East Asian Journal on Applied Mathematics
doi: 10.4208/eajam.231010.260711a

Vol. **1**, No. 4, pp. 379-402
November 2011

Incentive Effects of Multiple-Server Queueing Networks: The Principal-Agent Perspective

Sin-Man Choi[1,*], Ximin Huang[2], Wai-Ki Ching[3] and Min Huang[4]

[1] *Department of Industrial Engineering and Operations Research, University of California, Berkeley, US.*
[2] *College of Management, Georgia Institute of Technology, Atlanta, US.*
[3] *Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong.*
[4] *College of Information Science and Engineering, Northeastern University; State Key Laboratory of Synthetical Automation for Process Industries, (Northeastern University), Shenyang, Liaoning, 110819, China.*

Received 23 October 2010; Accepted (in revised version) 26 July 2011

Available online 23 September 2011

Abstract. A two-server service network has been studied from the principal-agent perspective. In the model, services are rendered by two independent facilities coordinated by an agency, which seeks to devise a strategy to suitably allocate customers to the facilities and to simultaneously determine compensation levels. Two possible allocation schemes were compared — viz. the common queue and separate queue schemes. The separate queue allocation scheme was shown to give more competition incentives to the independent facilities and to also induce higher service capacity. In this paper, we investigate the general case of a multiple-server queueing model, and again find that the separate queue allocation scheme creates more competition incentives for servers and induces higher service capacities. In particular, if there are no severe diseconomies associated with increasing service capacity, it gives a lower expected sojourn time in equilibrium when the compensation level is sufficiently high.

AMS subject classifications: 60K25, 68M20, 91A80

Key words: Capacity allocation, competition, incentive theory, Markovian queueing systems, Nash equilibrium, principal agent.

1. Introduction

Finding the optimal strategy and control policy for a queueing system is a traditional optimal control problem that is well studied in the literature — e.g. see [2, 11–14, 19].

*Corresponding author. *Email addresses:* kelly.smchoi@berkeley.edu (S.-M. Choi), hehe1121@gmail.com (X. Huang), wching@hku.hk (W.-K. Ching), mhuang@mail.neu.edu.cn (M. Huang)

An optimal control problem usually involves making decisions on such system parameters as service capacity, the waiting time or the sojourn time spent in the system, and the number of servers in the system under a specified cost structure (convex or concave). Service capacity is often an important competitive factor in system design — e.g. in telecommunication networks [6], data transmission systems [13], or Vendor-Managed Inventory (VMI) [3, 18] and other supply chain management [10]. In particular, current developments in supply chain management emphasize the coordination and integration of inventory and transportation logistics [4, 20]. VMI is a supply chain initiative where the distributor is responsible for all decisions regarding the selection of the retailers or agents, which creates a competitive environment for them in the market [16].

Kalai et al. [13] studied the service capacities of two servers competing for market share, assuming a Markovian queueing system. Markovian queueing systems are popular tools for modeling service systems, since they are more mathematically tractable than non-Markovian queueing systems [6, 7]. Game theory [17] is a popular and promising analytical approach [1, 5, 8, 10]. Kalai et al. [13] classified the relevant Nash equilibria into three different cases concerning the cost function and revenue per customer, with a finite waiting time and a unique symmetric equilibrium in one case. Although their model is simple, it included two important concepts. The first is the "competitive game of servers", and the second is "market share of a server in a multi-server facility". Furthermore, when the marginal cost of providing service is "high", they found there is a unique symmetric equilibrium and that the total service capacity is less than the mean demand rate. In such a case, each server actually behaves as if it were a monopolist, so there is no desirable competition. On the other hand, when the marginal cost of servicing is "low", a unique symmetric equilibrium exists and the total service capacity is greater than the mean demand rate.

In [14], a service network where a coordinating agency is responsible for satisfying the customers' total waiting and service time is studied. Two facilities (two servers) are considered, and two types of allocation policy — viz. a common queue with two servers, and two separate single-server queues. In some cases, the separate queue allocation scheme was found to have advantages over the common queue allocation scheme. In this paper, we extend the model in [13] to allow more than two servers, and are particularly interested in where the total service capacity exceeds the mean demand rate. Our analysis indicates that with multiple servers the separate queue allocation scheme gives more service incentives and induces higher service capacities. Moreover, when there are no severe diseconomies associated with increasing service capacity, the separate queue allocation scheme gives a lower expected sojourn time in equilibrium.

The remainder of this paper is structured as follows. In Section 2, we briefly review the two-server queueing system in [13] and the service system in [14]. The multiple-server common-queue model and our analysis of system performance is presented in Section 3. In Section 4, we discuss the multiple-server separate queue system and

analyse the system performance. The effect of the number of servers on the system equilibrium is discussed in Section 5. A numerical demonstration is given in Section 6 for the case of a 3-server queueing system; and then in Section 7, we compare server competition incentives to increase capacities under the two schemes and the resulting expected sojourn times. Concluding remarks and comment on further research are made in Section 8.

2. Review of the Two-Server Model

We briefly review the model studied by Gilbert and Weng [14], for a system where two independently operating servers are coordinated by one central agency. Customers arrive according to a Poisson process of rate λ, and each server i determines its own service capacity μ_i to maximize its individual profit. Service time is assumed to follow an exponential distribution with mean $1/\mu_i$, and the cost to operate at service capacity μ is $c(\mu)$. This operating cost function is assumed to be an increasing and strictly convex function (i.e. both $c'(\mu)$ and $c''(\mu)$ are positive), such as for example $c(\mu) = \mu^2$.

The goal of the coordinating agency is to maintain the expected sojourn time below a given level W at minimal cost. The coordinating agency determines a fixed amount R to compensate the servers for each unit of service rendered, and also chooses between two allocation systems — viz. the common queue and separate queue systems. The first allocates customers to a single First-In-First-Out (FIFO) queue, where any customer arriving when both servers are idle is assigned to either server with equal likelihood. The second maintains a separate queue for each server, where arriving customers are assigned such that the expected sojourn time (i.e. the total waiting and service time) is identical for each server. In the following subsections, we specifically discuss the queueing models in [13, 14].

2.1. The common queue model

The service system studied in [13] consists of two independently operating servers coordinated by one central agency. Customers arrive according to a Poisson process of rate λ, and the service times are assumed to follow the exponential distribution. Each server i operates independently and determines its own service capacity μ_i to maximize its profit. The servers share the same cost function $c(\mu)$ to operate at service capacity μ, and the coordinating agency determines a fixed amount R to compensate them for each unit of service rendered. The queueing system is a two-server FIFO queue, where a customer arriving when both servers are idle is assigned to either server with equal likelihood. No server is allowed to be idle when there is at least one customer in the queue, and a customer arriving when one server is idle and the other is busy is assigned to the idle server. Let us now briefly present the main results obtained by [13] for this two-server queueing model.

2.1.1. The market share

The market share of Server i is equal to the mean number of customers per time unit who enter service with Server i, so the fraction of all customers served by Server $i \in \{1,2\}$ is given by

$$\alpha_i(\mu_1,\mu_2) = \frac{\lambda\mu_i^2 + \mu_1\mu_2(\mu_1+\mu_2)}{\lambda(\mu_1+\mu_2)^2 + 2\mu_1\mu_2(\mu_1+\mu_2-\lambda)}. \tag{2.1}$$

2.1.2. The profit function

Under the respective market shares given by (2.1), the profit function $\pi_i^c(\mu_1,\mu_2)$ that defines the expected profit per time unit earned by Server $i \in \{1,2\}$ is

$$\pi_i^c(\mu_1,\mu_2) = \begin{cases} R\lambda\alpha_i(\mu_1,\mu_2) - c(\mu_i) & \text{if } \mu_1+\mu_2 > \lambda \\ R\mu_i - c(\mu_i) & \text{if } \mu_1+\mu_2 \le \lambda\,, \end{cases} \tag{2.2}$$

where $c(\mu)$ is the cost per time unit of providing service at capacity μ and R is the amount of compensation the server receives for each customer served.

2.1.3. The equilibrium

Kalai et al. [13] considered the situation to be a two-person strategic game, and found that finite waiting times exist at equilibrium if and only if

$$c'\left(\frac{\lambda}{2}\right) < \frac{R}{2}. \tag{2.3}$$

Moreover, if this condition is satisfied a unique equilibrium exists, where both servers select the same service capacity $\mu_c = \mu_1 = \mu_2$ such that

$$c'(\mu_c) = \frac{R\lambda^2}{2\mu_c(2\mu_c+\lambda)}.$$

2.2. The separate queue model

Gilbert and Weng [14] studied the separate queue model. To achieve the same expected sojourn time for both servers, the fraction of customers assigned to Server $i \in \{1,2\}$ is

$$\beta_i(\mu_1,\mu_2) = \frac{\mu_i - \mu_j + \lambda}{2\lambda} \quad \text{for } \mu_j - \lambda \le \mu_i \le \mu_j + \lambda\,, \tag{2.4}$$

where $j \in \{1,2\}$ and $i \ne j$. If μ_i falls outside of the bounds in (2.4), there is no possible allocation of customers to the two servers such that the expected sojourn times are equal. With $\beta_i(\mu_1,\mu_2)$ defined in (2.4), the profit for Server $i \in \{1,2\}$ is

$$\pi_i^s(\mu_1,\mu_2) = \begin{cases} R\lambda\beta_i(\mu_1,\mu_2) - c(\mu_i) & \text{if } \mu_1+\mu_2 > \lambda \\ R\mu_i - c(\mu_i) & \text{if } \mu_1+\mu_2 \le \lambda. \end{cases} \tag{2.5}$$

The following result determines the Nash equilibrium of the service capacities.

Proposition 2.1. *(Gilbert and Weng [14])*
Consider the separate queue system in which Server $i \in \{1, 2\}$ faces the profit function in (2.5).

(a) *At equilibrium, the expected sojourn time W is finite if and only if $R/2 > c'(\lambda/2)$.*

(b) *If $R/2 > c'(\lambda/2)$, there is a unique equilibrium where $\mu_1 = \mu_2 = \mu_s$ and μ_s satisfies $c'(\mu_s) = R/2$.*

For any given value of $R > 2c'(\lambda/2)$, Gilbert and Weng [14] concluded that the equilibrium service capacities are higher under the separate queue system than in the common queue system. This can be interpreted as the consequence of the more intensive competition between the servers for market share in the separate queue system. Further, Gilbert and Weng [14] compared the cost the coordinating agency incurs to maintain the expected sojourn time below a given level in the two systems. They found the separate queue allocation scheme is to be favored when there are no severe diseconomies associated with increasing service capacity. In particular, when the cost function is quadratic (i.e. $c(\mu) = a\mu^2$ where $a > 0$), the coordinating agency incurs lower costs with the separate queue than with common queue allocation.

3. The Common Queue Model with Multiple Servers

3.1. The n-server queueing system

We now extend from the two-server system in [13] and [14] to consider a multiple n-server system. It is assumed that customer arrival follows a Poisson process, and in the common queue model the arriving customers wait in a single FIFO queue if all of the servers are busy. No server is allowed to be idle when there is at least one customer in the queue; and if a customer arrives when more than one server is idle, the customer is assigned to any of the idle servers with equal probability. Once a server completes serving a customer, the first customer in the queue (if any) is assigned to that server, where each server i may choose its own service capacity μ_i and its service time follows the exponential distribution with mean $1/\mu_i$. It is assumed that the service capacity chosen is not observed by the coordinating agency, and therefore cannot be contracted. The servers are compensated by an amount R for each customer served, and each server incurs a cost of rate $c(\mu)$ at service capacity μ.

3.1.1. Market share

We first derive the market share of each server. When $\sum_{i=1}^{n} \mu_i \leq \lambda$, a steady-state probability distribution does not exist and each server receives customers at its service capacity. Otherwise, $\sum_{i=1}^{n} \mu_i > \lambda$ and all customers are served, where each server receives only a fraction of the arriving customers at a rate lower than its service capacity. The server's profit is thus affected by the fraction of all of the customers it serves —

i.e. its market share. The market share can be obtained by finding the expected value of the server's rate of receiving customers in different states of the system over the steady-state probabilities, and then dividing by the arrival rate λ. The following proposition illustrates this result.

Proposition 3.1. *(Ching et al.* [8]*)*
Suppose that $\sum_{i=1}^{n} \mu_i > \lambda$. *Then the market share* $\alpha_i(\mu_1, \mu_2, \ldots, \mu_n)$ *of Server* i *is given by*

$$\frac{\mu_i \left[\sum_{k=0}^{n-1} k! \lambda^{n-k-1} \left(\sum_{j_1<j_2<\ldots<j_k, j_p \neq i \, \forall p} \mu_{j_1} \mu_{j_2} \ldots \mu_{j_k} \right) + \lambda^{n-1} \left(\frac{\rho}{1-\rho} \right) \right]}{\sum_{k=1}^{n} k! \lambda^{n-k} \left(\sum_{j_1<j_2<\ldots<j_k} \mu_{j_1} \mu_{j_2} \ldots \mu_{j_k} \right) + \frac{\lambda^n}{1-\rho}}. \tag{3.1}$$

The following two propositions, involving the partial derivatives of the market share α_i with respect to μ_i, prove useful in characterizing the servers' decisions and determining the Nash equilibrium of the system in considering it to be an n-player strategic game.

Proposition 3.2. *(Ching et al.* [8]*)*
Suppose that $\sum_{i=1}^{n} \mu_i > \lambda$. *Then*

$$\frac{\partial \alpha_i(\mu_1, \mu_2, \ldots, \mu_n)}{\partial \mu_i} > 0. \tag{3.2}$$

Furthermore, when $\mu_i \to \infty$ *we have*

$$\frac{\partial \alpha_i(\mu_1, \ldots, \mu_n)}{\partial \mu_i} \to 0.$$

Proposition 3.3. *(Ching et al.* [8]*)*
Suppose that $\sum_{i=1}^{n} \mu_i > \lambda$. *Then*

$$\frac{\partial^2 \alpha_i(\mu_1, \mu_2, \ldots, \mu_n)}{\partial \mu_i^2} < 0. \tag{3.3}$$

Propositions 3.2 and 3.3 imply that market share α_i is increasing and concave with respect to μ_i $(i = 1, 2, \ldots, n)$.

3.2. The profit function

In deriving the server profit function, there are two cases to be considered — viz. when $\sum_{i=1}^{n} \mu_i > \lambda$, Server i receives customers at a rate of $\lambda \alpha_i(\mu_1, \mu_2, \ldots, \mu_n)$; but when $\sum_{i=1}^{n} \mu_i \leq \lambda$, Server i receives customer at a rate of μ_i. In both cases, Server

i incurs a cost of $c(\mu_i)$, so the rate of profit of Server i takes a similar form to that in [13] as follows:

$$\pi_i^c(\mu_1,\mu_2,\ldots,\mu_n)=\begin{cases} R\lambda\alpha_i(\mu_1,\mu_2,\ldots,\mu_n)-c(\mu_i) & \text{if} \quad \sum_{i=1}^{n}\mu_i>\lambda \\ R\mu_i-c(\mu_i) & \text{if} \quad \sum_{i=1}^{n}\mu_i\le\lambda. \end{cases} \tag{3.4}$$

When servers choose their service capacities, there is a tradeoff between increasing revenue and minimizing cost. From Propositions 3.2 and 3.3, it is straightforward to obtain the following proposition on the properties of the profit function π_i with respect to μ_i.

Proposition 3.4. *(Ching et al.* [8]*)*
For $i=1,2,\ldots,n$, for each fixed $\lambda>0$ and $\mu_j>0$ for $j\neq i$ the function $\pi_i^c(\mu_1,\mu_2,\ldots,\mu_n)$ is continuous and strictly concave in μ_i.

3.3. The equilibrium of the system

Servers' decisions on service capacities affect their respective profits, which we model as an n-player strategic game where each server i simultaneously chooses its service capacity μ_i to maximize its profit π_i. We then discuss the Nash equilibrium of the system. Similar to the two-server case in [13], we find there is a unique equilibrium where all servers choose the same service capacities, when the marginal cost is low enough. We first consider how the profit of Server i changes with its service capacity, when all of the other servers choose the same service capacity.

Proposition 3.5. *(Ching et al.* [8]*)*
For $\mu_c>\lambda/n$,

$$\frac{\partial}{\partial\mu_i}\alpha_i(\mu_1,\mu_2,\ldots,\mu_n)\Big|_{\mu_1=\mu_2=\ldots=\mu_n=\mu_c}=\frac{\lambda}{n^2\mu_c^2}\left[1-\frac{\lambda^{n-1}}{\sum_{k=0}^{n-1}(k+1)!\binom{n-1}{k}\lambda^{n-k-1}\mu_c^k}\right] \tag{3.5}$$

which is decreasing in μ_c. Also, we have

$$\lim_{\mu_c\to(\lambda/n)^+}\frac{\partial}{\partial\mu_i}\alpha_i(\mu_1,\mu_2,\ldots,\mu_n)\Big|_{\mu_1=\mu_2=\ldots=\mu_n=\mu_c}=\frac{n-1}{n\lambda}$$

and

$$\lim_{\mu_c\to\infty}\frac{\partial}{\partial\mu_i}\alpha_i(\mu_1,\mu_2,\ldots,\mu_n)\Big|_{\mu_1=\mu_2=\ldots=\mu_n=\mu_c}=0.$$

It is notable that for $\mu_c > \lambda/n$ Proposition 3.5 implies

$$\frac{\partial}{\partial \mu_i}\alpha_i(\mu_1, \mu_2, \ldots, \mu_n)\Big|_{\mu_1=\mu_2=\ldots=\mu_n=\mu_c} < \frac{n-1}{n\lambda}.$$

The following proposition gives the Nash equilibrium of the game, which represents the decision of the servers on their service capacities in the long-run.

Proposition 3.6. *(Ching et al.* [8]*)*
If $(n-1)R/n > c'(\lambda/n)$, there is a unique equilibrium where

$$\mu_1 = \mu_2 = \ldots = \mu_n = \mu_c \tag{3.6}$$

and μ_c is the unique solution that satisfies $\mu_c > \lambda/n$ and

$$R\lambda \frac{\partial}{\partial \mu_i}\alpha_i(\mu_1, \mu_2, \ldots, \mu_n)\Big|_{\mu_1=\mu_2=\ldots=\mu_n=\mu_c} = c'(\mu_c) \tag{3.7}$$

— i.e.

$$R\left(\frac{\lambda}{n\mu_c}\right)^2 \left[1 - \frac{\lambda^{n-1}}{\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix}\lambda^{n-k-1}\mu_c^k}\right] = c'(\mu_c). \tag{3.8}$$

If $(n-1)R/n \leq c'(\lambda/n)$, then the system has no equilibrium where the expected waiting time is finite.

This proposition indicates that, given the arrival rate of customers λ, the number of servers n and the revenue per customer R, all servers choose the same service capacity given by Equation (3.8) in the long-run if the condition

$$\frac{(n-1)R}{n} > c'\left(\frac{\lambda}{n}\right) \tag{3.9}$$

is satisfied. Proposition 3.6 is important and useful for determining the minimum value of compensation per customer R such that the system has a finite waiting time equilibrium.

4. The Separate Queueing Network Model

4.1. The n-separate-queue system

We now extend from the separate queueing system studied in [14] to consider n $M/M/1/\infty$ FIFO queues. It is assumed that customer arrival is a Poisson process. Each server i may choose its own service capacity μ_i, and service time follows an exponential

distribution with mean $1/\mu_i$. The coordinating agency allocates a fraction of the arriving customers to each of the queues such that each customer has the same expected sojourn time, independent of the server to which the customer is assigned. It is also assumed that the arrival of customers to each of the queues is a Poisson process. As before with the common queue system, the service capacity chosen is not observed by the coordinating agency and therefore cannot be contracted. The servers are compensated by an amount R for each customer served, and each of them incurs a cost $c(\mu)$ to operate at the service rate μ, where $c(.)$ is assumed to be increasing and strictly convex.

4.2. The allocation of customers

In this subsection we derive an expression for $\beta_i(\mu_1, \mu_2, \ldots, \mu_n)$, the proportion of the arriving customers allocated to Server i, such that the expected sojourn time for customers in each queue is the same. The sojourn time W_i of a customer in queue i depends on the rate of arrival to queue i, i.e. $\lambda\beta_i(\mu_1, \mu_2, \ldots, \mu_n)$ where μ_i is the service capacity of Server i. By using the standard results of an $M/M/1/\infty$ in queueing theory [6], we have

$$W_i = \frac{1}{\mu_i - \beta_i(\mu_1, \mu_2, \ldots, \mu_n)\lambda}.$$

Proposition 4.1. *If for all* $i = 1, 2, \ldots, n$

$$\frac{1}{n-1}\left(\sum_{j=1, j\neq i}^{n} \mu_j - \lambda\right) \leq \mu_i \leq \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \mu_j + \lambda\,, \tag{4.1}$$

then the proportion of arriving customers allocated to Server i *to achieve identical expected sojourn times for all servers is*

$$\beta_i(\mu_1, \mu_2, \ldots, \mu_n) = \frac{1}{n\lambda}\left[(n-1)\mu_i - \sum_{j=1, j\neq i}^{n} \mu_j + \lambda\right]. \tag{4.2}$$

Proof. To achieve the same expected sojourn time for the n servers, we must have $W_1 = W_2 = \ldots = W_n$ — i.e.

$$\frac{1}{\mu_1 - \beta_1(\mu_1, \mu_2, \ldots, \mu_n)\lambda} = \frac{1}{\mu_2 - \beta_2(\mu_1, \mu_2, \ldots, \mu_n)\lambda} = \ldots = \frac{1}{\mu_n - \beta_n(\mu_1, \mu_2, \ldots, \mu_n)\lambda}$$

or $\mu_1 - \beta_1(\mu_1, \mu_2, \ldots, \mu_n)\lambda = \mu_2 - \beta_2(\mu_1, \mu_2, \ldots, \mu_n)\lambda = \ldots = \mu_n - \beta_n(\mu_1, \mu_2, \ldots, \mu_n)\lambda$.

Moreover, we have

$$\beta_1(\mu_1, \mu_2, \ldots, \mu_n) + \beta_2(\mu_1, \mu_2, \ldots, \mu_n) + \ldots + \beta_n(\mu_1, \mu_2, \ldots, \mu_n) = 1. \tag{4.3}$$

Rearranging and combining these equations where (4.3) is also multiplied by λ, we have

$$\begin{bmatrix} \lambda & -\lambda & 0 & \cdots & & 0 \\ \lambda & 0 & -\lambda & 0 & & 0 \\ \lambda & 0 & 0 & -\lambda & 0 & 0 \\ \vdots & & & \ddots & \ddots & \\ \lambda & 0 & 0 & 0 & 0 & -\lambda \\ \lambda & \lambda & \lambda & \cdots & \lambda & \lambda \end{bmatrix} \begin{bmatrix} \beta_1(\mu_1,\mu_2,\ldots,\mu_n) \\ \beta_2(\mu_1,\mu_2,\ldots,\mu_n) \\ \vdots \\ \beta_n(\mu_1,\mu_2,\ldots,\mu_n) \end{bmatrix} = \begin{bmatrix} \mu_1-\mu_2 \\ \mu_1-\mu_3 \\ \mu_1-\mu_4 \\ \vdots \\ \mu_1-\mu_n \\ \lambda \end{bmatrix}. \tag{4.4}$$

Multiplying this equation on the left by the inverse of the first matrix on its left-hand side then yields the solution as follows:

$$\begin{bmatrix} \beta_1(\mu_1,\mu_2,\ldots,\mu_n) \\ \beta_2(\mu_1,\mu_2,\ldots,\mu_n) \\ \vdots \\ \beta_n(\mu_1,\mu_2,\ldots,\mu_n) \end{bmatrix} = \frac{1}{n\lambda} \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ -(n-1) & 1 & \cdots & 1 & 1 \\ 1 & -(n-1) & \cdots & 1 & 1 \\ & & \ddots & & \\ 1 & 1 & \cdots & -(n-1) & 1 \end{bmatrix} \begin{bmatrix} \mu_1-\mu_2 \\ \mu_1-\mu_3 \\ \mu_1-\mu_4 \\ \vdots \\ \mu_1-\mu_n \\ \lambda \end{bmatrix}$$

— i.e.

$$\beta_i(\mu_1,\mu_2,\ldots,\mu_n) = \frac{1}{n\lambda}\left[(n-1)\mu_i - \sum_{j=1,j\neq i}^{n} \mu_j + \lambda\right]$$

for

$$\frac{\sum_{j=1,j\neq i}^{n}\mu_j - \lambda}{n-1} \leq \mu_i \leq \frac{\sum_{j=1,j\neq i}^{n}\mu_j}{n-1} + \lambda.$$

□

Here we note that it is sufficient to require only the first inequality in (4.1). Suppose we have

$$\frac{\sum_{j=1,j\neq i}^{n}\mu_j - \lambda}{n-1} \leq \mu_i \quad \text{for all } i. \tag{4.5}$$

Then for any l, on summing these inequalities over all $i \neq l$ we have

$$\frac{(n-1)\mu_l + (n-2)\sum_{j=1,j\neq l}^{n}\mu_j - (n-1)\lambda}{n-1} \leq \sum_{i=1,i\neq l}^{n}\mu_i \tag{4.6}$$

or

$$\mu_l \leq \frac{\sum_{i=1,i\neq l}^{n}\mu_i}{n-1} + \lambda. \tag{4.7}$$

Thus it is sufficient to require

$$\frac{\sum_{j=1,j\neq i}^{n}\mu_j - \lambda}{n-1} \leq \mu_i \quad \text{for all } i. \tag{4.8}$$

Proposition 4.2. *If the constraint* (4.8) *is not satisfied, then it is impossible to make the expected sojourn time of all servers equal if every server is to receive a positive fraction of the customers.*

Proof. Suppose (4.8) is not satisfied. Then for some i

$$\mu_i < \frac{\sum_{j=1,j\neq i}^{n} \mu_j - \lambda}{n-1}. \tag{4.9}$$

Let $S_0 = \{1, 2, \ldots, n\}$ and $k_0 = |S| = n$. In the $(l+1)$-th iteration, while

$$\mu_i < \frac{\sum_{j\in S_l, j\neq i}^{n} \mu_j - \lambda}{k_l - 1} \tag{4.10}$$

for some $i \in S_l$ we remove the smallest $i \in S_l$ that satisfies inequality (4.10) to form S_{l+1} and let $k_{l+1} = |S_{l+1}|$. We repeat this process until we have S_m where every $i \in S_m$ satisfies

$$\mu_i \geq \frac{\sum_{j\in S_m, j\neq i}^{n} \mu_j - \lambda}{k_m - 1}. \tag{4.11}$$

In other words, we repeatedly eliminate servers which are too slow, until it is possible to allocate customers to the remaining servers such that the expected sojourn times are equal. We further note that for $i = 0, 1, 2, \ldots, m$

$$\frac{\sum_{j\in S_i}^{n} \mu_j - \lambda}{k_i} < \frac{\sum_{j\in S_m}^{n} \mu_j - \lambda}{k_m}, \tag{4.12}$$

since we always remove the server with the smallest value of μ_j. Then for all $i \notin S$ we have

$$\mu_i < \frac{\sum_{j\in S_m}^{n} \mu_j - \lambda}{k_m}. \tag{4.13}$$

We consider only the k_m servers with indices in S_m, and allocate customers to them such that the sojourn times are the same. Since we have

$$\mu_i \geq \frac{\sum_{j\in S_m, j\neq i}^{n} \mu_j - \lambda}{k_m - 1}$$

for all $i \in S_m$, we allocate a fraction

$$\beta_i(\mu_1, \mu_2, \ldots, \mu_n) = \frac{1}{k_m \lambda}\left[(k_m - 1)\mu_i - \sum_{j\in S_m, j\neq i}^{n} \mu_j + \lambda\right]$$

of customers to Server i where $i \in S_m$. The expected sojourn time of each of these k_m servers is then

$$\frac{1}{\mu_i - [(k_m - 1)\mu_i - \sum_{j,\in S_m, j\neq i} \mu_j + \lambda]/k_m} = \frac{k_m}{\sum_{j\in S_m} \mu_j - \lambda} < \frac{1}{\mu_l}$$

for any $l \notin S_m$, where the inequality follows from (4.13). We see that for any $l \notin S_m$ the expected service time of Server l is longer than the expected sojourn time of the servers with indices in S. It is therefore impossible to equalize the expected sojourn time of Server l with the other servers, and we set $\beta_l(\mu_1, \mu_2, \ldots, \mu_n) = 0$ for any servers $l \notin S$. □

We note that the service capacities of some servers are so low that it is possible to allocate all the customers to other servers and still achieve an expected sojourn time less than the expected service time of the slower servers. Thus it is undesirable to allocate any customers to those slow servers.

Similar to the case of the common server queue, the rate of profit of Server i is

$$\pi_i^s(\mu_1, \mu_2, \ldots, \mu_n) = \begin{cases} R\lambda\beta_i(\mu_1, \mu_2, \ldots, \mu_n) - c(\mu_i) & \text{if} \quad \sum_{i=1}^{n} \mu_i > \lambda \\ R\mu_i - c(\mu_i) & \text{if} \quad \sum_{i=1}^{n} \mu_i \leq \lambda. \end{cases} \tag{4.14}$$

We model the situation as an n-player strategic game, where each Server i chooses service capacity μ_i to maximize its profit as given by (4.14). We give the following result on the equilibrium service capacities.

Proposition 4.3. *Consider the separate queue system in which Server $i \in 1, 2, \ldots, n$ faces the profit function in* (4.14).

(a) *At equilibrium, the expected sojourn time W is finite if and only if*

$$\frac{(n-1)\,R}{n} > c'\left(\frac{\lambda}{n}\right).$$

(b) *If*

$$\frac{(n-1)\,R}{n} > c'\left(\frac{\lambda}{n}\right)$$

and $c'(\mu)$ is not bounded above by $(n-1)R/n$, then there is a unique equilibrium with $\mu_1 = \mu_2 = \ldots = \mu_n = \mu_s$ where μ_s satisfies

$$c'(\mu_s) = \frac{(n-1)R}{n}. \tag{4.15}$$

Proof. One may first note that in equilibrium the condition (4.1) must be satisfied, for otherwise at least one server i with capacity $\mu_i > 0$ receives no customers. Server i can then lower its service rate, for instance to $\mu_i/2$, to increase its profit without affecting the compensation as it does not serve any customers. Thus it is impossible to have a Nash equilibrium where condition (4.1) is not satisfied, and therefore we need only consider those cases where (4.1) is satisfied. The following is a generalization of the proof for the two-server case given in [14].

(a) Suppose W is infinite, so

$$\mu_1 + \mu_2 + \ldots + \mu_n \leq \lambda \quad \text{and} \quad \pi_i^s(\mu_1, \mu_2, \ldots, \mu_n) = R\mu_i - c(\mu_i).$$

To have an equilibrium, for $i = 1, 2, \ldots, n$ we must have

$$\frac{\partial \pi_i^s(\mu_1, \mu_2, \ldots, \mu_n)}{\partial \mu_i} = R - c'(\mu_i) = 0$$

— i.e. $R = c'(\mu_i)$. Since $\mu_1 + \mu_2 + \ldots + \mu_n \leq \lambda$, we have $\mu_i \leq \lambda/n$ for some i, and hence $R = c'(\mu_i) \leq c'(\lambda/n)$ by the convexity of $c(.)$. Since $(n-1)R/n < R$, we then have $(n-1)R/n < c'(\lambda/n)$. Now suppose W is finite, so

$$\mu_1 + \mu_2 + \ldots + \mu_n > \lambda \quad \text{and} \quad \pi_i^s(\mu_1, \mu_2, \ldots, \mu_n) = R\lambda\beta_i(\mu_1, \mu_2, \ldots, \mu_n) - c(\mu_i).$$

To have an equilibrium, for $i = 1, 2, \ldots, n$ we must have

$$\frac{\partial \pi_i^s(\mu_1, \mu_2, \ldots, \mu_n)}{\partial \mu_i} = R\lambda \frac{\partial \beta_i(\mu_1, \mu_2, \ldots, \mu_n)}{\partial \mu_i} - c'(\mu_i) = 0\ .$$

Substituting the partial derivative of $\beta_i(\mu_1, \mu_2, \ldots, \mu_n)$, for $i = 1, 2, \ldots, n$ we have

$$\frac{\partial \pi_i^s(\mu_1, \mu_2, \ldots, \mu_n)}{\partial \mu_i} = \frac{(n-1)R}{n} - c'(\mu_i) = 0 \tag{4.16}$$

— i.e. $(n-1)R/n = c'(\mu_i)$. Since $\mu_1 + \mu_2 + \ldots + \mu_n > \lambda$, we have $\mu_i > \lambda/n$ for some i and hence $(n-1)R/n = c'(\mu_i) > c'(\lambda/n)$ by the convexity of $c(.)$. Therefore $(n-1)R/n > c'(\lambda/n)$.

(b) It is given that $(n-1)R/n < c'(\lambda/n)$ and $c'(\mu)$ is not bounded above by $(n-1)R/n$. Since $c'(\mu)$ is increasing, (4.15) holds for some μ_s. From part (a), if $(n-1)R/n < c'(\lambda/n)$, then W is finite and $\mu_1+\mu_2+\ldots+\mu_n > \lambda$, and the equilibrium service capacities must satisfy (4.16). Since $c(.)$ is strictly convex, we must have $\mu_1 = \mu_2 = \ldots = \mu_n = \mu_s$ where μ_s satisfies (4.16). Note that in this case condition (4.1) is satisfied. □

5. Effect of the Number of Servers

We recall that the condition for the existence of a finite waiting-time equilibrium, in both the common queue system and the separate queue system, is

$$R > \frac{n}{n-1} \cdot c'\left(\frac{\lambda}{n}\right).$$

As n increases, $(n-1)R/n$ increases and $c'(\lambda/n)$ decreases, therefore the minimum value of R for which a finite waiting-time equilibrium exists decreases as n increases. As the number of servers increases, competition becomes more intense. This decreases the cost of the coordinating agency to achieve a finite-waiting time equilibrium. Moreover, for the separate queue system, when the above condition is satisfied we have $(n-1)R/n = c'(\mu_s)$, where the left-hand side is increasing with n. Hence the equilibrium value of μ_s increases with n, since $c(.)$ is convex. In other words, a rise in the number of servers increases competition incentives and therefore induces higher service capacities.

6. A Numerical Example for a Three-Server Queueing System

In this section, we present a numerical example for a three-server service system (i.e. for $n = 3$). We assume the cost function takes the form

$$c(\mu) = \mu^2 \,, \tag{6.1}$$

and that the condition for a stable queueing system holds — i.e.

$$\mu_1 + \mu_2 + \mu_3 > \lambda. \tag{6.2}$$

We note that $c'(\mu) > 0$ and $c''(\mu) > 0$ for $\mu > 0$, so $c(\mu)$ is strictly increasing and strictly convex.

6.1. Common queueing system

For the common queueing system we have

$$\alpha_i(\mu_1, \mu_2, \mu_3) = \frac{\mu_i \left[\lambda^2 + \lambda(\mu_j + \mu_l) + 2\mu_j\mu_l + \frac{\lambda^3}{\mu_i+\mu_j+\mu_l-\lambda}\right]}{\lambda^2(\mu_i + \mu_j + \mu_l) + 2\lambda(\mu_i\mu_j + \mu_i\mu_l + \mu_j\mu_l) + 6\mu_i\mu_j\mu_l + \frac{\lambda^3(\mu_i+\mu_j+\mu_l)}{\mu_i+\mu_j+\mu_l-\lambda}},$$

where $j, l \in \{1, 2, 3\}$ and i, j, l are distinct. Now

$$\frac{\partial}{\partial\mu_i}\alpha_i(\mu_1, \mu_2, \mu_3)\bigg|_{\mu_1=\mu_2=\mu_3=\mu_c} = \frac{2\lambda(2\lambda + 3\mu_c)}{9\mu_c(\lambda^2 + 4\mu_c\lambda + 6\mu_c^2)}.$$

If $2R/3 > c'(\lambda/n) = 2\lambda/3$ so $R > \lambda$, there is a unique equilibrium where

$$\mu_1 = \mu_2 = \mu_3 = \mu_c$$

and μ_c is the unique solution satisfying

$$\mu_c > \frac{\lambda}{3} \,, \quad \text{and} \quad \left[\frac{2\lambda^2(2\lambda + 3\mu_c)}{9\mu_c(\lambda^2 + 4\mu_c\lambda + 6\mu_c^2)}\right] R = c'(\mu_c) = 2\mu_c$$

$$\text{or} \;\; 54\mu_c^4 + 36\lambda\mu_c^3 + 9\lambda^2\mu_c^2 - 3R\lambda^2\mu_c - 2R\lambda^3 = 0.$$

6.2. Separate queueing system

For the separate queueing system, if

$$\frac{\mu_j + \mu_l - \lambda}{2} \le \mu_i \le \frac{\mu_j + \mu_l}{2} + \lambda$$

for all $i, j, l \in \{1, 2, 3\}$ and i, j, l distinct, we have

$$\beta_i(\mu_1, \mu_2, \mu_3) = \frac{1}{3\lambda}\left(2\mu_i - \mu_j + \mu_l + \lambda\right) \,.$$

In equilibrium, the expected sojourn time W is finite if and only if

$$\frac{2R}{3} > c'\left(\frac{\lambda}{n}\right) = \frac{2\lambda}{3},$$

or $R > \lambda$. If this condition is satisfied, there is a unique equilibrium where

$$\mu_1 = \mu_2 = \mu_3 = \mu_s \ ,$$

with μ_s satisfying

$$c'(\mu_s) = 2\mu_s = \frac{2R}{3}, \quad \text{or} \ \ \mu_s = \frac{R}{3} \ .$$

7. Comparison of Competition Incentives in the Two Queueing Systems

In this section, we consider the results of the common and separate queueing systems to compare how the independent servers choose their service capacities in each case, given the same level of compensation R large enough for a finite-waiting time equilibrium to exist.

Proposition 7.1. *If $(n-1)R/n > c'(\lambda/n)$ for fixed R, then unique symmetric equilibria exist for both the common queue and the separate queue systems. Further, if the equilibrium service capacity in each of these two systems is denoted by μ_c and μ_s respectively, then $\mu_s > \mu_c$.*

Proof. We have

$$c'(\mu_s) = \frac{(n-1)R}{n} > R\lambda \ \frac{\partial}{\partial \mu_i}\alpha_i(\mu_1, \mu_2, \ldots, \mu_n)\Big|_{\mu_1=\mu_2=\ldots=\mu_n=\mu_c} = c'(\mu_c) \ ,$$

where the inequality follows from Proposition 3.5. Now since $c(.)$ is strictly convex, $c'(\mu_s) > c'(\mu_c)$ implies $\mu_s > \mu_c$. □

This proposition indicates that, for a given value of

$$R > \frac{n}{n-1} \cdot c'\left(\frac{\lambda}{n}\right),$$

the equilibrium service capacity commonly chosen by the n servers in the separate queue system is higher than in the common queue system. Thus the servers have more incentives to work at a higher service capacity in a separate queue system than in a common queue system. As pointed out by Gilbert and Weng [14] in the two-server case, this can be interpreted as a consequence of more intensive competition for customers in the separate queue system. In the separate queue system, an increase in service capacity increases the server's rate of receiving customers whether idle or busy. On the other hand, an increase in service capacity in the common queue system only raises the server's rate of receiving customers when all servers are busy, since customers

are allocated to idle servers with equal probability. Proposition 7.1 shows this is also true for an n-server system — i.e. competition in the separate queue system provides more incentives for servers to work at a faster rate.

Nevertheless, a higher equilibrium service capacity in the separate queue system does not always imply a lower expected sojourn time for customers. In the two-server case, Gilbert and Weng [14] showed that the expected sojourn time under the separate queue allocation policy is always lower than that under the common queue allocation policy when $c(\mu) = a\mu^2$ and $R > c'(\lambda/2)$, but this does not hold when $n > 2$. For example, in the 3-server case suppose $c(\mu) = \mu^2, \lambda = 1$ and $R = 1.001$, so that $\mu_c = 0.3460$ and $\mu_s = 0.3667$ but $W_c = 27.4826 < W_s = 30$. From standard results for the $M/M/n/\infty$ queue in queueing theory, the expected sojourn time of the single n-server queue is

$$W_c(\mu_c) = \frac{1}{\mu_c}\left[\frac{a_c^n}{(n-1)!(n-a_c)^2} \times \left(\sum_{k=0}^{n-1}\frac{a_c^k}{k!} + \frac{a_c^n}{(n-1)!(n-a_c)}\right)^{-1} + 1\right], \tag{7.1}$$

where $a_c = \lambda/\mu_c$. For each queue in the separate queue system, the arrival rate is λ/n at equilibrium, so from standard $M/M/1/\infty$ queue results the expected sojourn time is

$$W_s(\mu_s) = \frac{1}{\mu_s - \lambda/n}. \tag{7.2}$$

As with Equation (7.1), the expression for the expected sojourn time of the n-server common queue is rather complicated. In order to investigate cases where the expected sojourn time W_s under the separate queue allocation policy is lower than the time W_c under the common queue allocation policy, we give a necessary and sufficient condition for W_c to be greater than W_s in the following lemma.

Lemma 7.1. *A necessary and sufficient condition for $W_c > W_s$ is*

$$\left(1 - \frac{n\mu_c}{\lambda R}c'(\mu_c)\right) \cdot \left(\frac{\mu_s - \lambda/n}{\mu_c - \lambda/n}\right) > 1. \tag{7.3}$$

Proof. Rearranging Equation (7.1), the expected sojourn time of the separate queue system can be written as

$$W_c = \frac{1}{\mu_c}\left[\frac{\lambda}{n\mu_c - \lambda}\left(\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix}\left(\frac{\mu_c}{\lambda}\right)^k\right)^{-1} + 1\right]. \tag{7.4}$$

Also from equation (7.2), the expected sojourn time for the separate queue system is

$$W_s = \frac{1}{\mu_s - \lambda/n}\,. \tag{7.5}$$

From equations (7.4) and (7.5),

$$\begin{aligned}
\frac{W_c}{W_s} =& \frac{1}{\mu_c}\left[\frac{\lambda}{n\mu_c-\lambda}\left(\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix}(\frac{\mu_c}{\lambda})^k\right)^{-1}+1\right] / \left(\frac{1}{\mu_s-\lambda/n}\right) \\
=& \frac{\mu_c-\lambda/n}{\mu_c}\left[\frac{\lambda}{n\mu_c-\lambda}\left(\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix}(\frac{\mu_c}{\lambda})^k\right)^{-1}+1\right]\cdot\frac{\mu_s-\lambda/n}{\mu_c-\lambda/n} \\
=& \left[\frac{\lambda}{n\mu_c}\left(\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix}(\frac{\mu_c}{\lambda})^k\right)^{-1}+1-\frac{\lambda}{n\mu_c}\right]\cdot\frac{\mu_s-\lambda/n}{\mu_c-\lambda/n} \\
=& \left[1-\frac{\lambda}{n\mu_c}\left(1-\frac{1}{\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix}(\frac{\mu_c}{\lambda})^k}\right)\right]\cdot\frac{\mu_s-\lambda/n}{\mu_c-\lambda/n} \\
=& \left[1-\frac{n\mu_c}{\lambda R}c'(\mu_c)\right]\cdot\frac{\mu_s-\lambda/n}{\mu_c-\lambda/n}\,,
\end{aligned}$$

where the last equality follows from Proposition 3.6. Therefore we have

$$W_c > W_s \quad \text{if and only if} \quad \left(1-\frac{n\mu_c}{\lambda R}c'(\mu_c)\right)\cdot\frac{\mu_s-\lambda/n}{\mu_c-\lambda/n} > 1.$$

□

To apply the condition in Proposition 7.1 with given values of R, λ and cost function $c(.)$, it should be noted that the equilibrium service capacities μ_c and μ_s must first be computed. However, there is no need to calculate the values of W_c and W_s after obtaining μ_c and μ_s. If we further assume that $c'(.)$ is concave, a sufficient condition for $W_c > W_s$ can be derived from the condition in Proposition 7.1.

Proposition 7.2. *Suppose $c'(\mu)$ is concave, i.e. $c''(\mu)$ is nonincreasing. Then a sufficient condition for $W_c > W_s$ is*

$$\frac{c'(\mu_c)\mu_c}{\lambda R}\left[\frac{\lambda}{\mu_c}+(n-1)\left(1-c'\left(\frac{\lambda}{n}\right)\Big/\frac{(n-1)R}{n}\right)\right] < \frac{(n-1)}{n}. \tag{7.6}$$

When $c'(.)$ is linear, $W_c > W_s$ if and only if the condition holds.

Proof. Given that $c'(\mu)$ is increasing, together with the additional assumption that $c''(\mu)$ is non-increasing, we know that $c'(\mu)$ is concave for $\mu > 0$. In that case, since $\lambda/n < \mu_c < \mu_s$ by concavity we have

$$c'(\mu_c) \geq \left(\frac{\mu_s-\mu_c}{\mu_s-\lambda/n}\right)c'(\lambda/n)+\left(\frac{\mu_c-\lambda/n}{\mu_s-\lambda/n}\right)c'(\mu_s) \tag{7.7}$$

such that
$$\frac{c'(\mu_s) - c'(\lambda/n)}{c'(\mu_c) - c'(\lambda/n)} \leq \frac{\mu_s - \lambda/n}{\mu_c - \lambda/n}\,. \tag{7.8}$$

Equality holds when $c'(.)$ is linear. Combining with equation (7.6), we have

$$\begin{aligned}\frac{W_c}{W_s} &\geq \left[1 - \frac{n\mu_c}{\lambda R}c'(\mu_c)\right] \cdot \frac{c'(\mu_s) - c'(\lambda/n)}{c'(\mu_c) - c'(\lambda/n)} \\ &= \left[1 - \frac{n\mu_c}{\lambda R}c'(\mu_c)\right] \cdot \frac{(n-1)R/n - c'(\lambda/n)}{c'(\mu_c) - c'(\lambda/n)} \\ &= \frac{\left[(n-1)R/n - c'(\mu_c)\left(\frac{(n-1)\mu_c}{\lambda} - \frac{n\mu_c}{\lambda R}c'(\lambda/n)\right) - c'(\lambda/n)\right]}{c'(\mu_c) - c'(\lambda/n)}.\end{aligned} \tag{7.9}$$

Again, equality holds if $c'(.)$ is linear. Thus we have $W_c > W_s$ if the condition
$$c'(\mu_c) < \frac{(n-1)R}{n} - c'(\mu_c)\left(\frac{(n-1)\mu_c}{\lambda} - \frac{n\mu_c}{\lambda R}c'(\frac{\lambda}{n})\right) \tag{7.10}$$
holds, which is equivalent to condition (7.6). When $c'(.)$ is linear, this is a necessary and sufficient condition for $W_c > W_s$. □

Note that condition (7.10) does not involve the separate-queue equilibrium service capacity μ_s. The condition is used to find the effect of the compensation level R on the comparative advantage of the separate queue allocation policy over the common queue allocation policy. In doing so, we first look at the effect of R on the left-hand side of the condition — in particular, let us consider how the expression $c'(\mu_c)\mu_c/(\lambda R)$ changes with R. Rearranging equation (3.8) in Proposition 3.6, we have

$$\frac{c'(\mu_c)\mu_c}{\lambda R} = \left(\frac{\lambda}{n^2\mu_c}\right)\left[1 - \frac{\lambda^{n-1}}{\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix}\lambda^{n-k-1}\mu_c^k}\right]. \tag{7.11}$$

Recall that for any fixed value of λ, μ_c is increasing in R and thus λ/μ_c is decreasing in R. Thus how the right-hand side of equation (7.11) changes with λ/μ_c will reflect how $c'(\mu_c)\mu_c/(\lambda R)$ changes with R. The following proposition gives the result.

Proposition 7.3. *Given $0 < a_c < n$ and $n \geq 2$, the expression*

$$\frac{a_c}{n^2}\left[1 - \frac{1}{\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix}a_c^{-k}}\right]$$

is increasing in a_c.

Proof. Let us consider $z = 1/a_c$ and $C(z) = \sum_{k=0}^{n-1}(k+1)! \left(\begin{array}{c} n-1 \\ k \end{array} \right) z^k$, and show that $\frac{1}{n^2 z}(1 - \frac{1}{C(z)})$ is decreasing in z for $z > 1/n$ and $n \geq 2$. Note that

$$\frac{d}{dz}\left[\frac{1}{n^2 z}\left(1 - \frac{1}{C(z)}\right).\right] = \frac{1}{n^2}\left[\frac{C'(z)}{z[C(z)]^2} - \frac{1}{z^2}\left(1 - \frac{1}{C(z)}\right)\right] = \frac{zC'(z) + C(z) - [C(z)]^2}{n^2 z^2 [C(z)]^2}.$$

Differentiating $C(z)$ with respect to z, $C'(z) = \sum_{k=0}^{n-1}(k+1)! \left(\begin{array}{c} n-1 \\ k \end{array} \right) k z^{k-1}$, so we have

$$zC'(z) + C(z) = \sum_{k=0}^{n-1}(k+1)(k+1)! \left(\begin{array}{c} n-1 \\ k \end{array} \right) z^k \, .$$

But

$$\begin{aligned}
[C(z)]^2 &= \left[\sum_{k=0}^{n-1}(k+1)! \left(\begin{array}{c} n-1 \\ k \end{array} \right) z^k\right]^2 \\
&> \sum_{m=0}^{n-1}\sum_{k=0}^{m}(k+1)! \left(\begin{array}{c} n-1 \\ k \end{array} \right) z^k (m-k+1)! \left(\begin{array}{c} n-1 \\ m-k \end{array} \right) z^{m-k} \\
&= \sum_{m=0}^{n-1}\frac{(n-1)!}{(n-m-1)!} z^m \sum_{k=0}^{m}(k+1)(m-k+1)\frac{(n-1)!}{(n-k-1)!}\frac{(n-m-1)!}{(n-m+k-1)!} \\
&= \sum_{m=0}^{n-1}\frac{(n-1)!}{(n-m-1)!} z^m \sum_{k=0}^{m}(k+1)(m-k+1)\prod_{i=0}^{k-1}\frac{n-i-1}{n-(m-k)-i-1} \\
&\geq \sum_{m=0}^{n-1}\frac{(n-1)!}{(n-m-1)!} z^m \sum_{k=0}^{m}(k+1)(m-k+1) \\
&= \sum_{m=0}^{n-1}\frac{(n-1)!}{(n-m-1)!} z^m \sum_{k=0}^{m}(km - k^2 + m + 1) \\
&= \sum_{m=0}^{n-1}\frac{(n-1)!}{(n-m-1)!} z^m \frac{(m+1)}{6}\left[m^2 + 5m + 6\right] \\
&= \sum_{m=0}^{n-1}\frac{(n-1)!}{(n-m-1)!} z^m (m+1)[m(m-1)/6 + (m+1)] \\
&\geq \sum_{m=0}^{n-1}(m+1)(m+1)!\frac{(n-1)!}{(n-m-1)!m!} z^m = \sum_{m=0}^{n-1}(m+1)(m+1)! \left(\begin{array}{c} n-1 \\ m \end{array} \right) z^m \\
&= zC'(z) + C(z) \, ,
\end{aligned}$$

so that

$$\frac{d}{dz}\left[\frac{1}{n^2 z}\left(1 - \frac{1}{C(z)}\right)\right] < 0 \, . \quad \square$$

Now we are ready to establish the main result concerning the expected sojourn times. The following proposition shows that, when $a_c = \lambda/\mu_c$ is sufficiently low, the expected sojourn time is lower under the separate queue allocation policy than under the common queue allocation policy.

Proposition 7.4. *Suppose $c'(\mu)$ is concave, i.e. $c''(\mu)$ is a non-increasing function. Let $a_c = \lambda/\mu_c$. Also let a_l be the unique solution to*

$$\frac{a_c}{n^2}\left[1 - \frac{1}{\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix} a_c^{-k}}\right] = \frac{n-1}{n(a_c+n-1)}.$$

Then $1 < a_l < n$, and whenever $a_c < a_l$ we have $W_c > W_s$.

Proof. Let $a_c = \lambda/\mu_c$. In equilibrium $\mu_c > \lambda/n$, so that $0 < a_c < n$. From (7.11),

$$\frac{c'(\mu_c)}{a_c R} = \frac{a_c}{n^2}\left[1 - \frac{1}{\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix} a_c^{-k}}\right].$$

Now consider the equation

$$\frac{a_c}{n^2}\left[1 - \frac{1}{\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix} a_c^{-k}}\right] = \frac{n-1}{n(a_c+n-1)}.$$

From Proposition 7.3, we know that the left-hand side is increasing in a_c. On the other hand, the right-hand-side is decreasing in a_c. Moreover, when $a_c \to n^-$ the left-hand-side approaches

$$(n-1)/n^2 > (n-1)/[n(n+n-1)].$$

When $a_c = 1$, the left-hand side of this equation is less than $1/n^2$ while the right-hand side is $(n-1)/n^2 > 1/n^2$, so the equation has a unique solution a_l such that $1 < a_l < n$.

Now for any $a_c < a_l$ we have

$$\frac{c'(\mu_c)}{a_c R} = \frac{a_c}{n^2}\left[1 - \frac{1}{\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix} a_c^{-k}}\right] < \frac{(n-1)}{n(a_c+n-1)}$$

From Proposition 7.2, we have $W_c > W_s$ if

$$\frac{c'(\mu_c)}{a_c R}\left[a_c + (n-1)\left(1 - c'(\frac{\lambda}{n})\Big/ \frac{(n-1)R}{n}\right)\right] < \frac{(n-1)}{n}.$$

Thus for $a_c < a_l$ we have

$$\begin{aligned}\frac{c'(\mu_c)}{a_c R}\left[a_c + (n-1)\left(1 - c'(\frac{\lambda}{n})\Big/ \frac{(n-1)R}{n}\right)\right] &< \frac{c'(\mu_c)}{a_c R}\left[a_c + (n-1)\right] \\ &< \frac{(n-1)}{n(a_c + n - 1)}\left[a_c + (n-1)\right] , \\ &= \frac{(n-1)}{n} ,\end{aligned}$$

so condition (7.6) is satisfied and $W_c > W_s$. □

The following two propositions can easily be deduced from Proposition 7.4.

Proposition 7.5. *Suppose $c'(\mu)$ is concave, i.e. $c''(\mu)$ is non-increasing. Then for any fixed λ there exists a constant R_l such that whenever $R > R_l$ we have $W_c > W_s$.*

Proof. Let $\mu_l = \lambda/\rho$ and take

$$R_l = \frac{n^2 c'(\mu_l)}{a_l^2}\left[1 - \frac{1}{\sum_{k=0}^{n-1}(k+1)!\begin{pmatrix} n-1 \\ k \end{pmatrix} a_l^{-k}}\right]^{-1} ,$$

whence the result from Proposition 7.4 and the fact that a_c decreases with R. □

Proposition 7.5 implies that when $c'(\mu)$ is concave and R is sufficiently large, the expected sojourn time is lower under the separate queue system than under the common queue system. In other words, the stronger competition incentive effect of a separate queue system more than offsets the risk-pooling benefits of a common queue system with such cost functions when the compensation level is sufficiently high. With a higher compensation level, the equilibrium service capacity is higher and the servers tend to be idle longer. As noted earlier, by increasing its capacity a server only receives more customers when the system is busy, under the common queue allocation policy. Thus when the compensation level increases, the relative advantage of the separate queue allocation policy becomes more significant.

Proposition 7.6. *Suppose $c'(\mu)$ is concave, i.e. $c''(\mu)$ is non-increasing. Let $R_c(W)$ and $R_s(W)$ be the level of compensation required to maintain $W_c = W$ and $W_s = W$, respectively. Then for any fixed λ, there exists a constant W_l such that whenever $W < W_l$ we have $R_c(W) > R_s(W)$.*

Proof. Let $W_l = W_c(\mu_l)$, where $\mu_l = \lambda/a_l$ and $W_c(.)$ are as given in (7.1). Then for any $W < W_l$, take μ_c such that $W_c(\mu_c) = W$. Since $W_c(.)$ is decreasing, we have $\mu_c > \mu_l$ and thus $a_c < a_l$. By Proposition 7.4, with a compensation level of $R = R_c(W)$ we have $W_s < W_c = W$. Finally, because the expected sojourn time W_s decreases with R, we know that W can be achieved with a small compensation level — i.e. $R_s(W) < R_c(W)$. □

If the coordinating agency aims to maintain the expected sojourn time at a given level W with minimum cost, the separate queue system is more advantageous when W is sufficiently low. In other words, smaller permissible waiting times favor the separate queue system. This generalises the observation in [14] for the two-server case — viz. that the separate queue system has an increasing advantage of shorter customer waiting times.

8. Concluding Remarks

In the earlier study of a two-server service system [14], a necessary and sufficient condition was found for the separate queue allocation to be less costly than the common queue allocation, when the coordinating agency maintains expected sojourn times under a given level. It was concluded that, with small permissible waiting times or no severe diseconomies associated with increasing capacity, the separate queue allocation scheme is favored.

We have investigated a multiple-server queueing model. Our analysis indicates that with multiple servers the separate queue allocation scheme creates more competition incentives for servers and therefore induces higher service capacities. In particular, when there are no severe diseconomies associated with increasing service capacity, the separate queue allocation scheme gives a lower expected sojourn time in equilibrium when the permissible expected sojourn time is sufficiently low. We conclude that when the operating cost function $c'(.)$ is concave, with small permissible waiting times the separate queue allocation scheme is favored.

Of interest is whether permissible waiting times and diseconomies associated with increasing capacity have similar effects as in [14] when $c'(.)$ is strictly convex. The analysis for the multi-server model is more complicated however, as the desired service capacity of the servers cannot be expressed explicitly in terms of the given constraint on the expected sojourn time. In Propositions 7.4, 7.5 and 7.6, we require $c'(.)$ to be concave such that $c'(.)$ does not increase too rapidly — indeed more strictly, that $c''(.)$ is non-increasing. Since $c'(.)$ represents the marginal cost to increase service capacity, this can be interpreted as requiring no severe diseconomies associated with increasing service capacity. This agrees with the conclusion in the two-server case [14], that with no severe diseconomies associated with increasing service capacity the separate queue system tends to be favored. We noted the condition that $c'(.)$ be concave is sufficient but not necessary in our analysis, and a future investigation may clarify whether similar results can be established under a cost function $c(.)$ where $c'(.)$ is strictly convex.

Acknowledgments

A preliminary discussion of the multiple-server queueing model considered in this paper was presented and published in the Proceedings of the 39th International Conference on Computers and Industrial Engineering (CIE39), Troyes, France [9]. Ching is supported in part by Hong Kong RGC Grant No. 7017/07P and the HKU Strategic Research Funding on Computational Sciences. Huang is supported in part by the National Natural Science Foundation of China under Grant No.71071028, No. 71021061, No.70931001 and No.61070162, Specialized Research Fund for the Doctoral Program of Higher Education under Grant No.20070145017 and No. 20100042110025, the Fundamental Research Funds for the Central Universities under Grant No.N090504006, No. N100604021 and No. N090504003

References

[1] E. ALTMAN, *Non-zero-sum Stochastic Games in Admission, Service and Routing Control in Queueing Systems*, Queueing Syst. Theory Appl. **23** (1996), pp. 259–279.

[2] S. ANDRADOTIR, H. AYHAN AND D. DOWN, *Server Assignment Policies for Maximizing the Steady-State Throughput of Finite Queueing Systems*, Manag. Sci., **47** (2001), pp. 1421–1439.

[3] M. BEN-DAYA AND M. HARIGA, *Integrated Single Vendor Single Buyer Model with Stochastic Demand and Variable Lead Time*, Int. J. Prod. Econ., **92** (2004), pp. 75–80.

[4] F. BERNSTEIN, F. CHEN AND A. FEDERGRUEN, *Coordinating Supply Chains with Simple Pricing Schemes: The Role of Vendor-Managed Inventories*, Manag. Sci., *52* (2006), pp. 1483–1492.

[5] W. CHING, *On Convergence of Asynchronous Greedy Algorithm with Relaxation in Multiclass Queueing Environment*, IEEE Communication Letters, **3** (1999), pp. 34–36.

[6] W. CHING, *Iterative Methods for Queuing and Manufacturing Systems*, Springer Monographs in Mathematics, (2001), Springer-Verlag, London.

[7] W. CHING AND M. NG, *Markov Chains: Models, Algorithms and Applications*, International Series on Operations Research and Management Science, (2006) Springer, New York.

[8] W. CHING, S. CHOI AND M. HUANG, *Optimal Service Capacity in A Multiple-Server Queueing System: A Game Theory Approach*, J. Ind. Manag. Optim., **6**, (2010), pp. 73–102.

[9] S. CHOI, W. CHING AND M. HUANG, *Incentive Effects of Common and Separate Queues with Multiple Servers: The Principal-Agent Perspective*, Proceedings of the 39th International Conference on Computers and Industrial Engineering (CIE39), Troyes, France, 6-8, July, (2009), pp. 1261–1267.

[10] W. CHING, S. CHOI AND X. HUANG, *Inducing High Service Capacities in Outsourcing via Penalty and Competition*, to appear in Int. J. Prod. Res., (2011).

[11] C. CRABILL, D. GROSS AND M. MAGAZINE, *A Classified Bibliography of Research on Optimal Control of Queues*, Oper. Res. **25** (1977), pp. 219–232.

[12] M. EL-TAHA AND B. MADDAH, *Allocation of Service Time in a Multiserver System*, Manag. Sci., **52** (2006), pp. 623–637.

[13] E. KALAI, M. KAMIEN AND RUBINOVITCH, *Optimal Service Speeds in a Competitive Environment*, Manag. Sci. **38**(8) (1992), pp. 1154–1163.

[14] S. GILBERT AND Z. WENG, *Incentive Effects Favor Nonconsolidating Queues in a Service System: The Principal-Agent Perspective*, Manag. Sci. **44**(12) (1998), pp. 1662–1669.
[15] J. LAFFONT, AND D. MARTIMORT, *The Theory of Incentives: the Principal-agent Model*, (2002) Princeton; Oxford: Princeton University Press.
[16] B. MISHRA AND S. RAGHUNATHAN, *Retailer vs. Vendor-Managed Inventory and Brand Competition*, Manag. Sci., **50** (2004), pp. 445–457.
[17] P. MORRIES, *Introduction to Game Theory*, (1994), New York, Springer-Verlag.
[18] A. TAI AND W. CHING, *A Quantity-time-based Dispatching Policy for a VMI System*, Lecture Notes in Computer Science, Springer. **3483** (2005), pp. 342–349.
[19] J. TEGHEM, Control of the Service Process in a Queueing System. Euro. J. Oper. Res., **23** (1986), pp. 141–158.
[20] D. THOMAS, *Coordinated Supply Chain Management*, Euro. J. Oper. Res., **94** (1996), pp. 1–15.