The prolonged service time at non-dedicated servers in a pooling system

Yanting Chen\textsuperscript{a}, Jingui Xie\textsuperscript{b}, T aozeng Zhu\textsuperscript{c,*}

\textsuperscript{a}School of Mathematics, Hunan University, Changsha, China 410082
\textsuperscript{b}School of Management, University of Science and Technology of China, Hefei, China 230026
\textsuperscript{c}School of Computing, National University of Singapore, Singapore 117417

Abstract

In this paper, we investigate the effect of the prolonged service time at the non-dedicated servers in a pooling system on the system performance. We consider the two-server loss model with exponential interarrival and service times. We show that if the ratio of the mean service time at the dedicated server and the mean prolonged service time at the non-dedicated server exceeds a certain threshold, pooling would become unfavourable. In particular, the threshold is explicitly provided. Moreover, when the degree of the prolonged service time is pre-specified, we show that the pooling system with prolonged service time at non-dedicated servers is not preferred when the work load in the system is greater than a certain threshold.

Keywords: Prolonged service time, Pooling

1. Introduction

Empirical studies indicate that the hidden (negative) consequences exist when the customer is served at the non-dedicated server. For instance, when the patients are assigned from a ward whose designated beds are fully occupied to an available bed in a unit designated for a different service, it has been shown in \cite{1} that this ‘off-service placement’ is associated with a substantial increase in remaining hospital length of stay, i.e., a prolonged service time. Therefore, it is of interest to investigate the effect of such prolonged service time at non-dedicated servers in a pooling system on the system performance. In particular, it is desirable to have quantitative results that capture the relationship between the degree of the prolonged service time at non-dedicated servers and the system performance. This research is not only of practical importance but also provides a theoretical attempt to understand the effect of the prolonged service time at non-dedicated servers in a pooling system on the system performance. We adopt a queueing approach. To the best of our knowledge, our attempt is the first to model and analyze this newly emerged effect in pooling thoroughly.

*Corresponding author

\textit{Email addresses:} yantingchen@hnu.edu.cn (Yanting Chen), xiej@ustc.edu.cn (Jingui Xie), zhutozng@mail.ustc.edu.cn (Taozeng Zhu)
2. Problem description and model formulation

We focus on the two-server loss system, i.e., there is no waiting space in front of the servers and the arriving customer who finds a busy server would be lost immediately. The dedicated servers for Type 1 and Type 2 customers are Server 1 and Server 2 respectively. We first describe the independent system, and we assume that the arrival processes of Type 1 and Type 2 customers are two independent Poisson processes with the same parameter $\lambda$. When a Type 1 (Type 2) customer finds Server 1 (Server 2) is busy upon arriving, this customer will leave immediately. No pooling is allowed here. The service time for a Type 1 customer at Server 1 is exponential with parameter $\mu$ and the service time for a Type 2 customer at Server 2 is exponential with parameter $\mu$ as well. We assume that $\lambda > 0$ and $\mu > 0$. The ratio defined as the occupation rate $\rho = \lambda/\mu$ satisfies $\rho > 0$. The loss system is always stable, see [2]. Therefore, the occupation rate $\rho$ is not required to be less than 1 here. The arrivals and services are mutually independent. The model is depicted in Figure 1.

$\lambda$

Type 1 customers

$\mu$

Server 1

Type 2 customers

$\mu$

Server 2

Figure 1: The independent system.

Before describing the pooling system, we first explain the prolonged service time at non-dedicated servers by introducing the prolonged coefficient which is denoted by $\gamma$. The service rate of any customer at the non-dedicated server is $\gamma$ times the service rate of this customer at the dedicated server. Due to the inefficiency of dealing with non-dedicated customers, the prolonged coefficient $\gamma$ satisfies $0 \leq \gamma \leq 1$ naturally. For example, if a Type 1 customer is allowed to enter Server 2 when Server 1 is busy, then the service rate of the Type 1 customer at Server 2 is $\gamma \mu$. This characterization allows us to define the following pooling system with prolonged service time at non-dedicated servers.

We now describe the pooling system with prolonged service time at non-dedicated servers. The arrival processes of Type 1 and Type 2 customers are two independent Poisson processes with the same parameter $\lambda$. When a Type 1 customer finds Server 1 is busy upon arriving, this Type 1 customer would immediately go to Server 2, if Server 2 is idle at this moment, this Type 1 customer would receive service at Server 2, otherwise, this Type 1 customer would leave the system immediately. The behaviour of the Type 2 customer is defined similarly. The service time for a Type 1 customer at Server 1 is exponential with parameter $\mu$ and at Server 2 is exponential with parameter $\gamma \mu$. Similarly, the service time for a Type 2 customer at Server 2 is exponential with parameter $\mu$ and at Server 1 is exponential with parameter $\gamma \mu$. Recall the assumption that $\lambda > 0$ and $\mu > 0$. The arrivals and services are mutually independent. The model is depicted in Figure 2.
Due to the Poisson and the exponential assumptions, this pooling system with prolonged service time at non-dedicated servers is a continuous-time Markov chain which can be denoted by $X(t) = (i(t), j(t))$ where $i(t), j(t) \in \{0, 1, 2\}$ for $t \geq 0$. When $i(t) = 0$, the Server 1 is empty at time $t$. When $i(t) = 1$, the Server 1 is serving a Type 1 customer at $t$. When $i(t) = 2$, the Server 1 is serving a Type 2 customer at time $t$. When $j(t) = 0, 1, 2$, the Server 2 is empty, serving a Type 1 customer and serving a Type 2 customer at time $t$, respectively. We know that the continuous-time Markov chain $X(t)$ is irreducible and positive recurrent, then the stationary probabilities which are denoted by $\pi(i, j)$ where $i, j \in \{0, 1, 2\}$ exist. In particular, the stationary probabilities of $X(t)$ are displayed in the next theorem.

**Theorem 1.** The stationary probabilities of the continuous-time Markov chain $X(t)$ are

\[
\begin{align*}
\pi(0, 0) &= \frac{\gamma^3 + \gamma^2 + 2\gamma^2 \rho}{\gamma^2 (\rho + 1)^3 + \gamma^4 (\rho + 1)^2 + \rho (\rho + \gamma)^2 + 2\gamma \rho^2 (\rho + \gamma)}, \\
\pi(0, 1) &= \frac{\rho^2}{\gamma^2 + \gamma + 2\gamma \rho} \pi(0, 0), \\
\pi(0, 2) &= \frac{\rho^2 + (\gamma + 1) \rho}{\gamma + 2\rho + 1} \pi(0, 0), \\
\pi(1, 0) &= \frac{\rho^3 + (\gamma + 1) \rho^2}{\gamma + 2\rho + 1} \pi(0, 0), \\
\pi(1, 1) &= \frac{\rho^3 + \gamma \rho^2}{\gamma^2 + \gamma + 2\gamma \rho} \pi(0, 0), \\
\pi(1, 2) &= \frac{\rho^3 + (\gamma + 1) \rho^2}{\gamma + 2\rho + 1} \pi(0, 0), \\
\pi(2, 0) &= \frac{\rho^2}{\gamma^3 + \gamma^2 + 2\gamma^2 \rho} \pi(0, 0), \\
\pi(2, 1) &= \frac{\rho^3 + \gamma \rho^2}{\gamma^3 + \gamma^2 + 2\gamma^2 \rho} \pi(0, 0), \\
\pi(2, 2) &= \frac{\rho^3 + \gamma^2 \rho^2}{\gamma^3 + \gamma^2 + 2\gamma^2 \rho} \pi(0, 0),
\end{align*}
\]

where $\rho = \frac{\lambda}{\mu}$. 

---

**Figure 2:** The pooling system with prolonged service time at non-dedicated servers where $0 \leq \gamma \leq 1$. 

![Diagram](Diagram.png)
3. The blocking probabilities

The blocking probabilities are used to evaluate the performances of the independent system and the pooling system with prolonged service time at non-dedicated servers. We first demonstrate the blocking probabilities here.

In the independent system, the blocking probability (see [2]) for each type of customer, which is denoted by $P_1$, is

$$P_1 = \frac{\rho}{\rho + 1} \quad \text{where} \quad \rho = \frac{\lambda}{\mu}.$$

In the pooling system with prolonged service time at non-dedicated servers, using PASTA property (Possion Arrivals See Time Averages, see [3, 4]), the blocking probability, which is
denoted by $P_2$, for either type of customers is $\pi(1, 1) + \pi(1, 2) + \pi(2, 1) + \pi(2, 2)$, i.e., the probability when an arriving customer finds both servers are occupied. Using Theorem 1, the detailed expression for $P_2$ is

$$P_2 = \rho^2(\gamma^3 + 3\gamma^2 + \gamma^2 \rho + 2\gamma \rho + \rho)/((\gamma^2 \rho^2 + 2\gamma \rho + \gamma^3 + \rho^2 + 5\gamma^2 \rho^2 + 4\gamma^2 \rho + \gamma^2 + 2\gamma \rho^2 + 2\gamma^2 + 2\gamma \rho + \rho^3)).$$

In the rest of the analysis, we first compare the performances of the independent system and the pooling system with prolonged service time at non-dedicated servers for the fixed occupation rate $\rho$ and different prolonged coefficient $\gamma$, then we compare the performances of the independent system and the pooling system with prolonged service time at non-dedicated servers for the pre-fixed prolonged coefficient $\gamma$ where the occupation rate $\rho$ is allowed to change.

4. Condition for pooling with prolonged service time at non-dedicated servers for fixed $\rho$

In this section, we first investigate the monotonicity property of the blocking probability $P_2$ for fixed $\rho$, then we provide the condition under which the pooling system with prolonged service time at non-dedicated servers is preferred when $\rho$ is fixed.

4.1. Property of blocking probability $P_2$ for fixed $\rho$

For fixed $\rho$, we investigate the property of the blocking probability $P_2$ while $\gamma$ changes. In the next lemma, we investigate the monotonicity of $P_2$ considered as a function of $\gamma$ and the extreme values of $P_2$ when $\gamma = 0$ and $\gamma = 1$, which are denoted by $P_2^{\rho=0}$ and $P_2^{\rho=1}$ respectively.

Lemma 2. For fixed $\rho$, the blocking probability $P_2$ is monotonically decreasing in $\gamma$ for $\gamma \in [0, 1]$. Moreover, we have $P_2^{\rho=0} \geq P_1$ and $P_2^{\rho=1} \leq P_1$ where $P_1 = \frac{\rho}{\rho+1}$.

Proof. Recall that

$$P_2 = \rho^2(\gamma^3 + 3\gamma^2 + \gamma^2 \rho + 2\gamma \rho + \rho)/((\gamma^2 \rho^2 + 2\gamma \rho + \gamma^3 + \rho^2 + 5\gamma^2 \rho^2 + 4\gamma^2 \rho + \gamma^2 + 2\gamma \rho^2 + 2\gamma^2 + 2\gamma \rho + \rho^3)).$$

For fixed $\rho$, let $u(\gamma)$ and $v(\gamma)$, which are functions of $\gamma$, denote the nominator and denominator of $P_2$ respectively. Specifically, we have $u(\gamma) = \rho^2(\gamma^3 + 3\gamma^2 + \gamma^2 \rho + 2\gamma \rho + \rho)$ and $v(\gamma) = \gamma^3 \rho^2 + 2\gamma^2 \rho + \gamma^2 + \gamma^2 \rho^3 + 5\gamma^2 \rho^2 + 4\gamma^2 \rho + \gamma^2 + 2\gamma \rho^2 + 2\gamma^2 + 2\gamma \rho + \rho^3$. For fixed $\rho$, let $L(\gamma)$ denote the nominator of the first derivative of $P_2$ regarding to $\gamma$, we have

$$L(\gamma) = u'(\gamma)v(\gamma) - u(\gamma)v'(\gamma)$$

$$= -\rho^2(3\gamma^2 \rho^3 + 2\gamma^3 \rho^2 + 4\gamma^2 \rho^2 + 2\gamma^2 \rho + \gamma^3 + 8\gamma^2 \rho^3 + 5\gamma^2 \rho + 4\gamma^2 \rho + 8\gamma \rho^2 + 2\gamma^2 + 2\gamma \rho + \rho^3).$$

Because the parameter $\rho$ is positive, it can be readily verified that $L(\gamma)$ is non-positive when $\gamma \in [0, 1]$. Moreover, the square of the denominator of $P_2$, which is $v^2(\gamma)$, is positive, we conclude that for fixed $\rho$, the blocking probability $P_2$ is monotonically decreasing in $\gamma$ for $\gamma \in [0, 1]$.

When $\gamma = 0$, we have $P_2^{\rho=0} = \rho^3/\rho^3 = 1$, this indicates that $P_2^{\rho=0} \geq P_1$ where $P_1 = \frac{\rho}{\rho+1}$. When $\gamma = 1$, the inequality $P_2^{\rho=1} \leq P_1$ holds immediately using Theorem 1 in [3], which completes the proof.
This monotonicity result holds intuitively because when the prolonged coefficient $\gamma$ becomes smaller, the service time at the non-dedicated server becomes longer, which would lead to more blocking in the system.

4.2. The comparison of blocking probabilities for fixed $\rho$

For fixed $\rho$, we compare the blocking probability of the independent system, $P_1$, with the blocking probability of the pooling system with prolonged service time at non-dedicated servers, $P_2$. When $P_1 < P_2$, the independent system is preferred, when $P_1 > P_2$, the pooling system with prolonged service time at non-dedicated servers is preferred. Denote the difference of the blocking probabilities $P_1$ and $P_2$ by $g(\gamma)$, we have

$$g(\gamma) = P_1 - P_2 = \rho/(\rho + 1) - \rho^3(g^3 + 3\gamma^2 + \gamma^2\rho + 2\gamma\rho + \rho)/(\gamma^3 + 3\rho^2 + 2\gamma\rho + \rho^2) + 2\gamma\rho + \rho(\rho + 1).$$

where $g(\gamma) = \rho(\gamma^3 + 3\gamma^2 + \gamma^2\rho + 2\gamma\rho + \rho)/(\gamma^3 + 3\rho^2 + 2\gamma\rho + \rho^2)$.

To compare $P_1$ and $P_2$, it is crucial to obtain the roots of $g(\gamma) = 0$, especially the potential root(s) in $[0, 1]$. For fixed $\rho$, we now consider $g(\gamma)$, $g_a(\gamma)$ and $g_d(\gamma)$ as functions of $\gamma$ where $\gamma$ is allowed to change from 0 to 1. Recall that $\rho$ is positive and the prolonged coefficient $\gamma$ is restricted to $[0, 1]$, we conclude that $g_a(\gamma) > 0$. This indicates that it is sufficient to investigate the root(s) in $[0, 1]$ of $g_a(\gamma) = 0$ if we are interested in the root(s) in $[0, 1]$ of $g(\gamma) = 0$. Apparently, for fixed $\rho$, the nominator $g_a(\gamma)$ is a cubic function of $\gamma$. We present the roots of $g_a(\gamma) = 0$ in the next lemma.

**Lemma 3.** For fixed positive $\rho$, the equation $g_a(\gamma) = 0$ has a unique root for $\gamma \in [0, 1]$. In particular, this root is $\gamma_1 = \frac{\rho}{\rho + 1}$. Moreover, the other two roots of $g_a(\gamma) = 0$ are $\gamma_2 = -1$ and $\gamma_3 = -\rho$.

**Proof.** Recall the expression for $g_a(\gamma)$, we have

$$g_a(\gamma) = \rho(\gamma^3 + 3\gamma^2 + \gamma^2\rho + 2\gamma\rho + \rho)/(\gamma^3 + 3\rho^2 + 2\gamma\rho + \rho^2) = \rho(\gamma + 1)(\gamma + \rho)/(\rho + 1).$$

Therefore, when $\rho > 0$, the roots of $g_a(\gamma) = 0$ are $\gamma_1 = \frac{\rho}{\rho + 1}$, $\gamma_2 = -1$ and $\gamma_3 = -\rho$, which completes the proof.

Based on the unique root of $g_a(\gamma) = 0$ when $\gamma \in [0, 1]$, we are now ready to prove the theorem that characterizes the relationship between the choice of $\gamma$ and the ordering of the blocking probabilities $P_1$ and $P_2$. 

6
Theorem 4. For fixed $\rho$, we have $P_1 < P_2$ for $\gamma \in [0, \frac{\rho}{\rho+1})$ and $P_1 > P_2$ for $\gamma \in (\frac{\rho}{\rho+1}, 1]$.

Proof. We know from Lemma 3 that the unique root of $g_\rho(\gamma) = 0$ belonging to $[0, 1]$ is $\frac{\rho}{\rho+1}$. Moreover, the expression $g_\rho(\gamma)$ is positive when $\rho > 0$ and $\gamma \in [0, 1]$. Therefore, we conclude that $\gamma_1 = \frac{\rho}{\rho+1}$ is the unique root of $g(\gamma) = P_1 - P_2 = 0$ belonging to $[0, 1]$. Recall from Lemma 2 that for fixed $\rho$ the blocking probability $P_2$ is monotonically decreasing in $\gamma$ for $\gamma \in [0, 1]$ and $P_2^{\gamma=0} \geq P_1$, $P_2^{\gamma=1} \leq P_1$. Moreover, the blocking probability $P_1$ is a constant for fixed $\rho$. We conclude that $P_1 < P_2$ for $\gamma \in [0, \frac{\rho}{\rho+1})$ and $P_1 > P_2$ for $\gamma \in (\frac{\rho}{\rho+1}, 1]$ when $\rho$ is fixed.

Notice that Theorem 4 can also be proved using the expression for $g_\rho(\gamma)$ and the property that $g_\rho(\gamma)$ is positive for $\gamma \in [0, 1]$ and $\rho > 0$. From Theorem 4 we see that the threshold of the prolonged coefficient $\gamma$ for allowing pooling with prolonged service time at non-dedicated servers is precisely the blocking probability in the independent system, i.e., $\frac{\rho}{\rho+1}$. We also learn that when the system gets busier (i.e., $\rho \uparrow$), the tolerance for the prolonged service time at the non-dedicated server becomes lower as the interval for allowing pooling $(\frac{\rho}{\rho+1}, 1]$ shrinks, see Figure 3. Therefore, we conclude that the pooling with prolonged service time at non-dedicated servers is suggested if the service rate at the non-dedicated server is greater than the service rate at the dedicated server times the blocking probability in the independent system. Otherwise, we suggest the two servers to work separately.

![Diagram](image_url)  
Figure 3: The preferred service scheme for different $\rho$.

In Figure 4 when $\rho = \frac{5}{8}$ the blocking probabilities $P_1$ and $P_2$ while $\gamma$ changes are illustrated. We see that the blocking probability $P_2$ increases when $\gamma$ comes to 0. This observation holds intuitively because when $\gamma$ is small, the customer which has been assigned to the non-dedicated server would induce a very long service time, which would lead to more congestion in both servers.

5. Condition for pooling with prolonged service time at non-dedicated servers for fixed $\gamma$

Notice that if the prolonged coefficient is given, i.e., $\gamma$ is fixed, we are also able to provide the condition under which the pooling system with prolonged service time at non-dedicated servers is preferred when the occupation rate, i.e., $\rho$, of the system changes.
5.1. Property of blocking probability $P_2$ for fixed $\gamma$

We first investigate the monotonicity of the blocking probability $P_2$ for fixed $\gamma$ in the next lemma.

**Lemma 5.** For fixed $\gamma \in [0, 1]$, the blocking probability $P_2$ is monotonically increasing in $\rho$ for $\rho > 0$.

**Proof.** Recall that

$$P_2 = \rho^2 (\gamma^3 \rho + 2\gamma \rho + \rho + \gamma^3 + 3\gamma^3)(\gamma^2 \rho^3 + 2\gamma \rho^3 + \rho^3 + \gamma^3 \rho^3 + 2\gamma^3 \rho^3 + 2\gamma^2 \rho^3 + 2\gamma^3 \rho^3 + 4\gamma^2 \rho + 4\gamma^2 \rho + \gamma^3 + \gamma^3).$$

For fixed $\gamma$, let $s(\rho)$ and $t(\rho)$, which are functions of $\rho$, denote the nominator and denominator of $P_2$ respectively. Specifically, we have $s(\rho) = \rho^2 (\gamma^2 \rho + 2\gamma \rho + \rho + 3\gamma^3)$ and $t(\rho) = \gamma^2 \rho^3 + 2\gamma \rho^3 + \rho^3 + 5\gamma \rho^3 + 2\gamma^2 \rho^3 + 2\gamma^3 \rho + 4\gamma^2 \rho + \gamma^3 + \gamma^3$. For fixed $\gamma$, let $Z(\rho)$ denote the nominator of the first derivative of $P_2$ regarding to $\rho$, we have

$$Z(\rho) = s'(\rho)t(\rho) - s(\rho)t'(\rho)$$

$$= \gamma \rho (2\gamma^3 \rho^3 + 6\gamma^2 \rho^3 + 6\gamma^3 \rho + 2\rho^3 + 4\gamma^4 \rho^2 + 16\gamma^3 \rho^2 + 20\gamma^2 \rho^2 + 8\gamma^3 \rho + 13\gamma^4 \rho + 21\gamma^3 \rho + 9\gamma^2 \rho + 3\gamma \rho + 6\gamma^3 + 2\gamma^5 + 8\gamma^4).$$

Because the prolonged coefficient $\gamma$ satisfies $0 \leq \gamma \leq 1$, it can be readily verified that $Z(\rho)$ is non-negative for $\rho > 0$. Moreover, the square of the denominator of $P_2$, which is $t^2(\rho)$, is positive, we conclude that for fixed $\gamma$, the blocking probability $P_2$ is monotonically increasing in $\rho$ for $\rho > 0$.

This monotonicity result holds intuitively because when the occupation rate becomes greater, i.e., heavier work load, there would be more blocking in the system.
5.2. The comparison of blocking probabilities for fixed $\gamma$

For fixed $\gamma$, we again need to compare the blocking probabilities $P_1$ and $P_2$ when the occupation rate of the system, $\rho$, is allowed to change. We now denote the difference of the blocking probabilities $P_1$ and $P_2$ by $h(\rho)$, where

$$h(\rho) = P_1 - P_2 = \rho/(\rho + 1) - \rho^2(\gamma^2 \rho + 2\gamma \rho + \gamma^3 + 3\gamma^2)/(\gamma^2 \rho^3 + 2\gamma \rho^2 + \gamma^3 \rho^2 + 5\gamma^2 \rho^2 + 2\gamma \rho^2 + 2\gamma^3 \rho + 4\gamma^2 \rho + 4\gamma^2 \rho + \gamma^3 + 3\gamma^2)(\rho + 1)$$

where $h_0(\rho) = \rho(\gamma^2 \rho^3 + 2\gamma \rho^2 + \gamma^3 \rho^2 + 5\gamma^2 \rho^2 + 2\gamma \rho^2 + 2\gamma^3 \rho + 4\gamma^2 \rho + \gamma^3 + 3\gamma^2)(\rho + 1)$ and $h_d(\rho) = (\rho + 1)(\gamma^2 \rho^3 + 2\gamma \rho^2 + \gamma^3 \rho^2 + 5\gamma^2 \rho^2 + 2\gamma \rho^2 + 2\gamma^3 \rho + 4\gamma^2 \rho + \gamma^3 + 3\gamma^2).

To compare $P_1$ and $P_2$, it is crucial to obtain the roots of $h(\rho) = 0$, especially the potential positive root(s). For fixed $\gamma$, we now consider $h(\rho)$, $s(\rho)$, $t(\rho)$ as functions of $\rho$ where $\rho$ is positive. Recall that $\gamma$ satisfies $\gamma \in [0, 1]$, we conclude that $h_d(\rho) > 0$. This indicates that it is sufficient to investigate the positive root(s) of $h_d(\rho) = 0$ if we are interested in the positive root(s) of $h(\rho) = 0$. For fixed $\gamma$, it can be readily verified that the nominator $h_a(\rho)$ is a cubic function of $\rho$ (the coefficient for $\rho^3$ is 0), we present the roots of $h_a(\rho) = 0$ in the next lemma. Here we focus on the case where $\gamma \neq 1$ because the case $\gamma = 1$, i.e., there is no prolonged service time, has been extensively studied in [5].

**Lemma 6.** For fixed $\gamma \in [0, 1)$, the equation $h_a(\rho) = 0$ has a unique root for $\rho \in (0, \infty)$. In particular, this root is $\rho_1 = \frac{1}{1+\gamma}$. Moreover, the other two roots of $h_a(\rho) = 0$ are $\rho_2 = 0$ and $\rho_3 = -\gamma$.

**Proof.** Recall the expression for $h_a(\rho)$, we have

$$h_a(\rho) = \rho(\gamma^2 \rho^3 + 2\gamma \rho^2 + \gamma^3 \rho^2 + 5\gamma^2 \rho^2 + 2\gamma \rho^2 + 2\gamma^3 \rho + 4\gamma^2 \rho + \gamma^3 + 3\gamma^2)(\rho + 1)$$

$$(\gamma^2 - 1)\rho(\rho - \frac{1}{\gamma - 1})(\rho + \gamma).$$

Therefore, for fixed $\gamma \in [0, 1)$, the roots of $h_a(\rho) = 0$ are $\rho_1 = \frac{1}{1+\gamma}$, $\rho_2 = 0$ and $\rho_3 = -\gamma$, which completes the proof.

Based on the unique positive root of $h_a(\rho) = 0$, we are now ready to prove the theorem that characterizes the relationship between the occupation rate $\rho$ and the ordering of the blocking probabilities $P_1$ and $P_2$.

**Theorem 7.** For fix $\gamma$ satisfying $\gamma \in [0, 1)$, we have $P_1 < P_2$ for $\rho \in (\frac{1}{1+\gamma}, \infty)$ and $P_1 > P_2$ for $\rho \in (0, \frac{1}{1+\gamma})$.
Proof. We know from Lemma 6 that the unique positive root of \( h_n(\rho) = 0 \) is \( \frac{\gamma}{1 - \gamma} \). Moreover, the expression \( h_d(\gamma) \) is positive when \( \gamma \in [0, 1) \) and \( \rho > 0 \). Therefore, we conclude that \( \rho_1 = \frac{\gamma}{1 - \gamma} \) is the unique positive root of \( h(\rho) = P_1 - P_2 = 0 \). Recall that

\[
h_n(\rho) = (P_1 - P_2)h_d(\rho) = (\gamma^2 - 1)\rho(\rho - \frac{\gamma}{1 - \gamma})(\rho + \gamma),
\]

we know that \( \gamma^2 - 1 < 0, \rho > 0 \) and \( \rho + \gamma > 0 \) for \( \gamma \in [0, 1) \) and \( \rho > 0 \). Moreover, the expression \( h_d(\rho) \) is positive for \( \gamma \in [0, 1) \) and \( \rho > 0 \). Therefore, we conclude that \( P_1 < P_2 \) for \( \rho \in (\frac{\gamma}{1 - \gamma}, \infty) \) and \( P_1 > P_2 \) for \( \rho \in (0, \frac{\gamma}{1 - \gamma}) \).

From Theorem 7, we see that if the prolonged coefficient \( \gamma \) is fixed, the pooling system with prolonged service time is not encouraged when the system gets busier, see Figure 5. In Figure 6, for \( \gamma = \frac{5}{8} \), the blocking probabilities \( P_1 \) and \( P_2 \) when the occupation rate \( \rho \) increases are illustrated. We see that when the work load in the system is light, the pooling system with prolonged service time at non-dedicated servers would lead to less blocking compared with the independent system. However, when the system becomes rather busy, the blocking probability \( P_2 \) will overtake the blocking probability \( P_1 \). Hence, it is suggested to keep the services independent when the work load in the system is heavy.

![Figure 5: The preferred service scheme for different \( \gamma \).](image)

6. Conclusion and discussion

Based on the two-server loss queueing model, we have investigated the effect of the prolonged service time at non-dedicated servers in a pooling system on the system performance. In particular, we have compared the blocking probabilities in the independent system and the pooling system with prolonged service time at non-dedicated servers.

When the occupation rate \( \rho \) is fixed, we show that only when the prolonged coefficient \( \gamma \) is greater than a certain threshold, the pooling system with prolonged service time at non-dedicated servers is preferred. More precisely, this threshold is the blocking probability in the independent system. When the prolonged coefficient \( \gamma \) is fixed, we show that only when the occupation rate \( \rho \) is less than \( \frac{\gamma}{1 - \gamma} \), the pooling system with prolonged service time at non-dedicated servers is preferred. Moreover, we have also demonstrated the monotonicity properties for the blocking
The blocking probability in the pooling system with prolonged service time at non-dedicated servers when the occupation rate $\rho$ is fixed or the prolonged coefficient $\gamma$ is fixed.

The theoretical results suggest that the pooling with prolonged service time at non-dedicated server becomes unfavourable when the delay of the service at the non-dedicate server becomes too substantial or the system is too busy. In future work, it will be of interest to generalize our work to incorporating the prolonged service time at non-dedicated servers in a queueing system with buffer space, multiple servers and asymmetric partners.

Acknowledgement

Yanting Chen acknowledges support through the NSFC grant 71701066 and the Fundamental Research Funds for the Central Universities.

References

[1] H. Song, A. Tucker, R. Graue, S. Moravick, J. Yang, Capacity pooling in hospitals: The hidden consequences of off-service placement, Management Science (2019). doi:10.1287/mnsc.2019.3395
[2] L. Kleinrock, Queueing Systems, Vol. I: Theory, Wiley, New York, 1975.
[3] R. Wolff, Poisson arrivals see time averages, Operations Reserach 30 (1982) 223–231.
[4] R. Wolff, Stochastic modeling and the theory of queues, Prentice-Hall, London, 1989.
[5] D. Smith, W. Whitt, Resource sharing for efficiency in traffic systems, Bell System Technical Journal 60 (1) (1981) 39–55.