# Incentive Effects of Common and Separate Queues with Multiple Servers: The Principal-Agent Perspective

**Sin-Man Choi**[1], **Wai-Ki Ching**[1], **Min Huang**[2]

[1] Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong. (kellyci@hkusua.hku.hk, wching@hkusua.hku.hk)

[2] College of Information Science and Engineering, Northeastern University, Shenyang, 110004, China. Key Laboratory of Integrated Automation of Process Industry, Ministry of Education, Shenyang, 110004, China. (mhuang@mail.neu.edu.cn)

## ABSTRACT

A two-server service network has been studied by Gilbert and Weng [13] from the principal-agent perspective. In the model, services are rendered by two independent facilities coordinated by an agency. The agency must devise a strategy to allocate customers to the facilities and determine the compensation. A common queue allocation scheme and separate queue allocation scheme are then compared. It has been shown that the separate queue system gives more competition incentives to the independent facilities and induces a higher service capacity. The main aim of this paper is to extend the results of the two-server queueing model to the case of multiple-server queueing model. Our analysis shows that in the case of multiple servers the separate queue allocation scheme creates more competition incentives for servers to increase their service capacities. In particular, when there are not severe diseconomies associated with increasing service capacity, the separate queue allocation scheme gives a lower expected sojourn time in equilibrium.

***Keywords:*** Capacity Allocation, Competition, Incentive Theory, Markovian Queueing Systems, Nash Equilibrium, Principal Agent.

## 1. INTRODUCTION

The study of optimal strategy and control policy for a queueing system is a traditional mathematical problem and has been well studied in the literature, see for instance [2, 10, 11, 12, 13, 18]. In an optimal control problem, it usually involves making decisions on system parameters such as the system service capacity and number of servers in the system under a specified cost structure (convex or concave). Here service capacity is an important competitive factor in the design of a system, for example, in the areas of telecommunication networks [1], data transmission systems [12] and Vendor-Managed Inventory (VMI) system [3, 17]. In particular, the current development in supply chain management emphasizes the coordination and integration of inventory and transportation logistics [4, 19]. VMI is a supply chain initiative where the distributor is responsible for all decisions regarding the selection of retailers or agents. This creates a competitive environment for the agents and retailers to compete in the market [15].

Regarding the service capacity, Kalai et al [12] studied a strategic game of two servers competing for their market shares through determining their service capacities. A Markovian queueing system of two servers is used in their model and analysis. Markovian queueing systems are popular tools for modeling service systems as they are mathematically tractable [6, 7]. The problem is then analyzed using game theory [16]. Game theory is a popular and promising approach [1, 5] for the captured problem. They classified the Nash equilibria into three different cases concerning the cost function and the revenue per customer. The waiting time is finite in one of these cases and there is a unique symmetric equilibrium. Although their model is simple, it brings in two important concepts. The first one is the "competitive game of servers" and the second one is the "market share of a server in a multi-server facility". Furthermore, they also report that when the marginal cost of providing service is "high", then there is a unique symmetric equilibrium and 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. Competition therefore has no effect and this leads to an undesirable situation. On the other hand, when the marginal cost of providing service is "low", a unique symmetric equilibrium exists and the total service capacity is greater than the mean demand rate.

In [13], a service network in which a coordinating agency is responsible for satisfying customers in total waiting and service time is studied. They consider a network of two facilities (two servers) with two types of allocation policy: a common queue with two servers and two separate single-server queues. They conclude that in some cases the separate queue allocation scheme has advantages over the common queue allocation scheme though. Here we will extend the model in [12] by allowing the number of servers to be more than two. In particular, we are interested in the case when the total service capacity is greater than the mean demand rate. Our analysis indicates that in the case of multiple servers, the separate queue allocation scheme also gives more incentives to servers and induces higher service capacities. Moreover, when there are not severe diseconomies associated with increasing service capacity, the separate queue allocation scheme gives a lower expected sojourn time in equilibrium.

The remainder of the paper is structured as follows. In Section 2, we will give a brief review on the two-server queueing system discussed in [12] and the service system in [13]. We present the multiple-server common-queue model and our analysis on the system performance in Sec-

978-1-4244-4136-5/09/$25.00 2009 IEEE

tion 3. In Section 4, we then present the multiple-server separate queue system and give the analysis on the system performance. In Section 5, we discuss the effect of the number of servers on the system equilibrium. In Section 6, we compare the competition incentives for the servers to increase capacities in the two schemes and the resulting expected sojourn times. Finally concluding remarks are given to address further research issues in Section 7.

## 2. A REVIEW ON THE TWO-SERVER MODEL

The service system studied by Gilbert and Weng [13] consists of two independently operating servers coordinated by one central agency. Customers arrive according to a Poisson process of rate $\lambda$. Each of the server $i$ operates independently and determines its own service capacity $\mu_i$ so as to maximize its own individual profit. Its service time is then assumed to follow an exponential distribution with mean $1/\mu_i$. The cost to operate at service capacity $\mu$ is $c(\mu)$. Here the operating cost function $c(.)$ is assumed to be an increasing and strictly convex function, i.e., both $c'(\mu)$ and $c''(\mu)$ are positive and an example of such a function is $c(\mu) = \mu^2$.

The goal of the coordinating agency is to maintain the expected sojourn time below a given level $W$ with a minimal cost. The coordinating agency determines a fixed amount $R$ to compensate the servers for each unit of service rendered.

The agency also chooses between two allocation systems, namely the common queue system and the separate queue system. The first one allocates customers to a single queue, which is First-In-First-Out (FIFO). If a customer arrives when both servers are idle, he/she is assigned to either server with equal likelihood. The second allocation policy maintains a separate queue for each server, and arriving customers are assigned so that the expected sojourn time (i.e., the total waiting and service time) is identical for each server. In the followings, we give a brief review on the queueing models discussed in [12, 13].

### 2.1. The Common Queue Model

The service system studied in Kalai et al [12] consists of two independently operated servers coordinated by one central agency. Customers arrive according to a Poisson process of rate $\lambda$ and the service times are assumed to follow the exponential distribution. Each of the server $i$ operates independently and determines its own service capacity $\mu_i$ so as to maximize its profit. They share the same cost function $c(\mu)$ to operate at service capacity $\mu$. The coordinating agency then determines a fixed amount $R$ to compensate the servers for each unit of service rendered. The queueing system is a two-server FIFO queue. If a customer arrives when both servers are idle, the customer will be assigned to either server with equal likelihood. No server is allowed to be idle when at least one customer is in the system. If a customer arrives when one server is idle and the other is busy, he/she will be assigned to the idle server. We then briefly present the main results obtained in [12] concerning the two-server queueing model.

**The Market Share** Computing the market share of Server $i$ is equivalent to computing the mean number of customers per time unit that entered service with Server $i$. The fraction of all customers served by Server $i (i = 1, 2)$, is given by

$$\alpha_i(\mu_1, \mu_2) = \frac{\lambda\mu_i^2 + \mu_1\mu_2(\mu_1 + \mu_2)}{\lambda(\mu_1 + \mu_2)^2 + 2\mu_1\mu_2(\mu_1 + \mu_2 - \lambda)}. \quad (1)$$

**The Profit Function** Given the market shares of the servers, the profit function $\pi_i^c(\mu_1, \mu_2)$ of Server $i \in \{1, 2\}$, the expected profit per time unit earned by Server $i$, is then given by

$$\pi_i^c(\mu_1, \mu_2) = \begin{cases} R\lambda\alpha_i(\mu_1, \mu_2) - c(\mu_i) & \text{if } \mu_1 + \mu_2 > \lambda \\ R\mu_i - c(\mu_i) & \text{if } \mu_1 + \mu_2 \leq \lambda. \end{cases} \quad (2)$$

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

**The Equilibrium** Kalai et al. [12] considered the situation as a two-person strategic game and found that finite waiting times exist at equilibrium if and only if

$$c'\left(\frac{\lambda}{2}\right) < \frac{R}{2}. \quad (3)$$

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

$$c'(\mu_c) = \frac{R\lambda^2}{2\mu_c(2\mu_c + \lambda)}. \quad (4)$$

### 2.2. The Separate Queue Model

Gilbert and Weng [13] studied the separate queue model and obtained the following results. To achieve the same expected sojourn time for both servers, we have $\beta_i(\mu_1, \mu_2)$, the fraction of customer requests that are assigned to Server $i (i = 1, 2)$, given by

$$\beta_i(\mu_1, \mu_2) = \frac{\mu_i - \mu_j + \lambda}{2\lambda} \quad \text{for } \mu_j - \lambda \leq \mu_i \leq \mu_j + \lambda \quad (5)$$

where $j \in \{1, 2\}$ and $i \neq j$. In cases where $\mu_i$ falls outside of the bounds, there does not exist an allocation of customers for which the expected sojourn times are equal for the two servers. Using $\beta_i(\mu_1, \mu_2)$ defined in (5), we have the profit for Server $i$ as follows:

$$\pi_i^s(\mu_1, \mu_2) = \begin{cases} R\lambda\beta_i(\mu_1, \mu_2) - c(\mu_i) & \text{if } \mu_1 + \mu_2 > \lambda \\ R\mu_i - c(\mu_i) & \text{if } \mu_1 + \mu_2 \leq \lambda \end{cases} \quad (6)$$

They proved the following result for determining the Nash equilibrium of the service capacities in the separate queue system.

**Proposition 1** (Gilbert and Weng [13]). *Consider the separate queue system in which Server $i \in \{1, 2\}$ faces the profit function in (6).*

- **(a)** *At equilibrium, the expected sojourn time $W$ is finite if and only if $R/2 > c'(\lambda/2)$.*
- **(b)** *If $R/2 > c'(\lambda/2)$, then there will be a unique equilibrium in which $\mu_1 = \mu_2 = \mu_s$, where $\mu_s$ satisfies: $c'(\mu_s) = R/2$.*

Gilbert and Weng [13] then concluded that for a given value of $R > 2c'(\lambda/2)$, the equilibrium service capacities will be higher under the separate-queue system than in the common queue system. The result can be interpreted as the consequence of the more intensive competition between the servers for market share in the separate queue system. They also compared the cost that the coordinating agency incurs to maintain expected sojourn time below a given level $W$ in the two systems. It is found that cases with not severe diseconomies associated with increasing service capacity favor the separate queue allocation scheme. In particular, the coordinating agency incurs lower costs with the separate queue allocation than with the common queue allocation when the cost function is quadratic, i.e., $c(\mu) = a\mu^2, a > 0$.

## 3. The Common Queue Model with Multiple Servers

### 3.1. The $n$-server Queueing System

In this section, we extend the two-server queueing system studied in [12] and [13] to a $n$-server queueing system. The arrival process of customers is assumed to be a Poisson process. In this queueing system, arriving customers wait in a single FIFO queue if all servers are busy. No server is allowed to be idle when there is at least one customer in the queueing system. If a customer arrives when more than one server is idle, the customer is assigned to any of the idle servers with equal likelihood. Once a server completes the service of a customer, the first customer in the queue, if any, is assigned to the server. Each server $i$ may choose its own service capacity $\mu_i$, 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 of $R$ for each customer served, and each of them incurs a cost of $c(\mu)$ to operate at service capacity $\mu$.

**The Market Share** We derive the market share of each server from the steady-state distribution in [8]. We note that when $\sum_{i=1}^{n} \mu_i \leq \lambda$, the steady-state probability distribution does not exist. In this case, each server receives customers at its service capacity. Otherwise, $\sum_{i=1}^{n} \mu_i > \lambda$ and all customers will be served. Each server only receives 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 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 state of the systems, over the steady-state probabilities, then dividing by the arrival rate $\lambda$. We have the following proposition.

**Proposition 2.** *(Ching et al [8]) If $\sum_{i=1}^{n} \mu_i > \lambda$, the market share of Server $i$, $\alpha_i(\mu_1, \mu_2, \ldots, \mu_n)$ is given by*

$$\frac{\mu_i \left[ \sum_{k=0}^{n-1} k! \lambda^{n-k-1} \left( \sum_{\substack{j_1<j_2<\ldots<j_k,\\ j_p \neq i \,\forall p}} \mu_{j_1}\mu_{j_2}\cdots\mu_{j_k} \right) + \lambda^{n-1}\left(\frac{\rho}{1-\rho}\right) \right]}{\sum_{k=1}^{n} k! \lambda^{n-k} \left( \sum_{j_1<j_2<\ldots<j_k} \mu_{j_1}\mu_{j_2}\cdots\mu_{j_k} \right) + \frac{\lambda^n}{1-\rho}}. \tag{7}$$

We note that when $\mu_i \to \infty$, we have $\alpha_i(\mu_1, \mu_2, \ldots, \mu_n)$, the market share of Server $i$ $(i = 1, 2, \ldots, n)$ tend to the following limit

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

It is shown in Ching et al [8] that the market share $\alpha_i$ is increasing and concave with respect to $\mu_i$ $(i = 1, 2, \ldots, n)$. This will be useful in characterizing the servers' decisions and determining the Nash equilibrium of the system when we considered the system as a $n$-player strategic game.

### 3.2. The Profit Function

In deriving the profit function of the servers, there are two cases to be considered. When $\sum_{i=1}^{n} \mu_i > \lambda$, Server $i$ receives customers at a rate of $\lambda\alpha_i(\mu_1, \mu_2, \ldots, \mu_n)$. When

$$\sum_{i=1}^{n} \mu_i \leq \lambda,$$

Server $i$ receives customer at a rate of $\mu_i$. In both cases, Server $i$ incurs a cost of $c(\mu_i)$. Therefore, the rate of profit of Server $i$ takes a similar form as the one in [12] and is given by

$$\pi_i^c(\mu_1, \mu_2, \ldots, \mu_n) = \begin{cases} R\lambda\alpha_i(\mu_1, \mu_2, \ldots, \mu_n) - c(\mu_i) & \text{if } \sum_{i=1}^{n} \mu_i > \lambda \\ R\mu_i - c(\mu_i) & \text{if } \sum_{i=1}^{n} \mu_i \leq \lambda. \end{cases} \tag{8}$$

When servers choose their service capacities, there is a tradeoff between increasing revenue and minimizing cost. From the fact that $\alpha_i$ is increasing and concave with respect to $\mu_i$, we readily obtain the following proposition describing the properties of the profit function $\pi_i$ with respect to $\mu_i$.

**Proposition 3.** *(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 $\mu_i$.*

### 3.3. The Equilibrium of the System

Since servers' decisions of their service capacities would affect the profit of each other, we model the situation as an $n$-player strategic game, in which each server $i$ chooses its service capacity $\mu_i$ to maximize its profit $\pi_i$. Here we discuss the Nash equilibrium of the system. In our analysis, we will show that, similar to the two-server case in [12], when the marginal cost is low enough, there is a unique equilibrium, in which all servers choose the same service capacities. In the following, we will first look at how the profit of Server $i$ changes with its service capacity when all other servers choose the same service capacities.

**Proposition 4.** *For $\mu_c > \lambda/n$,*

$$\frac{\partial}{\partial \mu_i}\alpha_i(\mu_1,\mu_2,\ldots,\mu_n)\Big|_{\mu_1=\mu_2=\ldots=\mu_n=\mu_c} = \frac{\lambda}{n^2\mu_c^2}\left[1-\frac{\lambda^{n-1}}{\sum_{k=0}^{n-1}(k+1)!\binom{n-1}{k}\lambda^{n-k-1}\mu_c^k}\right] \tag{9}$$

*which is decreasing in $\mu_c$. 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 should be noted that proposition 4 implies that for $\mu_c > \lambda/n$, we have

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

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 5.** *If $(n-1)R/n > c'(\lambda/n)$ then there is a unique equilibrium where*

$$\mu_1 = \mu_2 = \ldots = \mu_n = \mu_c \tag{10}$$

*and $\mu_c$ unique solution that satisfies $\mu_c > \lambda/n$ and*

$$R\lambda\frac{\partial}{\partial \mu_i}\alpha_i(\mu_1,\mu_2,\ldots,\mu_n)\Big|_{\mu_1=\mu_2=\ldots=\mu_n=\mu_c} = c'(\mu_c), \tag{11}$$

*i.e.,*

$$R\left(\frac{\lambda}{n\mu_c}\right)^2\left[1-\frac{\lambda^{n-1}}{\sum_{k=0}^{n-1}(k+1)!\binom{n-1}{k}\lambda^{n-k-1}\mu_c^k}\right] = c'(\mu_c). \tag{12}$$

*If $(n-1)R/n \le c'(\lambda/n)$ then the system has no equilibrium in which the expected waiting time is finite.*

The proposition shows that, given the arrival rate of customer $\lambda$, the number of servers $n$ and the revenue per customer $R$, all servers will choose the same service capacity given by Equation (12) in the long run if the condition

$$\frac{(n-1)R}{n} > c'(\frac{\lambda}{n}) \tag{13}$$

is satisfied. The proposition is useful for determining the minimum value of compensation per customer $R$ for which the system will have a finite-waiting time equilibrium.

## 4. The Separate Queueing Model

### 4.1. The $n$-separate-queue System

In this subsection, we extend the separate queueing system studied by [13] to a system of $n$ $M/M/1/\infty$ FIFO queues. The arrival of customers is assumed to be a Poisson process. Again, each server $i$ may choose its own service capacity $\mu_i$, and the 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 which server he/she is being assigned to. It is assumed that the arrival of customers to each of the queues is also a Poisson process. Similar to the case of 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 of $R$ for each customer served, and each of them incurs a cost of $c(\mu)$ to operate at service rate $\mu$, where $c(.)$ is increasing and strictly convex.

### 4.2. Proportion of Customers Allocated

Let $\beta_i(\mu_1,\mu_2,\ldots,\mu_n)$ be the proportion of arriving customers allocated to Server $i$. We derive an expression for $\beta_i$ so 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. $\lambda\beta_i(\mu_1,\mu_2,\ldots,\mu_n)$ and $\mu_i$, the service capacity of Server $i$. By using standard results in an $M/M/1/\infty$ queue theory, we have $W_i = \frac{1}{\mu_i-\beta_i(\mu_1,\mu_2,\ldots,\mu_n)\lambda}$. The proof of the following proposition and the remark can be found in [9].

**Proposition 6.** *If for all $i = 1,2,\ldots,n$,*

$$\frac{1}{n-1}\left(\sum_{j=1,j\neq i}^{n}\mu_j-\lambda\right) \le \mu_i \le \frac{1}{n-1}\sum_{j=1,j\neq i}^{n}\mu_j+\lambda \tag{14}$$

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

$$\beta_i(\mu_1,\mu_2,\ldots,\mu_n) = \frac{1}{n\lambda}\left[(n-1)\mu_i - \sum_{j=1,j\neq i}^{n}\mu_j+\lambda.\right] \tag{15}$$

**Remark 1.** *If the constraint (14) is not satisfied, then it is impossible to make the expected sojourn time of all servers equal with all servers receiving a positive fraction of customers. The service capacities of some servers are low to an extent 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. It is therefore 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

$$\pi_i^s(\mu_1,\mu_2,\ldots,\mu_n) = \begin{cases} R\lambda\beta_i(\mu_1,\mu_2,\ldots,\mu_n)-c(\mu_i) & \text{if } \sum_{i=1}^{n}\mu_i > \lambda \\ R\mu_i - c(\mu_i) & \text{if } \sum_{i=1}^{n}\mu_i \le \lambda. \end{cases} \tag{16}$$

We model the situation as a $n$-player strategic game, in which each Server $i$ chooses service capacity $\mu_i$ to maximize its profit given by (16). We give the following result

on the equilibrium service capacities. The proof can be found in [9].

**Proposition 7.** *Consider the separate queue system in which Server $i \in 1, 2, \ldots, n$ faces the profit function in (16).*

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

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

**(b)** *If*

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

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

$$c'(\mu_s) = \frac{(n-1)R}{n}. \tag{17}$$

For a numerical example of a 3-server queue system, refer to [9].

## 5. The Effect of the Number of Servers

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

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

It is worth noting that 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 lowers the cost of the coordinating agency to achieve finite-waiting time equilibrium.

Moreover, for the separate queue system, when the condition above 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 $\mu_s$ increases with $n$, since $c(.)$ is convex. In other words, a rise in the number of servers increases competition incentives and induces higher service capacities.

## 6. A Comparison of Competition Incentives in the Two Systems

Combining the results of the common queue system and the separate queue system, one can compare how the independent servers choose their service capacities in the two cases given the same level of compensation $R$, when $R$ is large enough for a finite-waiting time equilibrium to exist.

**Proposition 8.** *For fixed $R$, if $(n-1)R/n > c'(\lambda/n)$, unique symmetric equilibriums exist for both the common queue system and the separate queue system. Denote the equilibrium service capacity in the two systems by $\mu_c$ and $\mu_s$ respectively, then we have $\mu_s > \mu_c$.*

*Proof.*

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

where the inequality follows from Proposition 4. Since $c(.)$ is strictly convex, $c'(\mu_s) > c'(\mu_c)$ implies $\mu_s > \mu_c$. □

This proposition shows that, for a given value of $R > \frac{n}{n-1} \cdot c'(\frac{\lambda}{n})$, the equilibrium service capacity commonly chosen by the $n$ servers in the separate queue system will be higher than that in the common queue system. In other words, the servers have more incentives to work at a higher service capacity in a separate queue system than in a common queue system.

As suggested by Gilbert and Weng [13] in the two-server case, this can be interpreted as a consequence of more intense competition among servers for customers in the separate queue system. In the separate queue system, increasing the service capacity will increase the server's rate of receiving customers both when it is idle and busy. However, in the common queue system, since customers are allocated to idle servers with equal probability, increasing service capacities only raise the server's rate of receiving customers when all servers are busy. Our proposition shows that this result is also true for an $n$-server system and competition in the separate queue system provides more incentives for servers to work at a higher service capacity.

However, a higher equilibrium service capacity in the separate queue system does not always imply a lower expected sojourn time for customers. In the following, we give a condition on $c(.)$ for which the expected sojourn time in equilibrium is always lower in the separate queue system than in the common queue system. The proof can be found in [9].

**Proposition 9.** *Suppose $c'(\mu)$ is concave, i.e. $c''(\mu)$ is non-increasing. Then for any fixed $R$, if $(n-1)R/n > c'(\lambda/n)$, unique symmetric equilibriums exists for both the common queue system and the separate queue system. Denote the expected sojourn time of the two systems by $W_c$ and $W_s$ respectively, then we have $W_c > W_s$.*

The proposition states that for any increasing and strictly convex cost function $c(.)$ with $c'(.)$ being concave, the separate queue system always yields a lower expected sojourn time than the common queue system. In other words, the stronger competition incentive effect of a separate queue system will more than offset the risk-pooling benefits of a common queue system with such cost functions. Since a rise in the level of compensation $R$ reduces both $W_c$ and $W_s$, the result also implies that it is less costly for the coordinating agency to maintain expected sojourn time below a given level in a separate queue system in these cases. When $n = 2$ and $c(\mu) = a\mu^2, a > 0$, this coincides with the results obtained in Gilbert and Weng [13].

The condition that $c'(.)$ is concave is a requirement that $c'(.)$ does not increase too rapidly, or to be precise, that

$c''(.)$ is non-increasing. As $c'(.)$ represents the marginal cost to increase service capacity, this can be interpreted as requiring that there are not severe diseconomies associated with increasing service capacity. This agrees with the conclusion in [13].

It should be noted for cost function $c(.)$ where $c'(.)$ is strictly convex, whether the separate queue system or the common queue system gives a lower expected sojourn time may depend on the level of compensation $R$.

## 7. Concluding Remarks

In this paper, we extend the analytic results and conclusion of the two-server queueing model in [13] to the case of multiple-server queueing model. Our analysis shows that in the case of multiple servers, the separate queue allocation scheme creates more competition incentives for servers and induces higher service capacities. In particular, when there are not severe diseconomies associated with increasing service capacity, the separate queue allocation scheme gives a lower expected sojourn time in equilibrium.

In the study of the service system of two servers by Gilbert and Weng [13], they obtain a necessary and sufficient condition for which the separate queue allocation is less costly than the common queue allocation for the coordinating agency to maintain expected sojourn times under a given level. In particular, they conclude that cases with small permissible waiting times or not severe diseconomies associated with increasing capacity favor the separate queue allocation scheme. In our analysis, we conclude that cases where $c'(.)$ is concave favor the separate queue allocation scheme. It is of interest whether the permissible waiting times and diseconomies associated with increasing capacity have similar effects as in [13] when $c'(.)$ is strictly convex. However, the analysis becomes more complicated as the desired service capacity of servers cannot be expressed explicitly in terms of the given constraint of the expected sojourn time. This can be further investigated in the future.

**Acknowledgment:** Ching is supported in part by Hong Kong RGC Grant No. 7017/07P and the HKU Strategic Research Funding on Computational Sciences, HKU Hung Hing Ying Physical Science Research Grant, and HKU CRCG Grants. Huang is supported in part by the National Natural Science Foundation of China (Project no. 70671020, 70721001, 70431003, 60673159), the Program for New Century Excellent Talents in University (Project no.NCET-05-0295,NCET-05-0289), Specialized Research Fund for the Doctoral Program of Higher Education (20070145017, ,20060145012).

## REFERENCES

[1] Altman, E., "Non-zero-sum Stochastic Games in Admission, Service and Routing Control in Queueing Systems", *Queueing Systems Theory Appl.*, Vol. 23, pp259-279, 1996.

[2] Andradotir, S., Ayhan, H. and Down, D., "Server Assignment Policies for Maximizing the Steady-State Throughput of Finite Queueing Systems", *Manag. Sci.*, Vol. 47, pp1421-1439, 2001.

[3] Ben-Daya, M. and Hariga, M, "Integrated Single Vendor Single Buyer Model with Stochastic Demand and Variable Lead Time", *International Journal of Production Economics*, Vol. 92, pp75-80, 2004.

[4] Bernstein, F., Chen, F. and Federgruen, A., "Coordinating Supply Chains with Simple Pricing Schemes: The Role of Vendor-Managed Inventories", *Manag, Sci.*, Vol. 52, pp1483-1492, 2006.

[5] Ching, W., "On Convergence of Asynchronous Greedy Algorithm with Relaxation in Multiclass Queueing Environment", *IEEE Communication Letters*, 3, pp34-36, 1999.

[6] Ching, W., "Iterative Methods for Queuing and Manufacturing Systems", *Springer Monographs in Mathematics, Springer-Verlag London, Ltd., London*, 2001.

[7] Ching, W. and Ng, M., "Markov Chains : Models, Algorithms and Applications", *International Series on Operations Research and Management Science, Springer, New York*, 2006.

[8] Ching, W., Choi, S. and Huang, M., "Optimal Service Capacity of Multiple Queueing Systems: A Game Theory Approach", 2008.

[9] Ching, W., Choi, S. and Huang, M., "Incentive Effects of Common and Separate Queues with Multiple Servers: The Principal-Agent Perspective", Preprint, 2009. Available at "http://hkumath.hku.hk/ wkc/cchpaper2.pdf"

[10] Crabill, C., Gross, D. and Magazine, M., "A Classified Bibliography of Research on Optimal Control of Queues", *Oper. Res.*, Vol. 25, pp219-232, 1977.

[11] El-Taha, M. and Maddah, B., "Allocation of Service Time in a Multiserver System", *Manag. Sci.*, Vol. 52, pp623-637, 2006.

[12] Kalai, E., Kamien, M. and Rubinovitch, M., "Optimal Service Speeds in a Competitive Environment", *Manag. Sci.*, Vol. 38, No. 8, pp1154-1163, 1992.

[13] Gilbert, S. and Weng, Z., "Incentive Effects Favor Nonconsolidating Queues in a Service System: The Principal-Agent Perspective", *Manag. Sci.*, Vol. 44, No. 12, pp1662-1669, 1998.

[14] Laffont, J. and Martimort, D, "The Theory of Incentives : the Principal-agent Model", *Princeton ; Oxford : Princeton University Press*, 2002.

[15] Mishra, B. and Raghunathan, S., "Retailer vs. Vendor-Managed Inventory and Brand Competition", *Manag. Sci.*, Vol. 50, pp445-457, 2004.

[16] Morries, P., "Introduction to Game Theory", *Springer-Verlag, New York*, 1994.

[17] Tai, A. and Ching, W., "A Quantity-time-based Dispatching Policy for a VMI System", *Lecture Notes in Computer Science, Springer*, 3483, pp342-349, 2005.

[18] Teghem, J., "Control of the Service Process in a Queueing System", *Euro. J. of Oper. Res.* Vol. 23, pp141-158, 1986.

[19] Thomas, D., "Coordinated Supply Chain Management", *European Journal of Operational Research*, Vol. 94, pp1-15, 1996.