$\eta_C = \lim_{L\to\infty} \frac{1}{L} \int_0^{T_L} f(\boldsymbol{n}(t))dt, \tag{1}$$

where T_L is the time when the queueing network has exactly served L customers, $\boldsymbol{n}(t)$ is the system state at time t. The *time-average performance* η_T is defined as

$$\eta_T = \lim_{L\to\infty} \frac{1}{T_L} \int_0^{T_L} f(\boldsymbol{n}(t))dt. \tag{2}$$

η_C and η_T has the following relationship

$$\eta_C = \lim_{L\to\infty} \frac{T_L}{L}\frac{1}{T_L} \int_0^{T_L} f(\boldsymbol{n}(t))dt = \eta_I \eta_T, \tag{3}$$

where $\eta_I := \lim_{L\to\infty} \frac{T_L}{L}$ is a special case of η_C when $f(\boldsymbol{n}) \equiv 1$ for all $\boldsymbol{n}$. Actually, the reciprocal of η_I is the average throughput of the network which is denoted as η_{th}. From (3), we observe that if the change of service rates affects both η_T and η_I, the optimal service rates for η_T and η_C are generally different [13]. Thus, it is necessary to discuss the optimization for time-average and customer-average, respectively.

The central idea of PA is to measure the effect of a single perturbation of system parameters on the whole system performance, and use such fundamental measurements as building blocks to study the performance sensitivity w.r.t. parameters. In queueing systems, *perturbation realization factor* $c^{(f)}(\boldsymbol{n}, i)$ is such a measurement and it quantifies the average effect of a unit delay Δ of service time of server i at state $\boldsymbol{n}$ on the whole system performance. It is defined as follows.

$$c^{(f)}(\boldsymbol{n}, i) = \lim_{\substack{L\to\infty \\ \Delta\to 0}} E\left\{\frac{1}{\Delta}\left[\int_0^{T'_L} f(\boldsymbol{n}'(t))dt - \int_0^{T_L} f(\boldsymbol{n}(t))dt\right]\right\}, \tag{4}$$

where $\boldsymbol{n}'(t)$ is the perturbed system process with delay Δ and T'_L is the time when the perturbed system has served L customers.

With the perturbation realization factor, PA gives the difference equation of customer-average performance of queueing systems. Consider state-dependent service rates $\mu_{i,\boldsymbol{n}}$ are changed to $\mu'_{i,\boldsymbol{n}}$, for all $i = 1, 2, \cdots, M$ and $\boldsymbol{n} \in \mathcal{S}$, the change of η_C can be measured by the following difference equation [13]

$$\begin{aligned} \eta'_C - \eta_C &= \eta'_I \sum_{\boldsymbol{n}\in\mathcal{S}} \pi'(\boldsymbol{n})\Big\{\sum_{i=1}^{M} \frac{\mu_{i,\boldsymbol{n}} - \mu'_{i,\boldsymbol{n}}}{\mu_{i,\boldsymbol{n}}} c^{(f)}(\boldsymbol{n}, i) \\ &\quad + [f'(\boldsymbol{n}) - f(\boldsymbol{n})]\Big\}, \end{aligned} \tag{5}$$

where the parameters with superscript "$'$" indicate the corresponding values of the system with changed service rates and $\pi'(\boldsymbol{n})$ denotes the steady-state probability at state $\boldsymbol{n}$.

The above introduction is mainly about PA for queueing models. As per [3], [4], this theory is also valid for Markov models.

Consider a continuous-time Markov process and its state at time t is denoted as X_t, $t \geq 0$. The system state space is finite and denoted as $\mathcal{S} = \{1, 2, \cdots, S\}$, where S is the size of state space. At each state transition epoch, we choose an action a from an action space $\mathcal{A}$ according to a policy $\mathcal{L}$. Here, we consider the Markovian and deterministic policy $\mathcal{L}$ which is a mapping function $\mathcal{L} : \mathcal{S} \to \mathcal{A}$. Different policies have different infinitesimal generators $B = [b_{uv}]_{u,v=1}^{S}$. The steady-state probability distribution is a row vector denoted as $\pi = (\pi(1), \pi(2), \cdots, \pi(S))$. We have $Be = 0$ and $\pi B = 0$, where e is a column vector whose elements are all 1. The cost function is a column vector denoted as $f = (f(1), f(2), \cdots, f(S))^T$ where the superscript "T" means the transpose operation. The time-average performance is $\eta_T = E\{f(X_t)\}$ and we have $\eta_T = \pi f$.

Similar to perturbation realization factors, there is a fundamental quantity called *performance potential* g. g is a column vector and it is also called bias or relative value function in MDP theory. The element $g(u)$, $u \in \mathcal{S}$, quantifies the long-term bias from the system performance η_T if the initial state is u. $g(u)$ is defined as

$$g(u) = \lim_{T\to\infty} \int_0^T [f(X_t) - \eta_T]dt \Big| X_0 = u. \tag{6}$$

With the performance potentials as building blocks, PA gives the following difference equation of time-average performance when the policy is changed from $\mathcal{L}$ to $\mathcal{L}'$ [3], [4].

$$\eta'_T - \eta_T = \pi' \left[(B' - B)g + (f' - f) \right], \tag{7}$$

where the superscript "$'$" means the corresponding parameters under the policy $\mathcal{L}'$.

In the next section, we use the difference equations (5) and (7) to discuss the optimality of service rates in closed Jackson networks.

III. Optimality Property of Service Rates

In this section, based on the difference equation of PA theory and the special structure of queueing networks, we discuss the problem of service rate control in closed Jackson networks. Our results cover both the load-dependent and state-dependent service rates under both the time-average and customer-average criteria. We first discuss the case with load-dependent service rates and time-average performance criterion. The other 3 cases will be briefly discussed later with similar analysis.

Suppose that the load-dependent service rate μ_{i,n_i} has a value domain denoted as $[a_{i,n_i},\ b_{i,n_i}]$, $i = 1, 2, \cdots, M$, $n_i = 1, 2, \cdots, N$ ($n_i = 0$ is omitted since the corresponding μ_{i,n_i} is definitely 0). The cost function f is determined by the state and service rates. Therefore, $f(\boldsymbol{n})$ in (5) and (7) is further notated as $f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}})$ and $f'(\boldsymbol{n})$ is notated as $f(\boldsymbol{n}, \vec{\mu}'_{\boldsymbol{n}})$, where $\vec{\mu}_{\boldsymbol{n}} := (\mu_{1,n_1}, \mu_{2,n_2}, \cdots, \mu_{M,n_M})$. The objective is to choose a set of proper service rates μ_{i,n_i}, $i = 1, 2, \cdots, M$, $n_i = 1, 2, \cdots, N$, which maximizes the system performance η_T.

This problem looks like an MDP. However, it is not a standard MDP because it violates the *independent-action assumption* in MDP theory [9]. That is, the actions cannot be chosen independently at different states in this problem. For example, in a closed Jackson network with 3 servers and 4 customers, the queue lengths of server 1 at two states $\boldsymbol{n} = (2, 0, 2)$ and $\boldsymbol{n}' = (2, 1, 1)$ are identical. Thus, the service rates (actions) of server 1 at this two states should be the same as $\mu_{1,2}$ and they cannot be chosen independently. Therefore, we cannot use MDP approach to solve this problem.

Below, we derive a useful property of closed Jackson network based on its product-form solution to solve the aforementioned difficulty. In a load-dependent closed Jackson network, the visit ratio of server i, v_i, is determined as follows.

$$v_i = \sum_{j=1}^{M} q_{ji} v_j, \qquad i = 1, 2, \cdots, M. \tag{8}$$

Let $A_i(0) = 1$ and

$$A_i(n_i) = \prod_{j=1}^{n_i} \mu_{i,j}, \tag{9}$$

where $i = 1, 2, \cdots, M$ and $n_i = 1, 2, \cdots, N$. We define

$$G_m(n) = \sum_{n_1 + \cdots + n_m = n} \prod_{j=1}^{m} \frac{v_j^{n_j}}{A_j(n_j)}, \tag{10}$$

where $m = 1, 2, \cdots, M$ and $n = 0, 1, \cdots, N$. It is well known that the steady-state distribution of closed Jackson networks has the following product-form solution

$$\pi(\boldsymbol{n}) = \frac{1}{G_M(N)} \prod_{j=1}^{M} \frac{v_j^{n_j}}{A_j(n_j)}. \tag{11}$$

With (11), we study the property of conditional probability $\pi(\boldsymbol{n}|n_i)$, where $n_i = 0, 1, \cdots, N$, $\boldsymbol{n} \in \mathcal{S}_{n_i}$, and $\mathcal{S}_{n_i}$ is the set of states where the number of customers at server i is fixed as n_i.

$$\begin{aligned}
\pi(\boldsymbol{n}|n_i) &= \frac{\pi(\boldsymbol{n})}{\sum_{\boldsymbol{n} \in \mathcal{S}_{n_i}} \pi(\boldsymbol{n})} = \frac{\prod_{j=1}^{M} \frac{v_j^{n_j}}{A_j(n_j)}}{\sum_{\boldsymbol{n} \in \mathcal{S}_{n_i}} \prod_{j=1}^{M} \frac{v_j^{n_j}}{A_j(n_j)}} \\
&= \frac{\frac{v_i^{n_i}}{A_i(n_i)} \prod_{j \neq i}^{M} \frac{v_j^{n_j}}{A_j(n_j)}}{\frac{v_i^{n_i}}{A_i(n_i)} \sum_{\boldsymbol{n} \in \mathcal{S}_{n_i}} \prod_{j \neq i}^{M} \frac{v_j^{n_j}}{A_j(n_j)}} = \frac{\prod_{j \neq i}^{M} \frac{v_j^{n_j}}{A_j(n_j)}}{\sum_{\boldsymbol{n} \in \mathcal{S}_{n_i}} \prod_{j \neq i}^{M} \frac{v_j^{n_j}}{A_j(n_j)}}.
\end{aligned} \tag{12}$$

From (12), we observe that the conditional probability $\pi(\boldsymbol{n}|n_i)$ has no relation with the service rates of server i. Therefore, if we only change the service rates of a particular server i from μ_{i,n_i} to μ'_{i,n_i}, $n_i = 1, 2, \cdots, N$, and fix other servers' service rates, the conditional probability has the following property

$$\pi'(\boldsymbol{n}|n_i) = \pi(\boldsymbol{n}|n_i), \qquad \text{when only } \mu_{i,n_i} \text{ is changed.} \tag{13}$$

This property has a vital role in studying the optimality of service rates in closed Jackson networks. Consider a particular service rate μ_{i,n_i} is changed to μ'_{i,n_i} and all the other service rates are fixed, the corresponding infinitesimal generator B will also be changed to B'. According to the structure of closed Jackson networks, we obtain the elements of matrix $\Delta B = B' - B$ as follows. For each $\boldsymbol{n} \in \mathcal{S}_{n_i}$, we have $\Delta B(\boldsymbol{n}, \boldsymbol{n}) = \mu_{i,n_i} - \mu'_{i,n_i}$; $\Delta B(\boldsymbol{n}, \boldsymbol{n}_{ij}) = q_{ij}(\mu'_{i,n_i} - \mu_{i,n_i})$ for all $j = 1, 2, \cdots, M$; and all the other elements of $\Delta B(\boldsymbol{n}, :)$ are 0, where $\boldsymbol{n}_{ij} = (n_1, \cdots, n_i - 1, \cdots, n_j + 1, \cdots, n_M)$ is called the neighboring state of $\boldsymbol{n}$. For all the other $\boldsymbol{n} \notin \mathcal{S}_{n_i}$, the elements of $\Delta B(\boldsymbol{n}, :)$ are always 0. Therefore, the difference equation (7) can be rewritten as follows.

$$\begin{aligned}
\eta'_T - \eta_T &= \sum_{\boldsymbol{n} \in \mathcal{S}_{n_i}} \pi'(\boldsymbol{n}) \Big\{ (\mu'_{i,n_i} - \mu_{i,n_i}) \sum_{j=1}^{M} q_{ij} [g(\boldsymbol{n}_{ij}) - g(\boldsymbol{n})] \\
&\qquad + [f(\boldsymbol{n}, \vec{\mu}'_{\boldsymbol{n}}) - f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}})] \Big\} \\
&= \pi'(n_i) \Big\{ (\mu'_{i,n_i} - \mu_{i,n_i}) \sum_{\boldsymbol{n} \in \mathcal{S}_{n_i}} \pi'(\boldsymbol{n}|n_i) \sum_{j=1}^{M} q_{ij} [g(\boldsymbol{n}_{ij}) - g(\boldsymbol{n})] \\
&\quad + \sum_{\boldsymbol{n} \in \mathcal{S}_{n_i}} \pi'(\boldsymbol{n}|n_i) [f(\boldsymbol{n}, \vec{\mu}'_{\boldsymbol{n}}) - f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}})] \Big\},
\end{aligned} \tag{14}$$

where $\pi'(n_i)$ is the marginal probability that the number of customers at server i is n_i. Substituting (13) into (14), we

obtain

$$\eta'_T - \eta_T = \pi'(n_i)\Big\{(\mu'_{i,n_i} - \mu_{i,n_i}) \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i) \sum_{j=1}^{M} q_{ij} \big[g(\boldsymbol{n}_{ij}) - g(\boldsymbol{n})\big] + \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)[f(\boldsymbol{n},\vec{\mu}'_{\boldsymbol{n}}) - f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})]\Big\}. \quad (15)$$

Since $\pi(\boldsymbol{n}|n_i)$ and g are only based on the current system behavior and not related to the perturbed system, we define

$$\tilde{g}(n_i,i) = \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i) \sum_{j=1}^{M} q_{ij}[g(\boldsymbol{n}_{ij}) - g(\boldsymbol{n})]. \quad (16)$$

Difference equation (15) can be rewritten as follows.

$$\eta'_T - \eta_T = \pi'(n_i)\Big\{[\mu'_{i,n_i} - \mu_{i,n_i}]\tilde{g}(n_i,i) + \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)[f(\boldsymbol{n},\vec{\mu}'_{\boldsymbol{n}}) - f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})]\Big\}. \quad (17)$$

Suppose that the current service rate is μ^0_{i,n_i}. With (17), we study the performance difference when the service rate is changed to a new value as $\mu_{i,n_i} \in [a_{i,n_i}, b_{i,n_i}]$. Equation (17) can be rewritten as follows.

$$\eta_T|_{\mu_{i,n_i}} - \eta_T|_{\mu^0_{i,n_i}} = \pi(n_i)|_{\mu_{i,n_i}}\Big\{[\mu_{i,n_i} - \mu^0_{i,n_i}]\tilde{g}(n_i,i)|_{\mu^0_{i,n_i}} + \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)|_{\mu^0_{i,n_i}}[f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})|_{\mu_{i,n_i}} - f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})|_{\mu^0_{i,n_i}}]\Big\}. \quad (18)$$

Assuming that f is differentiable w.r.t. μ_{i,n_i}, we use the *Taylor Expansion* to represent $f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})|_{\mu_{i,n_i}}$ when $\mu_{i,n_i} \to \mu^0_{i,n_i}$. We obtain

$$f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})|_{\mu_{i,n_i}} = f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})|_{\mu^0_{i,n_i}} + \frac{df(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu_{i,n_i}}\Big|_{\mu^0_{i,n_i}} (\mu_{i,n_i} - \mu^0_{i,n_i}) + \frac{1}{2}\frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}}\Big|_{\mu^0_{i,n_i}} (\mu_{i,n_i} - \mu^0_{i,n_i})^2 + o(|\mu_{i,n_i} - \mu^0_{i,n_i}|^2). \quad (19)$$

Substituting (19) into (18) yields

$$\eta_T|_{\mu_{i,n_i}} - \eta_T|_{\mu^0_{i,n_i}} = \pi(n_i)|_{\mu_{i,n_i}}\Big\{[\mu_{i,n_i} - \mu^0_{i,n_i}]\tilde{g}(n_i,i)|_{\mu^0_{i,n_i}} + \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)|_{\mu^0_{i,n_i}}\Big[\frac{df(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu_{i,n_i}}\Big|_{\mu^0_{i,n_i}} (\mu_{i,n_i} - \mu^0_{i,n_i}) + \frac{1}{2}\frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}}\Big|_{\mu^0_{i,n_i}} (\mu_{i,n_i} - \mu^0_{i,n_i})^2 + o(|\mu_{i,n_i} - \mu^0_{i,n_i}|^2)\Big]\Big\}. \quad (20)$$

Now, we study the first-order derivative of η_T w.r.t. μ_{i,n_i}. Taking the derivative operation w.r.t. μ_{i,n_i} on both sides of (20), we obtain the following equation (we assume that η_T, π, and f are differentiable, which is true for common cases.)

$$\frac{d\eta_T}{d\mu_{i,n_i}}\Big|_{\mu_{i,n_i}} = \pi(n_i)|_{\mu_{i,n_i}}\Big\{\tilde{g}(n_i,i)|_{\mu^0_{i,n_i}} + \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)|_{\mu^0_{i,n_i}} \Big[\frac{df(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu_{i,n_i}}\Big|_{\mu^0_{i,n_i}} + \frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}}\Big|_{\mu^0_{i,n_i}} (\mu_{i,n_i} - \mu^0_{i,n_i}) + o(|\mu_{i,n_i} - \mu^0_{i,n_i}|)\Big]\Big\} + \frac{d\pi(n_i)}{d\mu_{i,n_i}}\Big|_{\mu_{i,n_i}} \Big\{[\mu_{i,n_i} - \mu^0_{i,n_i}]\tilde{g}(n_i,i)|_{\mu^0_{i,n_i}} + \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)|_{\mu^0_{i,n_i}}\Big[\frac{df(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu_{i,n_i}}\Big|_{\mu^0_{i,n_i}} (\mu_{i,n_i} - \mu^0_{i,n_i}) + \frac{1}{2}\frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}}\Big|_{\mu^0_{i,n_i}} (\mu_{i,n_i} - \mu^0_{i,n_i})^2 + o(|\mu_{i,n_i} - \mu^0_{i,n_i}|^2)\Big]\Big\}. \quad (21)$$

In the above analysis, *note that* $\tilde{g}(n_i,i)$ *and* $\pi(\boldsymbol{n}|n_i)$ *are independent of* μ_{i,n_i}*, so their derivatives are both* 0. Let $\mu_{i,n_i} \to \mu^0_{i,n_i}$, ignore the infinitesimal terms including $\mu_{i,n_i} - \mu^0_{i,n_i}$ and (21) becomes

$$\frac{d\eta_T}{d\mu_{i,n_i}}\Big|_{\mu^0_{i,n_i}} = \pi(n_i)|_{\mu^0_{i,n_i}}\Big\{\tilde{g}(n_i,i)|_{\mu^0_{i,n_i}} + \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)|_{\mu^0_{i,n_i}} \frac{df(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu_{i,n_i}}\Big|_{\mu^0_{i,n_i}}\Big\}. \quad (22)$$

This is the first-order derivative at the current service rate μ^0_{i,n_i}. Furthermore, we study the second-order derivative. Taking the derivative operation again on both sides of (21) and letting $\mu_{i,n_i} \to \mu^0_{i,n_i}$, we omit the infinitesimal terms and obtain the following second-order derivative

$$\frac{d^2\eta_T}{d\mu^2_{i,n_i}}\Big|_{\mu^0_{i,n_i}} = \pi(n_i)|_{\mu^0_{i,n_i}} \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)|_{\mu^0_{i,n_i}} \frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}}\Big|_{\mu^0_{i,n_i}} + 2 \cdot \frac{d\pi(n_i)}{d\mu_{i,n_i}}\Big|_{\mu^0_{i,n_i}} \Big\{\tilde{g}(n_i,i)|_{\mu^0_{i,n_i}} + \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)|_{\mu^0_{i,n_i}} \frac{df(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu_{i,n_i}}\Big|_{\mu^0_{i,n_i}}\Big\}. \quad (23)$$

Therefore, we have obtained the first-order and second-order derivatives at the current point μ^0_{i,n_i} and their formulas are (22) and (23), respectively. Further using the difference equation (18), we derive the following theorems.

Theorem 1: For a particular server i and n_i, if $\frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}} \equiv 0$ for all $\boldsymbol{n} \in \mathcal{S}_{n_i}$ and $\mu_{i,n_i} \in [a_{i,n_i}, b_{i,n_i}]$, the system performance is monotonic w.r.t. μ_{i,n_i} and the optimal μ^*_{i,n_i} can be either a_{i,n_i} or b_{i,n_i}.

Proof: Condition $\frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}} \equiv 0$ means that the cost function is linear to service rate μ_{i,n_i} and can be represented as $f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}}) = \phi_1(\boldsymbol{n},\vec{\mu}^{(-i)}_{\boldsymbol{n}}) + \mu_{i,n_i} \cdot \phi_2(\boldsymbol{n},\vec{\mu}^{(-i)}_{\boldsymbol{n}})$, where $\vec{\mu}^{(-i)}_{\boldsymbol{n}} := (\mu_{1,n_1}, \cdots, \mu_{i-1,n_{i-1}}, \mu_{i+1,n_{i+1}}, \cdots, \mu_{M,n_M})$, ϕ_1 and ϕ_2 are any functions. Therefore, the difference equation (18) can be rewritten as follows.

$$\eta_T|_{\mu_{i,n_i}} - \eta_T|_{\mu^0_{i,n_i}} = \pi(n_i)|_{\mu_{i,n_i}}[\mu_{i,n_i} - \mu^0_{i,n_i}] \Big\{\tilde{g}(n_i,i)|_{\mu^0_{i,n_i}} + \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)|_{\mu^0_{i,n_i}} \phi_2(\boldsymbol{n},\vec{\mu}^{(-i)}_{\boldsymbol{n}})\Big\}. \quad (24)$$

From (13), we know that $\pi(\boldsymbol{n}|n_i)$ is constant when only μ^0_{i,n_i} is varying. Thus, we define

$$\begin{aligned}\alpha &:= \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)|_{\mu^0_{i,n_i}} \phi_2(\boldsymbol{n}, \vec{\mu}^{(-i)}_{\boldsymbol{n}}) \\ &= \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)\phi_2(\boldsymbol{n}, \vec{\mu}^{(-i)}_{\boldsymbol{n}}), \end{aligned} \tag{25}$$

where α is a constant if we only change μ_{i,n_i} and fix other service rates. Therefore, the difference equation can be rewritten as

$$\eta_T|_{\mu_{i,n_i}} - \eta_T|_{\mu^0_{i,n_i}} = \pi(n_i)|_{\mu_{i,n_i}}[\mu_{i,n_i} - \mu^0_{i,n_i}] \left\{\tilde{g}(n_i,i)|_{\mu^0_{i,n_i}} + \alpha\right\}. \tag{26}$$

Exchanging the parameters μ_{i,n_i} and μ^0_{i,n_i}, (26) can be written as

$$\eta_T|_{\mu^0_{i,n_i}} - \eta_T|_{\mu_{i,n_i}} = \pi(n_i)|_{\mu^0_{i,n_i}}[\mu^0_{i,n_i} - \mu_{i,n_i}] \left\{\tilde{g}(n_i,i)|_{\mu_{i,n_i}} + \alpha\right\}. \tag{27}$$

Since the system process $\boldsymbol{n}(t)$ is ergodic, $\pi(n_i)$ is always positive. Comparing (26) with (27), it is notable that the sign of $\tilde{g}(n_i,i)|_{\mu^0_{i,n_i}} + \alpha$ and $\tilde{g}(n_i,i)|_{\mu_{i,n_i}} + \alpha$ must be the same. That is, the sign of $\tilde{g}(n_i,i)|_{\mu_{i,n_i}} + \alpha$ is fixed whatever μ_{i,n_i} is. Since $\frac{df(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu_{i,n_i}} = \phi_2(\boldsymbol{n}, \vec{\mu}^{(-i)}_{\boldsymbol{n}})$, the first-order derivative (22) can be further rewritten as follows.

$$\left.\frac{d\eta_T}{d\mu_{i,n_i}}\right|_{\mu^0_{i,n_i}} = \pi(n_i)|_{\mu^0_{i,n_i}} \left\{\tilde{g}(n_i,i)|_{\mu^0_{i,n_i}} + \alpha\right\}. \tag{28}$$

Therefore, the sign of the first-order derivative is also fixed if we only change the value of μ_{i,n_i}. The performance η_T is monotonic w.r.t. the service rate μ_{i,n_i} and the optimal μ^*_{i,n_i} can be either a_{i,n_i} or b_{i,n_i}. ■

The condition in Theorem 1 means that f is linear to the service rate. One of the typical examples is $f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}}) = \phi(\boldsymbol{n}) + \sum_{i=1}^{M} \mu_{i,n_i}$ or $f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}}) = \phi(\boldsymbol{n}) + \prod_{i=1}^{M} \mu_{i,n_i}$, where $\phi(\boldsymbol{n})$ can be any function.

Theorem 2: For a particular server i and n_i, if $\frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}} > 0$ for all $\boldsymbol{n} \in \mathcal{S}_{n_i}$ and $\mu_{i,n_i} \in [a_{i,n_i}, b_{i,n_i}]$, the optimal μ^*_{i,n_i} for $\max\{\eta_T\}$ can be either a_{i,n_i} or b_{i,n_i}.

Proof: For a differentiable η_T, the optimal μ^*_{i,n_i} must either make $d\eta_T/d\mu_{i,n_i} = 0$ or be on the boundary of the value domain of μ_{i,n_i}. Below, we prove that the first situation, μ_{i,n_i} making $d\eta_T/d\mu_{i,n_i} = 0$, cannot obtain $\max\{\eta_T\}$.

Suppose $d\eta_T/d\mu_{i,n_i} = 0$ when the service rate is μ^0_{i,n_i}. Then, with (22), we have

$$\pi(n_i)|_{\mu^0_{i,n_i}} \left\{\tilde{g}(n_i,i)|_{\mu^0_{i,n_i}} + \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)|_{\mu^0_{i,n_i}} \left.\frac{df(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu_{i,n_i}}\right|_{\mu^0_{i,n_i}}\right\} = 0. \tag{29}$$

Since $\pi(n_i)$ is always positive, (29) means that the term in the big bracket of (29) should equal 0. Substituting this conclusion into (23), we obtain

$$\left.\frac{d^2\eta_T}{d\mu^2_{i,n_i}}\right|_{\mu^0_{i,n_i}} = \pi(n_i)|_{\mu^0_{i,n_i}} \sum_{\boldsymbol{n}\in\mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i)|_{\mu^0_{i,n_i}} \left.\frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}}\right|_{\mu^0_{i,n_i}}. \tag{30}$$

Since $\frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}}$ is always positive as a given condition and $\pi(n_i)$ and $\pi(\boldsymbol{n}|n_i)$ are always positive, $\frac{d^2\eta_T}{d\mu^2_{i,n_i}}$ is also positive at any point μ^0_{i,n_i} which makes $d\eta_T/d\mu_{i,n_i} = 0$. We observe that such μ^0_{i,n_i} makes η_T get the local minimum. We can further prove that such μ^0_{i,n_i} is unique and it makes η_T obtain its global minimum. Therefore, μ^0_{i,n_i} is not the optimal μ^*_{i,n_i} for $\max\{\eta_T\}$ and μ^*_{i,n_i} must be on the boundary, i.e., a_{i,n_i} or b_{i,n_i}. ■

We further observe that a cost function with $\frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}} > 0$ for every $\boldsymbol{n}$ and μ_{i,n_i} must be a *strictly convex* function w.r.t. μ_{i,n_i}, but the converse does not hold. One of the typical functions is $f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}}) = \phi(\boldsymbol{n}) + \sum_{i=1}^{M} (\mu_{i,n_i})^k$, where $k > 1$ or $k < 0$. Many other functions which are common in practice, such as exponential function $f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}}) = \phi(\boldsymbol{n}) + \sum_{i=1}^{M} e^{\mu_{i,n_i}}$, also satisfy this condition.

Theorem 3: For a particular server i and n_i, if $\frac{d^2 f(\boldsymbol{n},\vec{\mu}_{\boldsymbol{n}})}{d\mu^2_{i,n_i}} < 0$ for all $\boldsymbol{n} \in \mathcal{S}_{n_i}$ and $\mu_{i,n_i} \in [a_{i,n_i}, b_{i,n_i}]$, the optimal μ^*_{i,n_i} for $\min\{\eta_T\}$ can be either a_{i,n_i} or b_{i,n_i}.

Proof: The proof is similar to that of Theorem 2 with the notice that $\min\{\eta_T\} = -\max\{-\eta_T\}$ with $-f$. The detailed proof is omitted. We also conclude that μ^0_{i,n_i} letting $d\eta_T/d\mu_{i,n_i} = 0$ is unique and it makes η_T obtain its global maximum. ■

We find that the cost function satisfying the condition in Theorem 3 should be *strictly concave* w.r.t. μ_{i,n_i}. One of the typical functions is $f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}}) = \phi(\boldsymbol{n}) + \sum_{i=1}^{M} (\mu_{i,n_i})^k$, where $0 < k < 1$.

It is worth mentioning that there are many other functions which satisfy the conditions in Theorem 2 or 3 to make the Max-Min optimality valid, besides what we have listed above. To the best of our knowledge, the most general results in the existing literature only derive Theorem 1 [8], cannot obtain Theorem 2 and 3. Our results are more general than all of the previous studies.

For a special case where the minimal service rates are 0, i.e., $a_{i,n_i} = 0$, we further show that the optimal service rates may hold a threshold form. This is described by the following theorem.

Theorem 4: If $a_{i,n_i} = 0$ for all $i = 1, 2, \cdots, M$ and $n_i = 1, 2, \cdots, N$, the optimal service rates in Theorem 1, 2, and 3 may hold a threshold form. That is, when $n_i < \theta_i$, $\mu^*_{i,n_i} = 0$; otherwise, $\mu^*_{i,n_i} = b_{i,n_i}$, where θ_i is a constant and called the threshold of server i, $i = 1, 2, \cdots, M$.

Proof: The proof is similar to that of [8], here we give a brief explanation to make this paper self-contained. Suppose that for a particular μ_{i,n_i}, the optimal service rate is $\mu^*_{i,n_i} = 0$. We find that all the states in the set $\bigcup_{n'_i \le n_i} \mathcal{S}_{n'_i}$ are transient states, because n_i will definitely increase and never go down after the system state reaches $\mathcal{S}_{n_i}$. Since the transient states have no effect on the system average performance, the optimal service rates μ^*_{i,n'_i} with $n'_i \le n_i$ may obviously be chosen as 0 and the total optimal policy for service rates holds a threshold form. ■

The threshold-type policy is very common in the practice, because it is easy to be employed in the real life operation. The optimality of threshold-type policies is of great significance for

practical applications. The similar cases can be found in other research fields, such as the well known (S, s) policy in the inventory management problem [10]. The optimality of those threshold-type policies is usually proved from the optimality equation and the structure of value functions in Markov models. Although we can also obtain the optimality equation for this service rate control problem, the form of optimality equations is very complicated and it is hard to obtain enough information to prove the properties of optimal policy. Here, we prove it based on the difference equation, which explores a new and effective way to study the performance optimization of queueing systems.

Until now, we have proved our main results for the load-dependent service rates under the time-average performance criterion. Below, we briefly prove that all the above results are still valid for both the load-dependent and state-dependent service rates under both the time-average and customer-average performance criteria.

For the load-dependent service rates with customer-average performance criterion, the difference equation (5) can be rewritten as follows.

$$\eta'_C - \eta_C = \eta'_I \pi'(n_i) \Big\{ \frac{\mu_{i,n_i} - \mu'_{i,n_i}}{\mu_{i,n_i}} \sum_{\boldsymbol{n} \in \mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i) c^{(f)}(\boldsymbol{n}, i) + \sum_{\boldsymbol{n} \in \mathcal{S}_{n_i}} \pi(\boldsymbol{n}|n_i) [f(\boldsymbol{n}, \vec{\mu}'_{\boldsymbol{n}}) - f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}})] \Big\}. \quad (31)$$

The first-order and second-order derivatives of η_C w.r.t. μ_{i,n_i} can be also derived with a similar analysis as (22) and (23). During the analysis, note that η'_I and $\pi'(n_i)$ are always positive. The details are omitted for the limit of space. With a similar proof as those for the previous theorems, we obtain the following corollary.

Corollary 1: The analogous results to Theorem 1, 2, 3, and 4 are also valid for the load-dependent service rates under customer-average performance criterion.

Furthermore, we discuss the optimization problem for state-dependent service rates under both the time-average and customer-average criteria. Each state-dependent service rate has a value domain, i.e., $\mu_{i,\boldsymbol{n}} \in [a_{i,\boldsymbol{n}}, b_{i,\boldsymbol{n}}]$, $i = 1, 2, \cdots, M$, $\boldsymbol{n} \in \mathcal{S}$. Suppose that a particular service rate $\mu_{i,\boldsymbol{n}}$ is changed to $\mu'_{i,\boldsymbol{n}}$. We similarly analyze the change of the associated infinitesimal generator B. For the time-average performance criterion, (7) can be rewritten as follows.

$$\eta'_T - \eta_T = \pi'(\boldsymbol{n}) \Big\{ [\mu'_{i,\boldsymbol{n}} - \mu_{i,\boldsymbol{n}}] \sum_{j=1}^{M} q_{ij} [g(\boldsymbol{n}_{ij}) - g(\boldsymbol{n})] + [f(\boldsymbol{n}, \vec{\mu}'_{\boldsymbol{n}}) - f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}})] \Big\}, \quad (32)$$

where $\vec{\mu}_{\boldsymbol{n}} := (\mu_{1,\boldsymbol{n}}, \mu_{2,\boldsymbol{n}}, \cdots, \mu_{M,\boldsymbol{n}})$ and $\vec{\mu}'_{\boldsymbol{n}} := (\mu_{1,\boldsymbol{n}}, \cdots, \mu'_{i,\boldsymbol{n}}, \cdots, \mu_{M,\boldsymbol{n}})$ with a little abuse of notations.

For the customer-average performance criterion, (5) can be rewritten as follows.

$$\eta'_C - \eta_C = \eta'_I \pi'(\boldsymbol{n}) \Big\{ \frac{\mu_{i,\boldsymbol{n}} - \mu'_{i,\boldsymbol{n}}}{\mu_{i,\boldsymbol{n}}} c^{(f)}(\boldsymbol{n}, i) + [f(\boldsymbol{n}, \vec{\mu}'_{\boldsymbol{n}}) - f(\boldsymbol{n}, \vec{\mu}_{\boldsymbol{n}})] \Big\}. \quad (33)$$

Based on (32) and (33), we can also derive the first-order and second-order derivatives of system performance w.r.t. service rates. Similarly, the Max-Min optimality of service rates can be also proved. For the limit of space, we omit the details. However, the threshold-type optimality similar to Theorem 4 cannot be guaranteed. This is because a zero state-dependent service rate cannot induce a transient state, which is essential for the proof of Theorem 4. Therefore, we obtain the following corollary and it completes the main results of this paper.

Corollary 2: The analogous results to Theorem 1, 2, and 3 are also valid for the state-dependent service rates under both the time-average and customer-average performance criteria.

IV. Brief Conclusion

In this paper, we extend the Max-Min optimality of service rate control in closed Jackson networks to a much more general form of cost functions. This result is derived based on the difference equation for Markov systems. With this Max-Min optimality, optimization algorithms can focus on only the maximal and minimal values of service rates and it greatly reduces the optimization complexity.

Acknowledgment

The authors would like to thank Dr. X.-R. Cao, the associate editor, and three anonymous reviewers for the constructive comments on this technical note.

References

[1] C. G. Cassandras and S. Lafortune, *Introduction to Discrete Event Systems, 2^{nd} Edition*, New York: Springer Verlag, 2008.

[2] X. R. Cao, *Realization Probabilities – The Dynamics of Queueing Systems*, New York: Springer Verlag, 1994.

[3] X. R. Cao, *Stochastic Learning and Optimization – A Sensitivity-Based Approach*, New York: Springer, 2007.

[4] X. R. Cao and H. F. Chen, "Perturbation realization, potentials, and sensitivity analysis of Markov processes," *IEEE Transactions on Automatic Control*, Vol. 42, pp. 1382-1393, 1997.

[5] P. Glasserman, *Gradient Estimation via Perturbation Analysis*, Boston: Kluwer Academic Publishers, 1991.

[6] Y. C. Ho and X. R. Cao, *Perturbation Analysis of Discrete Event Systems*, Norwell: Kluwer Academic Publishers, 1991.

[7] A. Lazar, "Optimal flow control of a class of queueing networks in equilibrium," *IEEE Transactions on Automatic Control*, Vol. 28, pp. 1001-1007, 1983.

[8] D. J. Ma and X. R. Cao, "A direct approach to decentralized control of service rates in a closed Jackson network," *IEEE Transactions on Automatic Control*, Vol. 39, pp. 1460-1463, 1994.

[9] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, New York: John Wiley & Sons, 1994.

[10] H. Scarf, "The optimality of (S, s) policies in the dynamic inventory problem," *Mathematical Methods in the Social Sciences*, Stanford University Press, Stanford, California, CA, pp. 196-202, 1960.

[11] S. Stidham and R. Weber, "A survey of Markov decision models for control of networks of queues," *Queueing Systems*, Vol. 13, pp. 291-314, 1993.

[12] R. Weber and S. Stidham, "Optimal control of service rates in networks of queues," *Advances in Applied Probability*, Vol. 19, pp. 202-218, 1987.

[13] L. Xia, X. Chen, and X. R. Cao, "Policy iteration for customer-average performance optimization of closed queueing systems," *Automatica*, Vol. 45, pp. 1639-1648, 2009.

[14] L. Xia, M. Xie, W. Yin, and J. Dong, "MAX-MIN optimality of service rates in queueing systems with customer-average performance criterion," *Proceedings of the 2008 Winter Simulation Conference*, pp. 509-515.

[15] D. D. Yao and Z. Schechner, "Decentralized control of service rates in a closed Jackson network," *IEEE Transactions on Automatic Control*, Vol. 34, pp. 236-240, 1989.