Optimization of after-sales services with spare parts consumption and repairman travel

Abstract

This paper studies the optimization of discretionary after-sales services with spare parts consumption and repairman travel. Discretionary service means that the service quality and spare parts consumption are determined by the service provider. We assume that a long service time enables the service provider to perform a thorough repair, which leads to a high repair quality and low spare parts consumption. Each service encounter comprises both repairman travel and repair service steps. We find that the dual concerns of service quality and spare parts consumption lead to a counterintuitive impact of the spare parts price: the service provider would decrease the consumption of spare parts even if selling the spare parts becomes more profitable. In addition, we show that the repairman travelling time has different impacts on the optimal service time given the maximum sojourn time constraint. In particular, as the repairman travelling time increases, the service provider increases the repair time to keep the repair quality at a high level if no sojourn time is guaranteed; however, if a maximum sojourn time is guaranteed, the service provider decreases the repair time to meet the promise.

Keywords: after-sales services; repair time; spare parts consumption; repairman travelling time; maximum sojourn time

1. Introduction

After-sales services have become an important competitive strategy for improving customer satisfaction and increasing profits (Kurata and Nam, 2013; Kurata and Nam, 2010). For example, to improve customer satisfaction, Daikin Co., a world-famous air conditioner manufacturer, offers competitive after-sales services to customers, including dedicated service specialists, a comprehensive range of spare parts, and fast response times. The after-sales service market is believed to be up to four or five times larger than the new products market. Moreover, the gross profit generated from the after-sales service and consumption of spare parts is more than three times that from the original purchase (Saccani et al., 2007). Given the enormous prospects of after-sales services and spare parts consumption, it is crucial for manufacturers to optimally design their service processes.

A typical after-sales service process includes the repair and replacement of degraded parts. The replacement process comprises the disassembly of a degraded part and installation of a new one; however, in addition to the disassembly and installation, the repair process also includes polishing, lubrication, soldering, and so on. Thus, the replacement of a degraded part takes less time than repairing it in general (Eruguz et al., 2018; Sleptchenko et al., 2019). We do not distinguish between preventive maintenance and corrective maintenance because our setting is applicable to both as long as the maintenance service involves parts repair and replacement. In practice, customers cannot self-diagnose their problems and have to rely on the service provider's advice because of the technical complexity of products. Such information asymmetry enables the service provider to offer discretionary services to customers, i.e., the repair quality and spare parts consumption are determined by the service provider.

Consider the following air conditioner repair scenario. In summer, an air conditioner repair service provider faces numerous repair requests and hires repairmen to complete the service tasks. To satisfy as many customers as possible, the repairman cannot spend too much time on each individual customer. This customer-intensive characteristic is common in practice, especially in servicing seasonal items. As a result, the repairman may have an incentive to advise the customers to replace some parts that actually can be repaired because

replacing a part is always more time-saving and profitable than repairing it. In this setting, the parts repair-replacement tradeoff determines the service provider's decisions. The above information asymmetry, together with the customer intensiveness and parts repair-replacement tradeoff, motivate us to study the equilibrium outcome concerning the repair quality, spare parts consumption, and profit of after-sales services.

On the one hand, the service quality depends on the service provider's repair effort (Nourelfath et al., 2016). Naturally, devoting more repairing effort, i.e., spending more repair time in this paper, for a customer enables the service provider to repair more degraded parts rather than replacing them, thus improving the repair quality (Sun et al., 2020; Sun et al., 2021). However, a long repair time also contributes to a long waiting time in the service system, which results in high waiting costs for customers. As a consequence, customers may not always benefit from long repair times.

On the other hand, since part replacement requires less time than repairing a part, to accommodate more customers in a limited duration, the service provider has an incentive to advise customers to consume more parts. For example, a long friction time causes brake discs in automobiles to become rugged, which damages the braking efficiency. Disassembling and polishing the brake disc can satisfy the customer's requirement, but the service provider may advise the customer to replace the brake disc. However, if the service provider devotes more effort to repairing a part, i.e., a longer repair time, more degraded parts can be repaired rather than replaced, which will reduce the likelihood of overtreatment (Paç and Veeraraghavan, 2015). The above phenomenon also happens in automobile and electric appliance repair services (Alger and Salanie, 2006).

Moreover, many service firms generally guarantee a maximum mean total sojourn time to customers in after-sales service contracts (Kurz, 2016). For example, Gome is committed to a response time of no more than 72 hours for the air conditioner cleaning service. In addition, most automobile manufacturers, such as BMW, Volkswagen, and Ford, prominently highlight the average waiting time for car maintenance and repair (Rezapour et al., 2016). It is of critical interest to investigate how the guaranteed maximum mean total sojourn time affects the above tradeoff among the repair quality, spare parts consumption, and customer

waiting cost.

In practice, the service provider assigns a repairman with the required spare parts to serve each customer. The repairman first brings the parts and travels to the customer's site, and then performs the maintenance. Hence, each service comprises two steps, namely the repairman travel and repair service. Following Tong et al. (2016), we assume that the total service duration is exponentially distributed and customer arrivals follow a Poisson distribution. We first apply the *M/M/1* queuing model to study the service provider's optimal decisions in the case of an exogenous number of repairmen. We then apply the N parallel *M/M/1* queuing model to study the case of an endogenous number of repairmen. Finally, considering that a repairman may serve the customers originally assigned to another repairman when the latter is too busy, we examine the robustness of our results in the *M/M/N* queuing model. Specifically, we address the following questions: What are the equilibrium decisions of the service provider under different queuing structures? Is making customers replace more spare parts necessarily beneficial to the service provider? How does the repairman travelling time affect the service provider's decisions? Will the results be the same if the service provider promises the maximum mean total sojourn time to customers?

Our contributions are as follows: First, different from the prior literature that studies price and service capacity decisions in common service systems, we focus on after-sales services, which are often accompanied by spare parts consumption and repairman travel in practice. Our results contribute to understanding the roles that spare parts consumption and travelling time play in the after-sales service optimization setting. Second, compared with prior studies on after-sales services in the literature, we consider the customer intensiveness and parts overtreatment features. Third, our results provide important insights into the roles of spare parts consumption and repair travel in after-sales services. For example, the dual concerns of service quality and spare parts consumption lead to the following counterintuitive result: the service provider would enhance the service quality to reduce spare parts consumption even if the spare parts become more profitable, i.e., the spare parts price increases. In addition, we show that in different settings, i.e., with and without the maximum mean total sojourn time constraint, the repair travel time impacts the optimal service time,

service quality, and spare parts consumption in different ways.

We organize the rest of the paper as follows: In Section 2 we review the related literature, identify the research gap, and position our study. In Section 3 we introduce the model setup and discuss the assumptions. In Section 4 we characterize the structural properties of the optimal strategies with and without the maximum mean total sojourn time constraint. In Section 5 we extend the models. In Section 6 we discuss the managerial implications of the analytical findings for the after-sales decision-maker. Finally, in Section 7, we conclude the paper and suggest topics for future research. For ease of exposition, we present the proofs of all the results in Appendix B.

2. Literature Review

The following three streams of literature are relevant to our study, namely after-sales services with spare parts inventory optimization, quality-speed tradeoff in queuing, and service operations with lead time consideration.

First, our research is related to studies on after-sales services with spare parts inventory optimization. Integrated maintenance and spare parts inventory optimization models have been developed. Basten and van Houtum (2014) surveyed the literature on spare parts inventory control that uses system-oriented service measures. Eruguz et al. (2018) developed a joint optimization model to address the integrated maintenance and spare parts optimization problem for a single critical component. Sleptchenko et al. (2019) considered the problem where a service engineer and a necessary replacement part have to be allocated when performing the maintenance task. They studied the joint optimization of spare parts inventory and workforce allocation in a single-site maintenance system. Qin et al. (2020) studied the maintenance contract design problem by considering that part replacement can be reduced if the repair efficiency is high. For multiple identical items subject to silent failures, Panagiotidou (2014) investigated the joint optimization of spare parts ordering and maintenance policies. Basten and Ryan (2019) studied the impact of maintenance delay flexibility on spare parts inventory optimization. Topan et al. (2018) studied the effects of imperfect advanced information on lost-sales inventory systems with the option of returning inventory. For the trade-off between repair capacity and inventory level in spare part

networks, Sleptchenko et al. (2003) developed a greedy optimization procedure to find a near-optimal combination of the inventory level and repair capacity. To the best of our knowledge, these models do not incorporate the relationships among repair quality, spare parts consumption, repairman travel and system congestion from system design perspective.

Our paper is also related to the literature on the quality-speed tradeoff that involves pricing and capacity decisions in queuing systems. Anand et al. (2011) investigated the equilibrium service price and service rate, where the expected waiting cost and service quality are both dependent on the service rate. Li et al. (2016) extended their work by considering two competing service providers and the bounded rationally of customers. Kostami and Rajagopalan (2014) considered a dynamic setting where the service price and speed decisions in the current period would impact the potential customer arrival rate in the next period. Dai et al. (2016) investigated the impact of the medical insurance structure on a hospital's service price and speed decisions concerning imaging testing. Tong et al. (2016) studied a service system with two consecutive steps that have shared resources. They showed how the base step service speed affects the second step service speed and other equilibrium outcomes. The most relevant work to our model is that of Wang et al. (2019), who studied the speed-quality tradeoff in health care service with the consumption of medical goods. Differing from the extant research, we not only consider spare parts consumption but also the repairman travel. Hence, the whole after-sales service process in our context corresponds to a two-step system. Furthermore, we investigate how the repairman travelling time affects the equilibrium outcomes in different settings.

Finally, our research is related to studies on service operations with lead time consideration. Most of the research in this field focuses on lead time differentiation strategies. Jayaswal and Jewkes (2016) studied a duopoly market in which customers can be segmented as price- or time-sensitive. They investigated how competition influences firms' price and lead time differentiation under different operations strategies. Zhao et al. (2012) developed models to investigate the optimal service offering strategies, i.e., the uniform quotation mode and differentiated quotation mode, under specified conditions. The cannibalization between high-quality products and low-quality products can be lowered by service differentiation;

however, the capacity requirement is also improved. Jain and Bala (2018) studied this tradeoff. Nguyen and Wright (2015) studied the optimal lead time and capacity by taking into account not only how customers will respond to the lead time, but also whether the firm's capacity could meet its commitment. Our study differs from theirs in that we take into account the interplay between repair efficiency and spare parts consumption, both of which depend on repair time.

3. Model Setup

We begin this section by introducing the key notation of our model in Table 1. We use air conditioner after-sales services for illustration purposes, but the results are also applicable to other maintenance services.

Table 1. Notation

Notation	Description
Decision variables	
τ	Repair time
P	Service price
N	Number of servers
Parameters	
l	Repairman travelling time
V_b	Benchmark service value
τ_b	Benchmark service time
α	Service quality improvement rate when service time increases
β	Rate of reduction in spare parts consumption when service time increases
θ	Cost of adding a server
$k(K)$	Unit cost of ordinary (expensive) parts
ϕ	Probability that a replaced part is expensive
$\Lambda(\lambda)$	Potential (effective) arrival rate
c_w	Customers' waiting cost per unit time
δ	Share of revenue from spare parts consumption

d	Maximum mean total sojourn time
Functions	
W	Expected waiting time in queue
U	Customers' net expected utility
R	Profit of the service provider

Notes: Superscripts *ns* and *s* denote the scenarios without and with the maximum total sojourn time constraint, respectively; subscripts *1* and *N* denote the one-server queue and multi-server queue, respectively.

Consider an air conditioner maintenance provider facing a stream of requests from homogenous and rational customers. The service requests arrive at the system according to a Poisson process at an arrival rate Λ. All the customers independently determine whether to use the service and the aggregate arrival rate is λ $(\leq \Lambda)$. We consider the setting in which a long repair time with thorough maintenance and repair can increase customer's perceived service quality and reduce the spare parts consumption. In the following we introduce the key elements of our model.

Service queuing. The service provider must assign a repairman with the necessary parts for each service request. Moreover, the requests that are geographically adjacent are usually assigned to the same repairman, who takes the parts to the customer region and finishes the services in sequence (Legnani and Cavalieri, 2012). Hence, each maintenance service comprises two steps: the repairman travel step (Step 1) with average time l and repair service step (Step 2) with average time τ. For analytical tractability, we assume that the total duration of the two-step service follows the exponential distribution, but the individual time of each step could be arbitrarily distributed with average time l and τ respectively. In the related literature on customer-intensive services, researchers commonly use the exponentially distributed service time, e.g., Anand et al. (2011), Li et al. (2016), and Tong et al. (2016). Together with the Poisson arrivals, the service process corresponds to an *M/M/1* queue or *N* parallel *M/M/1* queues. The service duration consists of two parts, namely the repairman travelling time l (Step 1) and repair time τ (Step 2). We assume that l depends on the transport and is exogenous. In addition, the delays due to traffic jams and rush hours are also reflected in the travelling time. We do not involve the diagnostic process in our model

because the diagnosis can be performed through customers' descriptions online or via telephone calls. For example, the repairmen at Daikin Co. can learn about the customers' needs quickly through the company's 24-hour maintenance hotline. Hence, the timeline of the two-step service is shown in Figure 1.

Please insert Figure 1 here.

Figure 1. Illustration of the service process

Customers' perceived service value. For customer-intensive services, the quality of the service to a customer increases with the service time (Anand et al., 2011). In after-sales services, a long repair time enables the repairman to perform thorough maintenance and repair, thus improving customer satisfaction (Alizamir et al., 2013; Sun et al., 2021). Note that although the average total service duration of the two-step service queue is $\tau + l$, only the repair time τ generates a value to the customers. That is, customers' perceived service value depends on the repair time that follows a stochastic distribution with mean value τ (because we only assume the total service duration follows the exponential distribution). In particular, the repair quality is reflected in the service value function $V(\tau)$, which increases with the repair time τ. Moreover, the marginal value to customers from an increase in the repair time is diminishing. Hence, following the prior literature (Anand et al., 2011), we model the customer's perceived value of the service as a nondecreasing and concave function of the repair time as follows:

$$V(\tau) = V_b + \alpha\left(\frac{1}{\tau_b} - \frac{1}{\tau}\right), \tag{1}$$

where V_b is the minimum service value at the minimum repair time τ_b and $\alpha > 0$ captures how the service quality changes when the repair time deviates from τ_b. For after-sales services, τ_b can be viewed as the minimum repair time to recover the product to the working state. V_b reflects the customers' feelings when many new parts are consumed.

Expected spare parts expense. To save time and serve more customers in a limited duration, the service provider may have an incentive to replace rather than repair some failing parts. As a result, the customer incurred expense for spare parts increases. For example, the air conditioner maintenance service provider may advise the customer to replace a compressor that can be repaired with sufficient repair time. Customers have to make their

purchasing decisions based on the expected spare parts expense because they cannot know their exact requirements prior to the service provider's diagnosis. Specifically, Feeney and Sherbrook (1966) have proved that at a random point in time after reaching a steady state, the number of parts $O(\tau)$ to be replaced follows a Poisson distribution, whose mean depends on the repair effort τ. Following the prior literature (Wang et al., 2019), we model the expected consumption of parts as a decreasing function of the repair time, i.e., $\xi = E[O(\tau)] = 1 + \beta\left(\frac{1}{\tau} - \frac{1}{\tau_b}\right)$. Hence, the probability that the number of parts to be replaced is x equals $P(X = x) = \frac{\xi^x}{x!}e^{-\xi}$. The prices of the required parts may be high (price K, probability ϕ) or low (price k, probability $1 - \phi$). As a result, the expected expense that a customer incurs for spare parts is as follows:

$$C(\tau) = \sum \frac{\xi^x}{x!}e^{-\xi}x\big(K\phi + k(1-\phi)\big) \tag{2}$$

where τ_b is the basic service time, $\beta > 0$ measures how the spare parts to be replaced change with the repair time, and the number "1" denotes the benchmark spare parts consumption at the basic repair time. We assume that β is sufficiently small so that $\xi > 0$. Since the revenue from the spare parts is shared by all the agents in the supply chain (e.g., parts manufacturers, logistics providers, and the service provider), we assume that the service provider can retain a fraction $\delta \in (0,1)$ of $C(\tau)$ as part of its revenue.

Customer's net utility function. To model customer's decisions, we let c_w be the customer's waiting cost per unit time. In addition, the customers also incur a service price P and an expected expense for spare parts $C(\tau)$. Each customer encounters a two-step service process with an average repair time τ and the generated service value is $V(\tau)$. Customers are homogenous and rational, i.e., they know the service quality, the service fee, and the consequent spare parts expense, and they can deduce the effective arrival rate λ based on their participation decisions. We denote the induced expected waiting time in the queue by $W(\lambda; N, \tau)$, which strictly increases in λ. Therefore, the expected net utility of a customer is as follows:

$$U(N, P, \tau) = V(\tau) - c_w W(\lambda; N, l + \tau) - P - C(\tau). \tag{3}$$

Each customer has two pure strategies to choose from, namely, to join or balk the service. The joining probability $p \in [0,1]$ is a pure strategy if $p = 0$ or 1 and is a mixed strategy otherwise. Without loss of generality, we assume that customers obtain zero utility if they choose to balk the service. Thus, customers will join the service if their expected utility is $U(N,P,\tau) > 0$, and market equilibrium is achieved when $U(N,P,\tau) = 0$.

Stackelberg game sequence. The decision process corresponds to a two-stage Stackelberg game between the service provider and customers. In the first stage, the service provider acts as the leader and sets the service price P, average repair time τ, and number of servers N. In the second stage, the customers determine their purchasing decisions based on the given $N,P,$ and τ. Using backward induction, we first derive the customer equilibrium decision in Proposition 1 for given $N,P,$ and τ.

Proposition 1. Given the service provider's decision (N,P,τ)*, let* $\lambda(N,P,\tau)$ *denote the unique solution of* λ *satisfying* $U(N,P,\tau) = 0$*, where*

$$\lambda(N,P,\tau) = W^{-1}\left(\frac{V(\tau) - P - C(\tau)}{c_w}\right).$$

The customer mixed equilibrium strategy is as follows:

$$p^e(N,P,\tau) = \begin{cases} 1 & \text{if } \Lambda \le \lambda(N,P,\tau) \\ \frac{\lambda(N,P,\tau)}{\Lambda} & \text{if } 0 < \lambda(N,P,\tau) < \Lambda, \\ 0 & \text{if } \lambda(N,P,\tau) \le 0 \end{cases}$$

and the equilibrium arrival rate is $\lambda^e(N,P,\tau) = p^e(N,P,\tau)\Lambda$.

The above equilibrium defines three market outcomes, namely full participation, partial participation, and no participation. Detailed illustrations of the three market outcomes can be found in Anand et al. (2011), Wang et al. (2010), and Hu et al. (2017).

Our model is not confined to a specific maintenance treatment, e.g., preventive maintenance or corrective maintenance, and is applicable to all maintenance services that involve parts repair and replacement activities. We solve the after-sales service optimization problem in two different settings, which are distinguished by whether the service provider promises the maximum mean total sojourn time to customers as discussed in Section 1.

4. Analysis

It is useful to solve the problem under two different settings. In the first setting, we consider that the service provider does not promise a maximum total sojourn time to customers. Then, we introduce the maximum total sojourn time constraint in the second setting. We compare the equilibrium of these two settings. The results provide guidelines for after-sales service providers on managing different service contracts.

4.1 Without the Maximum Total Sojourn Time Constraint

We first examine the case where the number of servers is exogenously given, which represents the short-run decision setting in which the capacity, in terms of the number of servers N, is inflexible. Then, we consider the long-run decision setting in which N is endogenous. For example, an air conditioner maintenance provider hires repairmen to fulfill customer requests. In the short-run setting (e.g., a single summer or winter), the number of repairmen is relatively fixed, e.g., high in summer and low in winter; however, in the long-run setting (e.g., from summer to winter), the provider hires more repairmen in summer than in winter.

4.1.1 Exogenous Number of Servers

First, we analyze the service provider's equilibrium with an exogenous number of servers. In particular, we consider the special case $N = 1$, which serves the following two purposes: (1) The case is observed in practice, e.g., in some remote rural areas, air conditioner manufacturers or retailers often assign a local agent to sell the products and undertake maintenance services, so the individual agent acts as the sole repairman in this case. (2) The case serves as an important benchmark for the subsequent analysis. In this case, the service system corresponds to an *M/M/1* system with service duration $l + \tau$, but the effective repair time is τ. Three market outcomes, namely full, zero, and partial participation corresponding to the joining probabilities $p(P, \tau) = 1$, 0, and $0 < p(P, \tau) < 1$, respectively, exist in the *M/M/1* system, depending on the total market potential Λ (Anand et al., 2011). In this subsection, we focus on the case where $0 < p(P, \tau) < 1$. When $p(P, \tau) = 1$, the problem becomes a special case of endogenous number of servers, which we will discuss later. The customers will obtain zero utility if they choose to balk the service. We assume $\lambda < \frac{1}{l+\tau}$ so that the queue is not

infinite. As a result, the customer's expected waiting time equals $W = \frac{1}{\frac{1}{l+\tau}-\lambda}$. Substituting W into the customer's net utility in (3), we obtain the effective arrival rate as follows:

$$\lambda^e(P,\tau) = \frac{1}{l+\tau} - \frac{c_w}{V(\tau)-P-C(\tau)}. \quad (4)$$

Therefore, we obtain the service provider's objective function as

$$\max_{P,\tau} R(P,\tau) = \big(P + \delta C(\tau)\big)\lambda^e(P,\tau). \quad (5)$$

Solving the optimization problem in (5), we obtain the following results.

Proposition 2. When the service provider does not promise a maximum total sojourn time to customers and the number of servers is exogenous, the equilibrium can be characterized as follows:

(i) the optimal service time is $\tau_1^{ns} = \frac{1+\sqrt{1+l\eta}}{\eta}$;

(ii) the optimal service price is

$$P_1^{ns} = V_b + \frac{\alpha}{\tau_b} - \left(1 - \frac{\beta}{\tau_b}\right)\big(K\phi + k(1-\phi)\big) - \Big(\alpha + \beta\big(K\phi + k(1-\phi)\big)\Big)\frac{\left(\sqrt{1+l\eta}-1\right)}{l} -$$

$$\sqrt{c_w(1+l\eta)}\Big(\alpha + (1-\delta)\beta\big(K\phi + k(1-\phi)\big)\Big);$$

(iii) the induced equilibrium arrival rate is

$$\lambda_1^{ns} = \frac{\eta}{1+l\eta+\sqrt{1+l\eta}} - \sqrt{\frac{c_w}{\Big(\alpha+(1-\delta)\beta\big(K\phi+k(1-\phi)\big)\Big)(1+l\eta)}}, \text{ where } \eta = \frac{V_b-(1-\delta)\big(K\phi+k(1-\phi)\big)}{\alpha+(1-\delta)\beta\big(K\phi+k(1-\phi)\big)} + \frac{1}{\tau_b}.$$

Superscript *ns* denotes the scenario without the maximum total sojourn time constraint, while subscript *1* denotes the one-server queue. The analytical results in Proposition 2 can be used to determine the service price in the maintenance contracts. Based on Proposition 2, we study how the external parameters affect the above equilibrium. First, we investigate the impacts of the repairman travelling time l and present the results in Corollary 1.

Corollary 1: If the service provider does not promise a maximum total sojourn time, the optimal repair time τ_1^{ns} *decreases and the equilibrium arrival rate* λ_1^{ns} *increases when the repairman travelling time* l *decreases.*

The repairman travelling time l is non-negligible in practice. The above results imply that l and τ act as 'complements' because the optimal repair time in step 2 decreases when the travelling time decreases. When making join-or-balk decisions, customers take into

account the spare parts consumption. We refer to the service value minus the expense for spare parts ($V(\tau) - C(\tau)$ in (3)) as the individual customer benefit. A direct implication of this result is that when a logistics innovation reduces the average repairman travelling time in Step 1 (e.g., the repairman's transport tool changes from bicycle to truck), the service provider derives more benefits by offering the service to more customers, i.e., λ_1^{ns} increases with l. To keep the service away from the infinite queue, the service manager has to decrease the repair time τ in step 2. As a result, the individual customer benefit declines. That is, the service manager chooses a lower individual benefit but simultaneously decreases the system congestion in equilibrium when the repairman travelling time decreases.

Then, we elaborate on the impacts of the economic parameters $K(k)$ and δ, which are related to the spare parts consumption, on the service provider's decisions.

Corollary 2. If the number of servers is exogenous,
(i) the optimal repair time τ_1^{ns} decreases but the equilibrium arrival rate λ_1^{ns} increases when the share of revenue δ increases, and
(ii) the optimal repair time τ_1^{ns} increases but the equilibrium arrival rate λ_1^{ns} decreases as the price for spare parts K or k increases.

Corollary 2 (i) indicates that if the service provider can earn a large proportion of the revenue from the spare parts, i.e., δ increases, it is optimal for the service provider to decrease the repair time so as to make customers consume more spare parts. As a result, the service provider has flexibility in admitting more customers to join the service. However, our result in Corollary 2 (ii) may initially seem surprising in that even when selling parts becomes more profitable, i.e., K or k increase, the provider will improve the repair time to decrease the spare parts consumption in equilibrium. The underlying intuition is that customers anticipate spare parts expense when making their acquisition decisions. In particular, a customer incurs a cost of $C(\tau)$ for spare parts consumption, but only $\delta C(\tau)$ ($\delta < 1$) can be retained by the maintenance service provider. As a consequence, the revenue that the service provider obtains from spare parts cannot fully compensate for the expense that the customers incur (customer loss). Moreover, the deficit continues to widen as K or k increases. Hence, the service provider has an incentive to enhance the service quality and decrease the spare parts

consumption to maintain the service attractiveness to customers. Consequently, the induced effective arrival rate decreases because the repair time increases.

We conduct numerical studies to show how the parameters l, δ, and $K(k)$ affect the service provider's optimal price and profit. Figure 2 shows the numerical results intuitively (the figures in the same line share the same parameters).

Observation 1: *(i) As the repairman travelling time* l *increases, both the optimal service price* P_1^{ns} *and the profit* R_1^{ns} *are unimodal in* l*;*
(ii) as the share of revenue δ *increases, the optimal service price* P_1^{ns} *decreases but the profit* R_1^{ns} *increases;*
(iii) as the unit spare parts price K *or* k *increases, the optimal service price* P_1^{ns} *increases but the profit* R_1^{ns} *decreases.*

The above numerical analysis not only illustrates the effects of the parameters on the optimal price and profit but also clarifies the analytical results in Corollary 2. That is, an increase in the value of δ is beneficial to the service provider while an increase in the unit spare parts cost K or k is harmful to the service provider, even though the service provider can charge a higher price for the service.

Please insert Figure 2 here.

Figure 2. Effects of l, δ, K and k on the optimal service price and profit

4.1.2 Endogenous Number of Servers

In this subsection we study the service provider's equilibrium by considering an endogenous number of servers. An endogenous number of servers is found in practice. For example, when facing numerous requests in summer, the air-conditioner maintenance service provider hires more repairmen than in winter. In addition, customers that are geographically adjacent are often assigned to the same repairman, who picks up the necessary parts and completes the service requests in sequence. That is, customers cannot switch between different repairmen, and each customer is in essence in a dedicated *M/M/1* queue. Although N repairmen are involved, the service system corresponds to N parallel *M/M/1* queues with full capacity $\frac{N}{l+\tau}$. This queuing structure has been adopted in the literature, see, e.g., Tang et al. (2021) and

Tong et al. (2016). In Section 5 we extend the study to the model with an *M/M/N* queue in order to examine the robustness of our results.

As a result, the customer expected waiting time in the system is $W(N,\tau)=\frac{1}{\frac{N}{l+\tau}-\lambda}$. Thus, we obtain the effective arrival rate of the queue as follows:

$$\lambda^e(N,P,\tau)=\frac{N}{l+\tau}-\frac{c_w}{V(\tau)-P-C(\tau)}. \tag{6}$$

Furthermore, we formulate the service provider's objective function as follows:

$$\max_{N,P,\tau} R(N,P,\tau)=\big(P+\delta C(\tau)\big)\lambda^e(N,P,\tau)-\theta N, \tag{7}$$

where θ is the investment cost for adding a server to the service system. To solve the optimization problem (7), we first derive the optimal number of servers for given repair time τ and service price P in Proposition 3.

Proposition 3: If the service provider can obtain a positive profit, i.e., $R(N,P,\tau)>0$*, capturing all the potential demand into the service is the unique equilibrium for the provider. Moreover, the number of servers should be set at the minimum level to do so. Specifically,* $N(P,\tau)=(\tau+l)\left[\Lambda+\frac{c_w}{V(\tau)-P-C(\tau)}\right]$.

Recall from subsection 4.1.1 that in the single-server setting, the market outcome can be one of three cases, namely full participation, no participation, and partial participation. However, Proposition 3 indicates that full participation is the unique equilibrium in the presence of endogenous servers, resulting from the economies of scale embedded in the multi-server queuing system. The explanation in our model is as follows: From (6), we note that for given repair time τ and service price P, λ grows more than proportionately in N. Therefore, the service provider's profit increases in N until the effective demand reaches the potential demand Λ. Once $\lambda=\Lambda$, increasing N only increases the staffing cost. Hence, the optimal number of servers should be set at the minimum level that can exactly involve the entire potential demand for the service. Note that the full participation outcome in the *M/M/1* queue, i.e., $p(P,\tau)=1$, is a special case of the N parallel *M/M/1* queues (N=1).

Our optimization problem is significantly simplified with Proposition 3. The result shows that only the two decision variables, i.e., P and τ, can be determined independently.

Substituting $N(P,\tau)$ into the objective function, we have the following:

$$\max_{P,\tau} R(P,\tau) = \big(P + \delta C(\tau)\big)\Lambda - \theta(\tau + l)\left[\Lambda + \frac{c_w}{V(\tau) - P - C(\tau)}\right]. \quad (8)$$

We solve the above optimization problem and present the equilibrium results in Proposition 4.

Proposition 4. When the service provider does not promise a maximum total sojourn time to customers, the optimal decisions regarding the repair time, service price, and number of servers are as follows:

(i) the service provider chooses the repair time τ_N^{ns} *that solves the following equation*

$$\Lambda\big[\alpha + (1-\delta)\beta\big(K\phi + k(1-\phi)\big)\big]\tau^{-2} - \sqrt{c_w\Lambda\theta}(\tau + l)^{-\frac{1}{2}} - \theta\Lambda = 0;$$

(ii) the corresponding optimal service price equals

$$P_N^{ns} = V_b + \alpha\left(\frac{1}{\tau_b} - \frac{1}{\tau_N^{ns}}\right) - \left(1 + \beta\left(\frac{1}{\tau_N^{ns}} - \frac{1}{\tau_b}\right)\right)\big(K\phi + k(1-\phi)\big) - \sqrt{\frac{c_w\theta(\tau_N^{ns} + l)}{\Lambda}}; \text{ and}$$

(iii) the corresponding optimal number of servers is

$$N_N^{ns} = \Lambda(\tau_N^{ns} + l) + \sqrt{c_w\Lambda(\tau_N^{ns} + l)/\theta}.$$

Superscript *ns* denotes the scenario without the maximum total sojourn time constraint, while subscript *N* denotes the multi-server queue. These analytical results are useful for helping the service provider to determine the service price and average product downtime in maintenance contracts. In addition, from Proposition 4, we can readily obtain Corollaries 3 and 4.

Corollary 3 In the presence of an endogenous number of servers, as the travelling time l *increases:*

(i) the optimal repair time τ_N^{ns} *increases in* l*. Furthermore, there exists an upper threshold* $\overline{\tau}$ *when* $l \to \infty$ *and a lower threshold* $\underline{\tau}$ *when* $l \to 0$*. Specifically,* $\overline{\tau} = \sqrt{\alpha + (1-\delta)\beta\big(K\phi + k(1-\phi)\big)/\theta}$*, and* $\underline{\tau}$ *is the solution of* $\Lambda\big[\alpha + (1-\delta)\beta\big(K\phi + k(1-\phi)\big)\big]\tau^{-2} - \sqrt{c_w\Lambda\theta}\tau^{-\frac{1}{2}} - \theta\Lambda = 0;$

(ii) the relationship between repairman travelling time l and the optimal price P_N^{ns} is nonmonotonic. Specifically, when $\frac{l}{\tau_N^{ns}(l)} < \frac{\delta\beta(K\phi+k(1-\phi))}{\alpha+(1-\delta)\beta(K\phi+k(1-\phi))}$, P_N^{ns} increases in l; otherwise, P_N^{ns} decreases in l;

(iii) the induced number of servers N_N^{ns} increases in l.

Corollary 3 indicates that without the maximum total sojourn time constraint, the travelling time l and the repair time τ still act as "complements" when the number of servers is endogenous. In addition, we find that the optimal price is unimodal in l, which is also consistent with the result for the case of an exogenous number of servers (shown in Figure 2). As a result of the increase in repair time, the service provider has to hire more servers to ease system congestion since the optimal demand always equals the market potential Λ.

Next, we analyze the effects of δ and $K(k)$ on the equilibrium outcomes given in Proposition 3 and give the results in Corollary 4.

Corollary 4 If the service provider does not promise a maximum total sojourn time and the number of servers is endogenous, we have the following:

(i) as the share of revenue δ increases, the equilibrium repair time τ_N^{ns}, the service price P_N^{ns}, and the number of servers N_N^{ns} decrease;

(ii) as the unit spare parts price K or k increases, both the equilibrium repair time τ_N^{ns} and number of servers N_N^{ns} increase.

Corollary 4 (i) shows that the service provider will decrease the repair time to improve the spare parts consumption if it can retain more revenue from spare parts consumption. This result is consistent with the case of an exogenous number of servers. In addition, the service manager will reduce the number of servers because each server can effectively render more services as the repair time decreases. Moreover, Corollary 4(ii) indicates that the service provider increases the repair time to decrease the spare parts consumption in equilibrium when selling spare parts becomes more profitable. As a result, more servers are required to enlarge the service provider's capacity.

4.2 With the Maximum Total Sojourn Time Constraint

So far, we have assumed that the service provider can determine the congestion level independently. However, the service provider and customers often contractually define a maximum mean total sojourn time in practice, as illustrated in Section 1. From the analysis in Sections 4.1.1 and 4.1.2, we know that the main insights are basically consistent between the cases with exogenous and endogenous number of servers. Hence, in this section, we focus only on an endogenous number of servers.

4.2.1 Customers

If the service provider makes a commitment to the customers regarding a maximum total sojourn time, a rational customer will make his purchasing decision based on the commitment (Zhang et al., 2009). For example, when a customer calls for an air conditioner after-sales service in summer, the provider always provides him with a promised waiting time. Also, a maximum air conditioner downtime may be stipulated in the service contract. Let d be the committed maximum total sojourn time. Then, the customers that choose to join the queue derive the utility

$$U(N,P,\tau) = V(\tau) - c_w d - P - C(\tau). \tag{9}$$

We then analyze the effective arrival rate for given customer utility (9). Recall that without the sojourn time constraint, the effective arrival rate always equals the market potential Λ in equilibrium (Section 4.1.2). In addition, the true maximum customer utility is $U(0;N,P,\tau) = V(\tau) - P - C(\tau) - c_w(\tau + l)$, which denotes the situation where a customer can be served immediately upon his call, and the true minimum customer utility equals 0 in equilibrium. However, when requesting the service, customers do not know the true utility they can obtain. Hence, a rational customer will make his acquisition decision based on the expected net utility defined in (9). Hence, the effective demand is aggregated from customers whose utility range is between 0 and $U(N,P,\tau)$. As a result, the induced equilibrium $\lambda(N,P,\tau)$ is the solution of $\frac{U(N,P,\tau)-0}{U(0;N,P,\tau)-0} = \frac{\lambda}{\Lambda}$, i.e.,

$$\lambda^e = \Lambda\left[1 - \frac{c_w(d-(\tau+l))}{V(\tau)-P-C(\tau)-c_w(\tau+l)}\right]. \tag{10}$$

4.2.2 Service Provider

Based on the above analysis, we formulate the service provider's profit maximization problem as

$$\max_{N,P,\tau} R(N,P,\tau) = \big(P + \delta C(\tau)\big)\lambda^e - \theta N.$$

$$s.t.\ \ W(N,\tau) = \frac{1}{\frac{N}{\tau+l} - \lambda^e} \le d. \tag{11}$$

Constraint (11) stipulates that the service provider has to provide adequate capacity to ensure that the customers' expected waiting time never exceeds the committed maximum mean total sojourn time. The sojourn time constraint is widely used in the service capacity allocation research (see, e.g., Zhang et al., 2009; Nguyen and Wright, 2015). Similar to the analysis in Section 4.1.2, we first derive the optimal number of servers for a given repair time τ and price P in Proposition 5.

Proposition 5. With the maximum mean total sojourn time constraint, it is optimal for the service provider to set the number of servers at the level that makes the expected waiting time exactly equal to d. As a result, the optimal number of servers is

$$N(P,\tau) = \left[\Lambda\left(1 - \frac{c_w\big(d-(\tau+l)\big)}{V(\tau)-P-C(\tau)-c_w(\tau+l)}\right) + 1/d\right](\tau + l)\ .$$

With the formulation of the equilibrium demand λ^e and the number of servers N, the service provider's problem in (11) thus becomes the following:

$$\max_{P,\tau} R(P,\tau) = \big(P + \delta C(\tau)\big)\left[1 - \frac{c_w\big(d-(\tau+l)\big)}{V(\tau)-P-C(\tau)-c_w(\tau+l)}\right]\Lambda - \theta\left[\Lambda\left(1 - \frac{c_w\big(d-(\tau+l)\big)}{V(\tau)-P-C(\tau)-c_w(\tau+l)}\right) + \frac{1}{d}\right](\tau + l). \tag{12}$$

Proposition 6 gives the equilibrium of the optimization problem (12).

Proposition 6: Given the maximum mean total sojourn time d,

(i) when d is lower than the equilibrium waiting time without the sojourn time constraint, i.e., $d < W^*(\tau_N^{ns}) = \sqrt{\theta(\tau_N^{ns} + l)/c_w\Lambda}$,

(a) the optimal repair time τ_N^S is the solution of $\frac{\partial R(P^*(\tau),\tau)}{\partial \tau} = 0$, where $P^*(\tau) = B - \delta\left(1 + \beta\left(\frac{1}{\tau_N^S} - \frac{1}{\tau_b}\right)\right)\big(K\phi + k(1-\phi)\big) - \sqrt{DB}$;

(b) the optimal service price is $P_N^S = B(\tau_N^S) - \delta\left(1 + \beta\left(\frac{1}{\tau_N^S} - \frac{1}{\tau_b}\right)\right)\big(K\phi + k(1-\phi)\big) - \sqrt{D(\tau_N^S)B(\tau_N^S)}$;

(c) the corresponding number of servers is $N_N^S = (\tau_N^S + l)\left(\Lambda - \Lambda\sqrt{\frac{D(\tau_N^S)}{B(\tau_N^S)}} + \frac{1}{d}\right)$;

(d) the induced equilibrium demand is $\lambda_N^S = \Lambda\left(1 - \sqrt{\frac{D(\tau_N^S)}{B(\tau_N^S)}}\right)$,

where $B = V(\tau) - (1-\delta)\left(1 + \beta\left(\frac{1}{\tau} - \frac{1}{\tau_b}\right)\right)\big(K\phi + k(1-\phi)\big) - c_w(\tau + l)$ *and* $D = c_w\big(d - (\tau + l)\big)$.

(ii) *When* d *is higher than the equilibrium waiting time without the sojourn time constraint, i.e.,* $d > W^*(\tau_N^{ns}) = \sqrt{\theta(\tau_N^{ns} + l)/c_w\Lambda}$, *the equilibrium outcomes are those given in Proposition 4.*

Superscript *s* denotes the scenario with the maximum total sojourn time constraint, while subscript *N* denotes the multi-server queue. Proposition 6 can be used to determine the service price and service level (sojourn time) in maintenance service contracts. We perform numerical studies to illustrate Proposition 6 and perform a sensitivity analysis of the parameters. We set the parameters as follows: $V_b = 4,\ \alpha = 2,\ \tau_b = 0.5,\ \delta = 0.3,\ \beta = 0.5,\ K = 6,\ k = 4,\ c_w = 1,\ \Lambda = 2,\ \phi = 0.5$ and $\theta = 3$. We set three different travelling times, i.e., $l = 0.01,\ 0.5,$ and 1. We illustrate the effects of d in Figure 3 and summarize the results in Observation 2. Note that to investigate the effects of d, we only show the case where the sojourn time constraint is not binding (case (i) in Proposition 6).

Please insert Figure 3 here.

Figure 3. Effects of d on the equilibrium decisions

Observation 2: As the maximum mean total sojourn time d *increases,*

(i) the service provider improves the repair time to enhance the service quality and reduce the spare parts consumption;

(ii) the optimal service price is unimodal in d*; and*

(iii) the service provider admits fewer customers and hires more servers for the system.

We first observe that the optimal repair time increases with the defined maximum mean total sojourn time d. The reason is that the service provider has to improve the individual

benefit to eliminate the customers' concern regarding the increased waiting cost. Furthermore, the numerical examples show that a long maximum sojourn time may not necessarily suppress the service price. The underlying intuition is as follows: From Observation 1(i), we know that the individual benefit increases with d. When d is small, the increase in the individual benefit dominates the increase in the waiting cost. Hence, the service provider will improve the service price. However, the increase in the individual benefit is limited when d is large, which cannot fully compensate for the increased waiting cost. Thus, the service provider has to reduce the service price to attract customers. Because of the increase in repair time, the service provider has to hire more servers to ensure that the system congestion remains at a certain level.

Then, we want to explore whether the travelling time l influences the service provider's equilibrium decisions equally for the two cases with and without the maximum sojourn time constraint. In particular, we set the parameters as follows: $V_b = 3$, $\alpha = 2$, $\tau_b = 0.5$, $\delta = 0.3$, $\beta = 0.5$, $K = 6$, $k = 4$, $c_w = 1$, $\Lambda = 2$, $\phi = 0.5$, and $\theta = 3$. We consider three different values of d, i.e., $d = 1.5, 2, 3$. We summarize the results in Figure 4 and Observation 3.

Please insert Figure 4 here

Figure 4. Effect of l on the optimal repair time τ

Observation 3: Under the maximum mean total sojourn time constraint, the service provider will decrease the repair time to decline the service quality and improve the spare parts consumption as the repairman travelling time l increases.

In Section 4.1, we found that the optimal repair time and repairman travelling time are "complementary" since the optimal service time τ increases with the travelling time l. However, in Figure 4, we observe that the service time and travelling time are now "substitutable" because the optimal service time decreases when the travelling time increases. This implies that under the sojourn time constraint, the service provider will sacrifice the individual benefit to ease the congestion, i.e., $V(\tau) - C(\tau)$ decreases.

5. Extensions

5.1 Reusable parts

In the above analysis, we assume that the service provider uses a new part for replacement. However, in practice, the service provider may have an incentive to replace some parts that could still be repaired. Furthermore, these parts may be reusable after offline repair.

In Section 4, each customer incurs an expected spare parts expense $C(\tau)$ and the service provider retains $\delta C(\tau)$ as part of its revenue. In the presence of reusable parts, we assume that a fraction $r(0 < r < 1)$ of the parts are reusable and the service provider obtains all the revenue from these parts. Hence, the service provider's expected revenue from each replaced part is $rC(\tau) + (1 - r)\delta C(\tau)$. As a result, the service provider's objective function is:

$$\max_{N,P,\tau} R(N,P,\tau) = \big(P + rC(\tau) + (1-r)\delta C(\tau)\big)\lambda^e(N,P,\tau) - \theta N, \tag{13}$$

where $\lambda^e(N,P,\tau)$ is the effective demand in each of the three cases in Section 4. Note that when the number of servers N is exogenous, the server investment cost θN is ignored. By solving the above optimization problem and analyzing the equilibrium condition, we obtain the following result.

Proposition 7. In both the settings of exogenous and endogenous servers, as the proportion of reusable parts r increases, the service provider will decrease the repair time to improve the spare parts consumption.

When the parts could be repaired offline for later use again, the service provider has a greater incentive to encourage customers to replace more parts. The reason is that when reusable parts are involved, the increase in the provider's revenue from spare parts exceeds the decrease in customer's utility from additional spare parts expense. Although the effective demand decreases due to the customer utility loss, the provider can earn more profit from each served customer, i.e., $rC(\tau) + (1-r)\delta C(\tau) > \delta C(\tau)$. As a result, the service provider enjoys a higher profit margin with the additional spare parts consumption.

5.2 *M/M/N* queue vs. *N* parallel *M/M/1* queues

In Section 4, we model the service system as N parallel *M/M/1* queues, which is consistent with the setting of our model because each customer request is assigned to a dedicated

repairman. In this section, we consider the case where customers can be served by any repairmen, e.g., a repairman may serve the customers originally assigned to another repairman when the latter is too busy. Thus, the service system corresponds to the *M/M/N* queuing model. We examine whether our main results in Section 4 remain robust in this case.

Given the effective demand rate λ and service duration $l+\tau$, it is well known that the expected waiting time for an *M/M/N* queue is

$$W(\lambda; N, \tau) = \frac{1}{1+(N!(1-\rho)/(N^N\rho^N))\sum_{i=0}^{N-1}(N^i\rho^i/i!)}\left(\rho/\lambda(1-\rho)\right), \tag{14}$$

where $\rho = \lambda(l+\tau)/N$ denotes the system workload with $\rho < 1$. We first derive the equilibrium arrival rate of the *M/M/N* queue in Proposition 8.

Proposition 8. Without the maximum total sojourn time constraint, the customer equilibrium arrival rate equals to the potential arrival rate, i.e., $\lambda^e = \Lambda$. Moreover, the number of servers N should be set at the minimum level to do so. In particular, N is the minimum integer that satisfies $\frac{c_w}{1+(N!(1-\rho)/(N^N\rho^N))\sum_{i=0}^{N-1}(N^i\rho^i/i!)}\left(\rho/\Lambda(1-\rho)\right) \le V(\tau) - P - C(\tau)$.

Combining Proposition 3 and Proposition 8, we can conclude that without the maximum total sojourn time constraint, attracting all potential customers into the service is optimal for the provider if N is endogenous. The reason is that the economies of scale embedded in the *N* parallel *M/M/1* queues remain true in the *M/M/N* queue system. Customers will join the service when their expected net utility in Equation (3) is nonnegative; thus, the service provider should set the number of servers N at the minimum value to maintain a nonnegative customer net utility to extract more customer surplus.

Given the complicated waiting time in Equation (14), we are unable to derive further analytical results. Thus, we conduct numerical studies to check the robustness of our results based on Proposition 8. We use the data inferred from Jackson and Pascual (2008) as the parameter base values and test the impacts of changing parameter values around the bases in specific ranges. Jackson and Pascual (2008) studied the optimal service contract negotiation with aging industrial equipment, which is consistent with our research. The detailed numerical studies are shown in the Appendix-A, and we summarize the numerical results in Table 2.

Table 2 Summary of the effects of l, δ, K and k on service provider's equilibrium

	An *M/M/N* queue				*N* parallel *M/M/1* queues			
	τ^*	P^*	N^*	R^*	τ^*	P^*	N^*	R^*
l ↑	↑	↑↓	↑	↑↓	↑	↑↓	↑	↑↓
δ ↑	↓	↓	↓	↑	↓	↓	↓	↑
K ↑	↑	↑	↑	↓	↑	↑	↑	↓
k ↑	↑	↑	↑	↓	↑	↑	↑	↓

Increase (↑), decrease (↓), unimodal (↑↓).

We can observe that the structural results in the case of *N* parallel *M/M/1* queues remain robust in the case of an *M/M/N* queue.

In the presence of the maximum total sojourn time constraint, customers make their decisions based on the promised sojourn time d. Hence, the customer net utility and the resulting effective arrival rate are irrelevant to the queuing structure. That is, the effective arrival rate λ^e in the case of the *M/M/N* queue is the same as that in the case of *N* parallel *M/M/1* queues (Equation (10)). Then, we characterize the service provider's optimal decisions regarding N for given P and τ in Proposition 9.

Proposition 9. With the maximum mean total sojourn time constraint, it is optimal for the service provider to set the number of servers at the minimum integer that satisfies $W(\lambda^e; N, \tau) \leq d$.

Based on Proposition 9, we continue the numerical analysis to study the relationship between travelling time l and repair time τ. We also present the detailed numerical study in the Appendix-A. The numerical results clearly show that in both queuing structures, the travelling time l and repair time τ act as "substitutes" in that the equilibrium repair time decreases with the travelling time.

Therefore, by conducting the above numerical studies, we can conclude that our analytical results in the case of *N* parallel *M/M/1* queueing remain robust in the case of the M/M/N queuing.

6. Implications for After-sales Service Managers

We highlight the managerial implications of our results for after-sales service managers in this section. As a typical case, we use air conditioner maintenance services for illustration.

(i) *The impacts of the revenue from the spare parts*. We show that although an increase in the revenue share (δ) or parts prices (K, k) can improve the service provider's revenue from parts consumption, the two parameters have inverse impacts on the equilibrium outcomes. Specifically, as the share increases, the air conditioner maintenance provider decreases the average repair time to improve spare parts consumption. However, as the spare parts become more profitable, i.e., K or k increases, the service manager increases the repair time to enhance service quality and reduce spare parts consumption to ensure that the after-sales service remains attractive to the customers (Corollaries 2 and 4).

(ii) *The impacts of the travelling time*. If the air conditioner maintenance provider improves the repairmen's transport, i.e., travelling time l decreases, e.g., from bicycle to truck, the provider has to adjust the optimal repair time accordingly. In particular, the provider should exploit the benefits by decreasing the average repair time to reduce service quality and increase spare parts consumption if no maximum mean total sojourn time is committed. However, under the maximum mean total sojourn time constraint, it is profitable for the provider to increase the repair time to attract more customers for the service and keep customers' air conditioner downtime at a certain level. (Corollary 1, Corollary 3, and Observation 3).

(iii) *Market coverage strategy in different settings*. We also present the provider's optimal market coverage strategies in different settings. First, in the short-run setting without the promised maximum mean total sojourn time, the market equilibrium is one of the three cases, namely full, zero, or partial coverage. Second, in the long-run setting, it is optimal for the air conditioner maintenance provider to use the service to engage all the potential customers (Proposition 3). Third, when the air conditioner maintenance provider stipulates a maximum downtime in the service contract, we show that partial coverage is always optimal (Proposition 5).

(iv) *Pricing Strategy*. We show that it is not always optimal for the air conditioner maintenance provider to improve the service price as the repair quality increases. In particular, the relationship between the service price and service quality is rather ambiguous. The provider should set the service price by considering different external factors rather than

adjusting the price blindly as the service quality increases (Observation 1, Corollary 3, and Observation 2).

(v) *Service contracts*. Our results are readily applicable to service contract design. For example, we present the service price and expected waiting time decisions under different queuing structures (Propositions 1-6), which provide guidance for the air conditioner maintenance provider to set the price and average downtime in maintenance contract. In addition, we show that depending on whether an air conditioner maximum downtime is stipulated in the service contract, external factors affect the optimal decisions in different ways.

7. Conclusions

In practice, after-sales services are always accompanied by spare parts consumption and repairman travel. We examine the effects of such consumption and travel on the service provider's decisions of optimal repair time, price, and number of servers. We focus on the setting where the service quality increases and the spare parts consumption decreases as the average repair time increases. We conclude our study as follows:

First, the dual concerns of service quality and spare parts consumption lead to a counterintuitive impact of the spare parts price: the service provider will improve the repair quality and decrease the spare parts consumption when selling the spare parts becomes more profitable. The reason is that the revenue growth from spare parts consumption cannot fully compensate for the customer loss in equilibrium. Second, we reveal that in different settings, i.e., with and without the maximum mean total sojourn time constraint, the repairman travelling time affects the optimal repair time in different ways. In particular, the repair time and travelling time are complementary in the sense that the equilibrium average repair time increases in the travelling time without the sojourn time constraint. However, under the sojourn time constraint, the repair time and travelling time are substitutable in the sense that the equilibrium average repair time decreases with the travelling time. Third, we show that the service provider adopts different market coverage strategies in different settings. Finally, we also show the followings: (i) the service quality decreases but spare parts consumption increases when the service provider's share of the revenue from spare parts consumption

increases, (ii) the service price may not necessarily increase with the service quality, and (iii) the service quality increases but the spare parts consumption decreases if the maximum mean total sojourn time increases.

We close by stating some limitations of our study. Specifically, the repairman travelling time depends on the traffic jams or rush hours experienced by the repairman. Future research may incorporate a piecewise function to model the repairman travelling time and study its effects on equilibrium. In addition, the payment scheme of an after-sales service is often comprehensive, which may consider the reliability level and a warranty and involve multiple agents. Future studies may consider extending our work to consider other pertinent factors to produce more interesting and insightful results.

References

Alger I, Salanie F (2006) A theory of fraud and overtreatment in experts markets. *Journal of Economics & Management Strategy*, 15(4): 853-881.

Alizamir S, De Véricourt F, Sun P (2013) Diagnostic accuracy under congestion. *Management Science*, 59(1): 157-171.

Anand KS, Paç MF, Veeraraghavan S (2011) Quality-speed conundrum: Trade-offs in customer-intensive services. *Management Science*, 57(1): 40-56.

Basten R, van Houtum G (2014) System-oriented inventory models for spare parts. *Surveys in Operations Research and Management Science*, 19(1):34-55.

Basten R, Ryan JK (2019) The value of maintenance delay flexibility for improved spare parts inventory management. *European Journal of Operational Research*, 278(2): 646-657.

Dai T, Akan M, Tayur S (2016) Imaging room and beyond: The underlying economics behind physicians' test-ordering behavior in outpatient services. *Manufacturing & Service Operations Management*, 19(1): 99-113.

Eruguz AS, Tan T, van Houtum GJ (2018) Integrated maintenance and spare part optimization for moving assets. *IISE Transactions*, 50(3): 230-245.

Feeney G J, Sherbrooke C C (1966) The (s−1, s) inventory policy under compound Poisson demand. *Management Science*, 12(5): 391-411.

Hu M, Li Y, Wang J (2018) Efficient ignorance: Information heterogeneity in a queue. *Management Science*, 64(6): 2650-2671.

Jackson C, Pascual R, 2008. Optimal maintenance service contract negotiation with aging equipment. *European Journal of Operational Research*, 189(2): 387-398.

Jain A, Bala R (2018) Differentiated or integrated: Capacity and service level choice for differentiated products. *European Journal of Operational Research*, 266(3): 1025-1037.

Jayaswal S, Jewkes EM (2016) Price and lead time differentiation, capacity strategy and market competition. *International Journal of Production Research*, 54(9): 2791-2806.

Kostami V, Rajagopalan S (2014) Speed–quality trade-offs in a dynamic model. *Manufacturing & Service Operations Management*, 16(1): 104-118.

Kurata H, Nam SH (2010) After-sales service competition in a supply chain: Optimization of customer satisfaction level or profit or both? *International Journal of Production Economics*,127(1): 136-146.

Kurata H, Nam SH (2013) After-sales service competition in a supply chain: Does uncertainty affect the conflict between profit maximization and customer satisfaction? *International Journal of Production Economics*, 144(1): 268-280.

Kurz J (2016) Capacity planning for a maintenance service provider with advanced information. *European Journal of Operational Research*, 251(2): 466-477.

Legnani E, Cavalieri S (2012) Modelling and measuring after-sales service delivery processes. *IFAC proceedings volumes*,45(6): 1684-1689.

Li X, Guo P, Lian Z (2016) Quality-speed competition in customer-intensive services with boundedly rational customers. *Production and Operations Management*, 25(11): 1885-1901.

Nguyen TH, Wright M (2015) Capacity and lead-time management when demand for service is seasonal and lead-time sensitive. *European Journal of Operational Research*, 247(2): 588-595.

Nourelfath M, Nahas N, Ben-Daya M (2016) Integrated preventive maintenance and production decisions for imperfect processes. *Reliability Engineering and System Safety*, 148: 21–31.

Panagiotidou S (2014) Joint optimization of spare parts ordering and maintenance policies for multiple identical items subject to silent failures. *European Journal of Operational Research*, 235(1): 300-314.

Paç MF, Veeraraghavan S (2015) False diagnosis and overtreatment in services. Working Paper, University of Pennsylvania.

Qin X, Shao L, Jiang Z Z (2020) Contract design for equipment after-sales service with business interruption insurance. *European Journal of Operational Research*, 284(1): 176-187.

Rezapour S, Allen JK, Mistree, F (2016) Reliable product-service supply chains for repairable products. *Transportation Research Part E: Logistics and Transportation Review*, 95: 299-321.

Saccani N, Johansson P, Perona M (2007) Configuring the after-sales service supply chain: A multiple case study. *International Journal of Production Economics*, 110(1-2): 52-69.

Sleptchenko A, Heijden M C, Harten A (2003) Trade-off between inventory and repair capacity in spare part networks. *Journal of the Operational Research Society,* 54, 263-272.

Sleptchenko A, Al Hanbali A, Zijm H (2019) Joint planning of service engineers and spare parts. *European Journal of Operational Research*, 271(1): 97-108.

Stenström C, Norrbin P, Parida A, and Kumar U (2016) Preventive and corrective maintenance–cost comparison and cost–benefit analysis. *Structure and Infrastructure Engineering*,12(5): 603-617.

Sun M, Wu F, Zhao S (2020) Machine diagnostic service center design under imperfect diagnosis with uncertain error cost consideration. *International Journal of Production Research*, 58(10): 3015-3035.

Sun M, Wu F, Ng C T, Cheng T C E (2021) Effects of imperfect IoT-enabled diagnostics on maintenance services: A system design perspective. *Computers & Industrial Engineering*, published online.

Tang Y, Guo P, Tang C S and Wang Y (2021) Gender-Based operational issues arising from on-demand ride-hailing platforms: safety concerns and system configuration. *Production and Operations Management*, published online.

Tong C, Nagarajan M, Cheng Y (2016) Operational impact of service innovations in multi-step service systems. *Production and Operations Management*, 25(5): 833-848.

Topan E, Tan T, van Houtum G J, and Dekker R (2018) Using imperfect advance demand information in lost-sales inventory systems with the option of returning inventory. *IISE Transactions*, 50(3): 246-264.

Wang X, Debo LG, Scheller-Wolf A, Smith SF (2010) Design and analysis of diagnostic service centers. *Management Science*, 56 (11): 1873–1890.

Wang X, Wu Q, Lai G, Scheller-Wolf A (2019) Offering Discretionary Healthcare Services with Medical Consumption. *Production and Operations Management*, 28(9): 2291-2304.

Zhang Z, Tan Y, Dey D (2009) Price competition with service level guarantee in web services. *Decision Support Systems*, 47: 93-104.

Zhao X, Stecke KE, Prasad A (2012) Lead time and price quotation mode selection: Uniform or differentiated? *Production and Operations Management*, 21(1): 177-193.

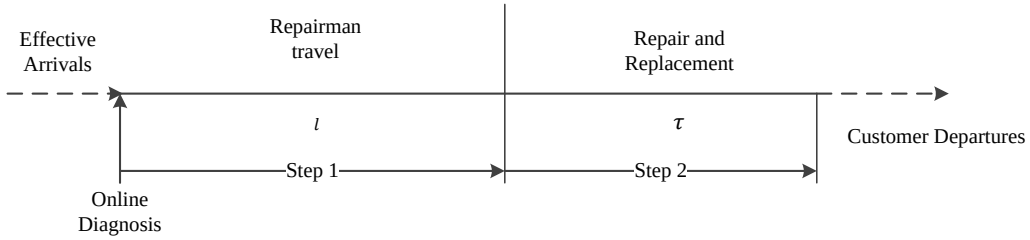


Figure 1. Illustration of the service process

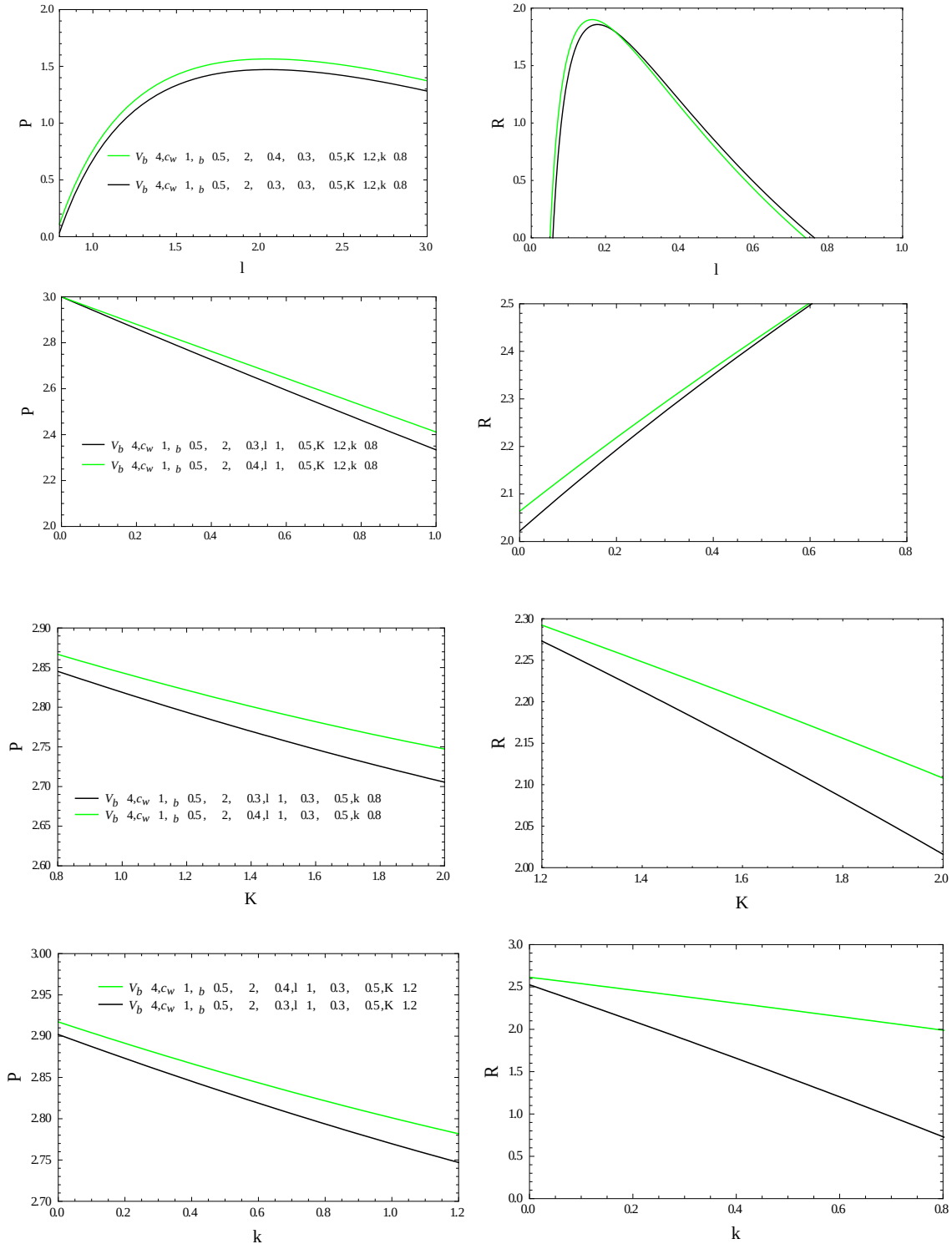


Figure 2. Effects of l, δ, K and k on the optimal service price and profit

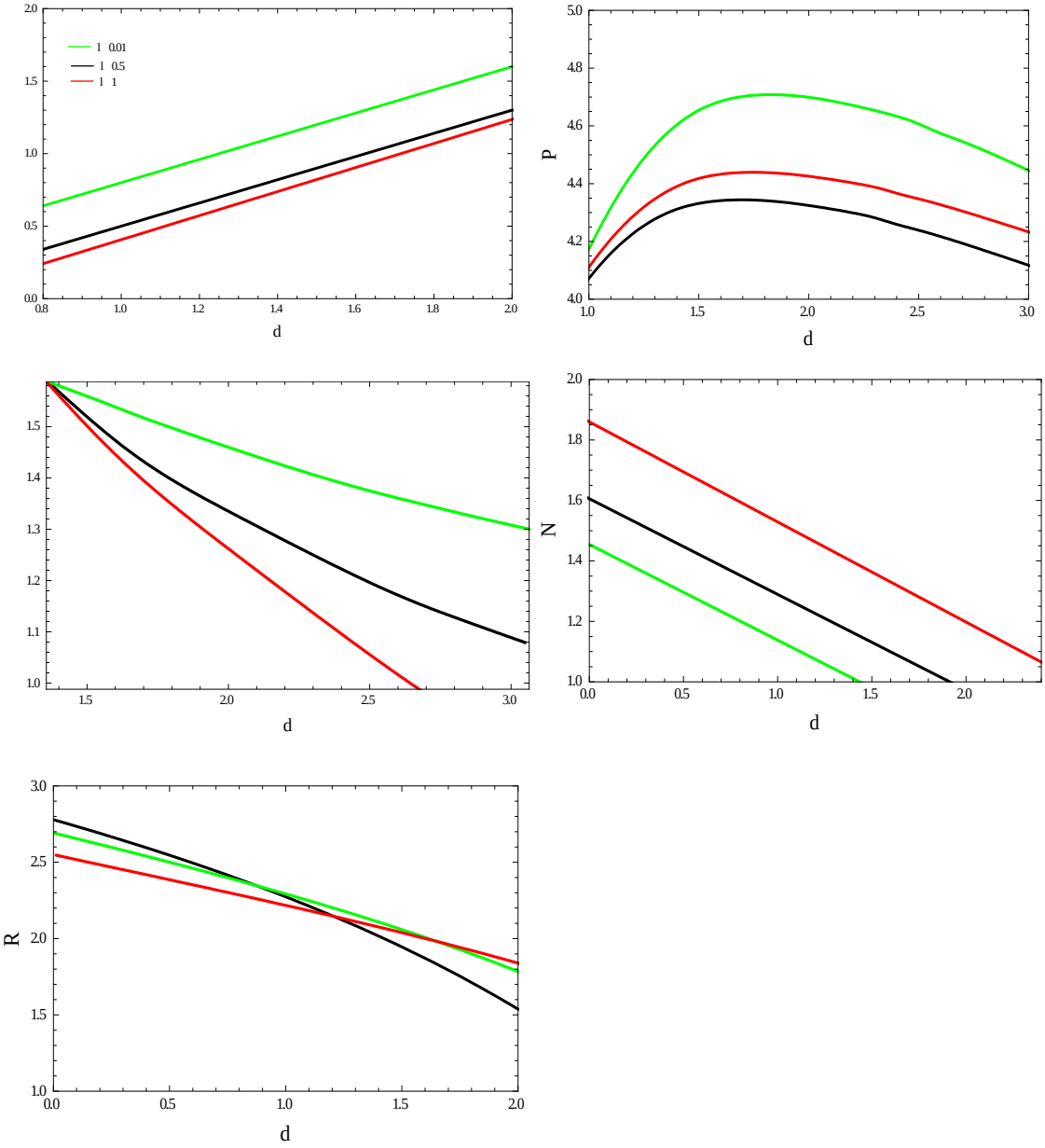
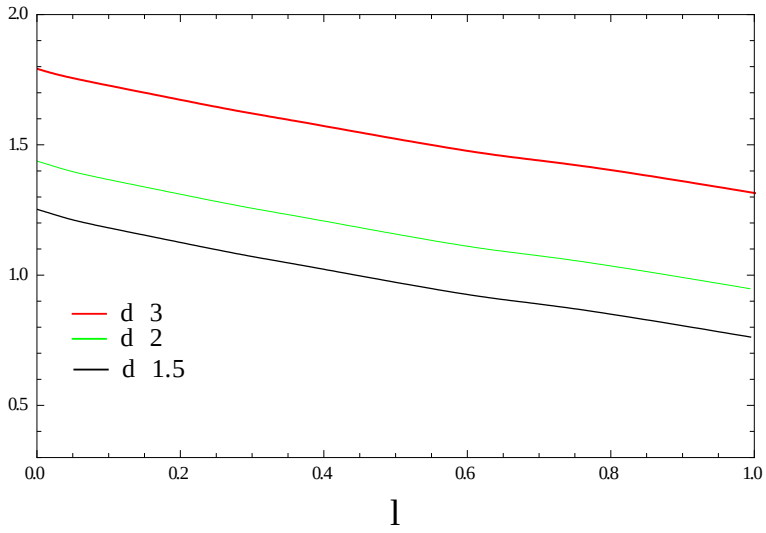


Figure 3. Effects of d on the equilibrium decisions



1

2 **Figure 4. Impact of l on the optimal repair time τ**

Appendix-A

Robustness checking of the M/M/N queuing and the N parallel M/M/1 queueing

In this section we conduct numerical studies using the real data of deteriorating industrial equipment inferred from Jackson and Pascual (2008). We use the data in Jackson and Pascual (2008) as base values of the parameters, and test the impact of changing parameter values around the base values. The detailed data settings are shown in Table A-1.

Table A-1 Parameter settings of the numerical studies

Parameters	Unit	Range	Base values
Waiting cost c_w	\$/Hour	40~80	60
Benchmark repair time τ_b	Hour	30~70	50
Promised total sojourn time d	Hour	40~100	70
Benchmark service value V_b	\$	$(700\sim800)*10^3$	$736.11*10^3$
Potential arrival rate Λ	1/Hour	$(0.02\sim0.1)*10^{-2}$	$0.08*10^{-2}$
Repairman travelling time (including spare parts lead time) l	Hour	0~30	20
Unit cost of expensive parts K	\$	$(4\sim12)*10^3$	$8*10^3$
Unit cost of ordinary parts k	\$	$(0.5\sim2)*10^3$	$1*10^3$
Share of revenue from spare parts consumption δ	--	0~1	--

Based on the data shown in Jackson and Pascual (2008), we relax the parameter base values to specific ranges in order to explore the impact of the parameters on the equilibrium decisions in different queuing structures. We take the average repair time (50 hours) in Jackson and Pascual (2008) as the benchmark repair time τ_b in our study. It is reasonable because 50 hours is associated to corrective actions in Jackson and Pascual (2008), which is defined as the minimum time that can restore the equipment to a working state. In addition, the allowed service duration in Jackson and Pascual (2008) is 70 hours; hence, we take the difference of these two durations as the base value of the repairman travelling time (including spare parts lead time). Finally, we use the customers' willingness to pay at the benchmark repair time (50 hours) in Jackson and Pascual (2008) as the benchmark service value V_b. For

α, β and ϕ, we let $\alpha = 2 * 10^3$ \$/Hour, $\beta = 0.3$ and $\phi = 0.5$ according to their definitions.

A.1 Changing the repairman travelling time

In this subsection, we examine that the results in Corollaries 1 and 3 remain robust in the *M/M/N* queuing setting. In the numerical example, we allow the repairman travelling time l to change from 0 to 30, while other parameters take base values in Table A-1.

Example 1. $\Lambda = 0.08 * 10^{-2}$, $V_b = 736.11 * 10^3$, $c_w = 60$, $\tau_b = 50$, $K = 8 * 10^3$, $k = 1 * 10^3$, $\delta = 0.7$, $\alpha = 2 * 10^3$, $\beta = 0.3$, and $\phi = 0.5$. The numerical results are shown in Figure A-1.

Observations. Figure A-1(a) confirms that in both the *M/M/N* queuing and *N* parallel *M/M/1* queuing settings, the optimal repair time τ increases in the repairman travelling time l. The intuition behind this result is clear: No matter the queueing structures, the service provider will enhance the service quality and reduce the spare parts consumption when the repairman travelling time increases, as illustrated in Corollaries 1 and 3. Moreover, from Figure A-1 (b) and (c), we can observe that changing the value of l has the same impacts on the service price and number of repairmen in both the *M/M/N* queuing and *N* parallel *M/M/1* queuing settings.

A.2 Changing the share of revenue from spare parts consumption δ

The share of revenue from spare parts consumption δ also affects the service equilibrium significantly, as shown in Corollaries 2(i) and 4(i). We conduct a numerical study to indicate if these analytical results remain true in the *M/M/N* queuing setting.

Example 2. $\Lambda = 0.08 * 10^{-2}$, $V_b = 736.11 * 10^3$, $c_w = 60$, $\tau_b = 50$, $l = 20$, $K = 8 * 10^3$, $k = 1 * 10^3$, $\alpha = 2 * 10^3$, $\beta = 0.3$ and $\phi = 0.5$. The numerical results are shown in Figure A-2.

Observations. Figure A-2 confirms that the analytical results in Corollaries 2(i) and 4(i) remain true in both *N* parallel *M/M/1* queuing and *M/M/N* queuing settings: As δ increases, the repair time, service price, and number of repairmen all decrease.

A.3 Changing the spare parts prices K and k

Corollaries 2(ii) and 4(ii) show the impact of spare parts prices K and k on the service provider's equilibrium decisions in the *N* parallel *M/M/1* queuing setting. In this subsection, we investigate whether these results remain robust in the *M/M/N* queuing setting.

Example 3. $\Lambda = 0.08 * 10^{-2}$, $V_b = 736.11 * 10^3$, $c_w = 60$, $\tau_b = 50$, $l = 20$, $\delta = 0.7$, $\alpha = 2 * 10^3$, $\beta = 0.3$ and $\phi = 0.5$. The numerical results are shown in Figure A-3.

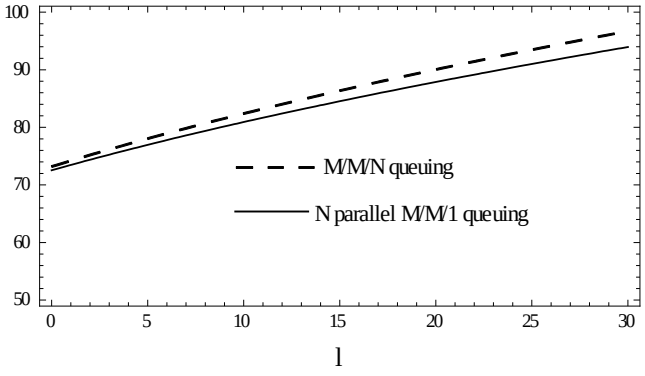
Observations. Figure A-3(a)~(c) reveal the impact of changing the value of expensive spare parts price K on the equilibrium repair time, service price, and number of repairmen, while Figure A-3(d)~(f) are associated with the ordinary spare parts price k. Our numerical study indicates that the results in Corollaries 2(ii) and 4(ii) remain true in the *M/M/N* queuing setting: If the spare parts prices K or k increases, then the equilibrium repair time increases, service price decreases, and number of repairmen increases.

A.4 Changing the repairman travelling time with maximum sojourn time constraint

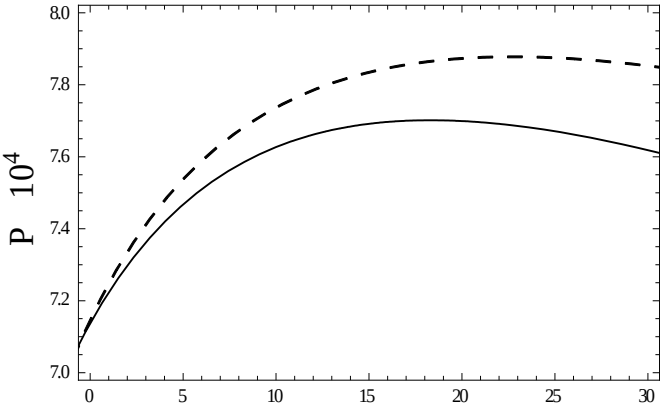
In Section 5, we stated that the optimal repair time τ and repairman travelling time are "complementary" without the maximum total sojourn time constraint, while they act as "substitutes" with the sojourn time constraint. In this subsection, we examine whether the above results remain true in the *M/M/N* queuing setting.

Example 4. $\Lambda = 0.08 * 10^{-2}$, $V_b = 736.11 * 10^3$, $c_w = 60$, $\tau_b = 50$, $d = 60$, $\delta = 0.7$, $\alpha = 2 * 10^3$, $\beta = 0.3$ and $\phi = 0.5$. The numerical results are shown in Figure A-4.

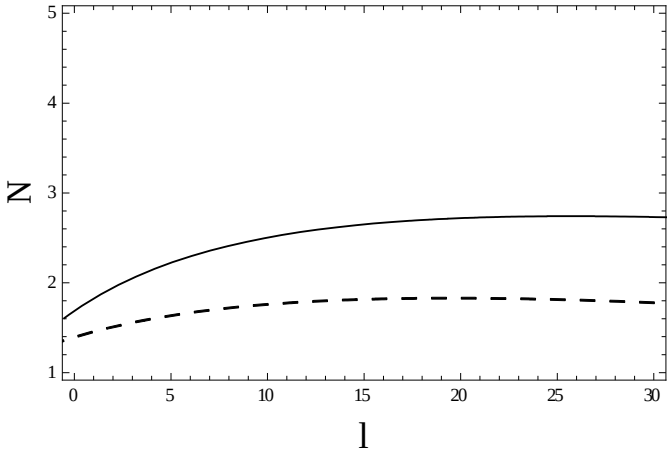
Observations. From Figure A-4, we note that in the presence of the maximum total sojourn time constraint, the repair time τ and repairman travelling time l are still "substitutes" in the *M/M/N* queuing in that the optimal repair time τ increases in the repairman travelling time l.



(a)

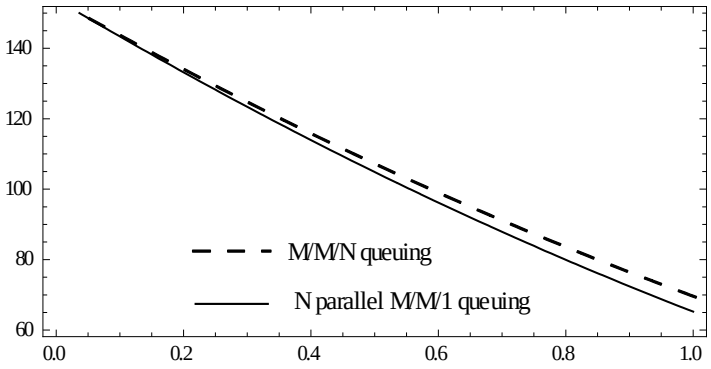


(b)

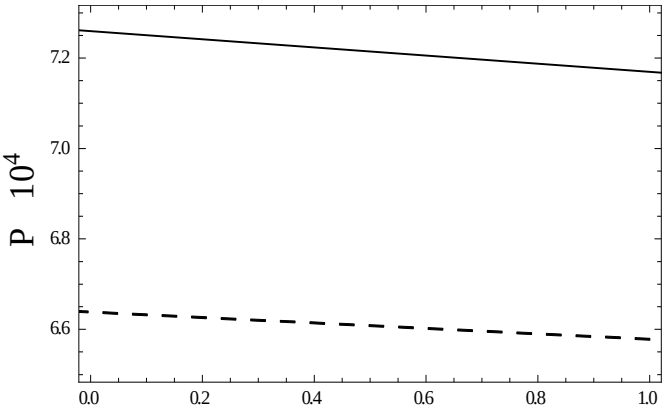


(c)

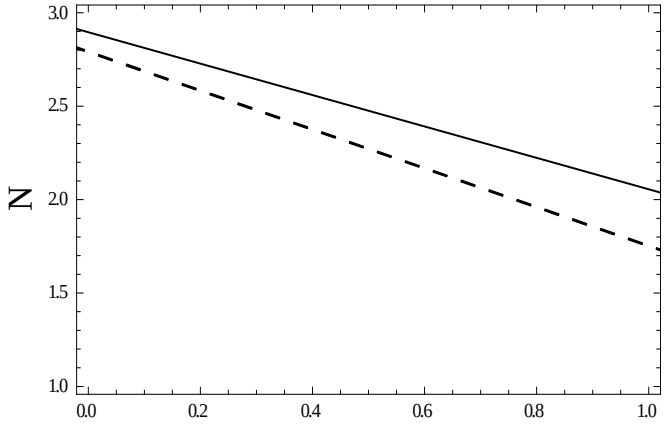
Figure A-1. Impact of l on the equilibrium decisions in different queuing structures



(a)



(b)



(c)

Figure A-2. Impact of δ on the equilibrium decisions in different queuing structures

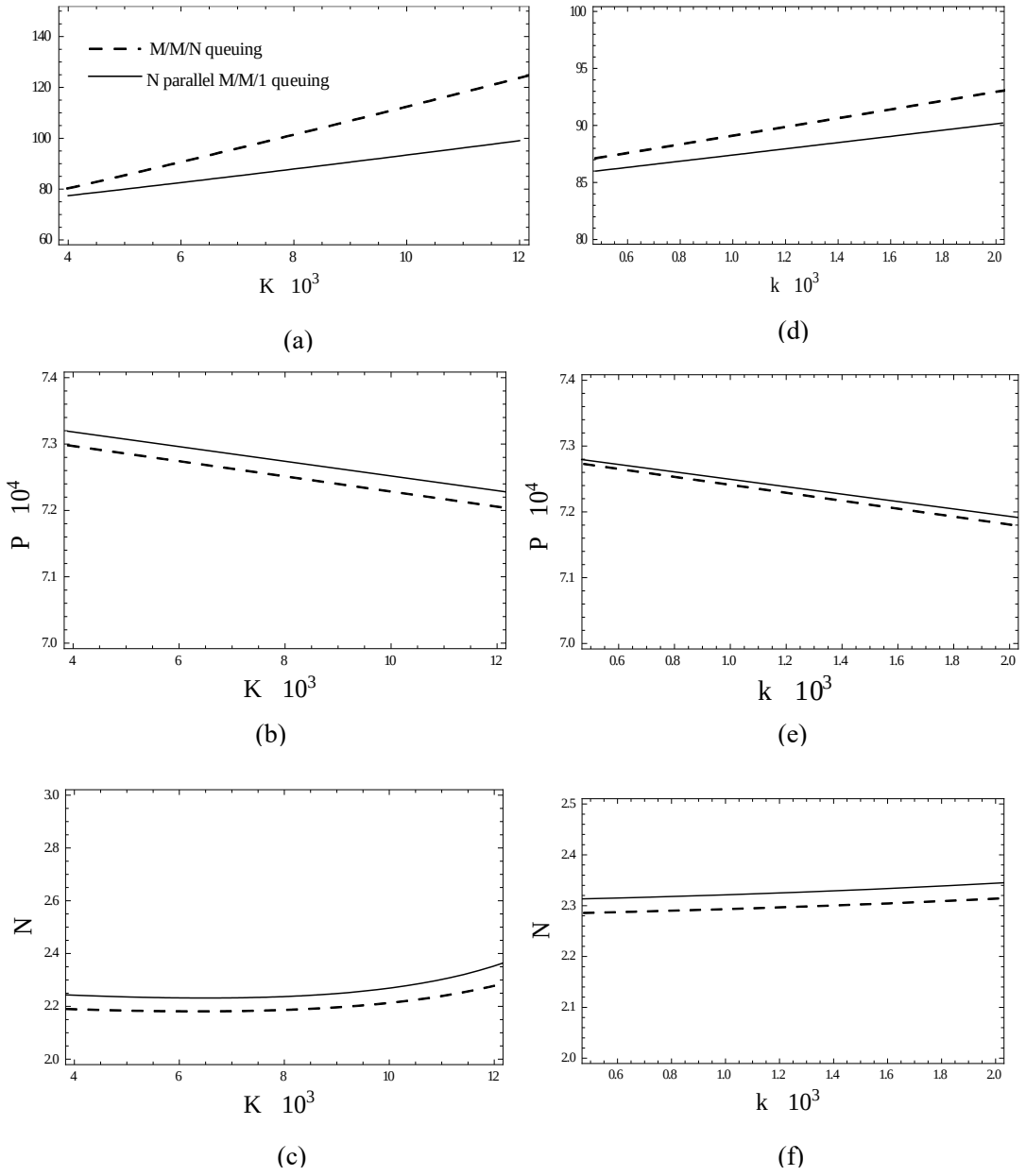


Figure A-3. Impact of $K(k)$ on the equilibrium decisions in different queuing structures

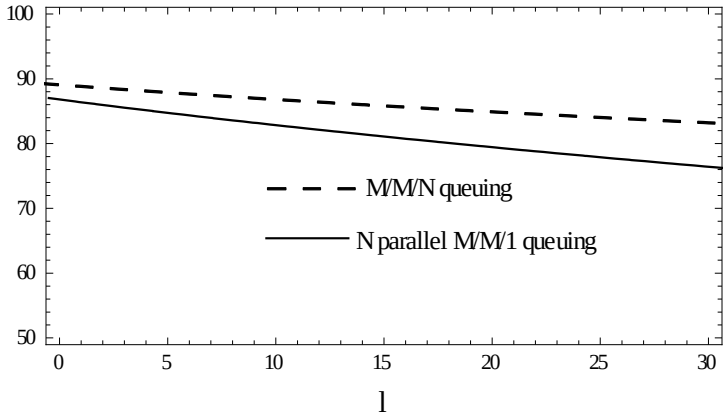


Figure A-4 Impact of l on the repair time in different queuing settings with sojourn time constraint

Appendix-B

Proof of Proposition 1

The expected utility for customers is:

$$U(N,P,\tau) = V(\tau) - c_w W(\lambda; N, l+\tau) - P - C(\tau).$$

Let $U(N,P,\tau) = 0$, we can get $\lambda(N,P,\tau) = W^{-1}\left(\frac{V(\tau)-P-C(\tau)}{c_w}\right)$. (Note that the waiting time W strictly increases in λ).

Case 1 (*Full participation*): In this case, the utility is positive even when all potential customers join the service. That is, when $\Lambda < \lambda(N,P,\tau)$, $V(\tau) - c_w W(\lambda; N, l+\tau) - P - C(\tau) > 0$ always holds. The customer unique equilibrium is $p(N,P,\tau) = 1$.

Case 2 *(No participation)*: In this case, the utility is negative for a customer even when no other customer joins the service. That is, when $V(\tau) - P - C(\tau) < c_w(\tau + l)$, no users will buy the service. The customer unique equilibrium is $p(N,P,\tau) = 0$.

Case 3 (*Partial participation*): This case falls into the above two conditions. The users will continuously join the service until the expected utility becomes zero. The mixed equilibrium strategy for the customers is $p(N,P,\tau) = \frac{\lambda(N,P,\tau)}{\Lambda} \in (0,1)$.

Proof of Proposition 2

The expected spare parts expense is $C(\tau) = \sum \frac{\xi^x}{x!} e^{-\xi} x\big(K\phi + k(1-\phi)\big)$, which equals to $C(\tau) = \left(1 + \beta\left(\frac{1}{\tau} - \frac{1}{\tau_b}\right)\right)\big(K\phi + k(1-\phi)\big)$ ultimately. Hence, the service provider's objective function becomes

$$\max_{P,\tau} R(P,\tau) = \left(P + \delta\left(1 + \beta\left(\frac{1}{\tau} - \frac{1}{\tau_b}\right)\right)\big(K\phi + k(1-\phi)\big)\right)\left(\frac{1}{l+\tau} - \frac{c_w}{V(\tau)-P-\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))}\right).$$

Let $q = P + \delta\left(1 + \beta\left(\frac{1}{\tau} - \frac{1}{\tau_b}\right)\right)\big(K\phi + k(1-\phi)\big)$. Then, the above objective function becomes

$$\max_{q,\tau} R(q,\tau) = q\left(\frac{1}{l+\tau} - \frac{c_w}{V(\tau)-q-(1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))}\right).$$

Further, we obtain

$$\frac{\partial R(P,\tau)}{\partial \tau} = \frac{\partial R(q,\tau)}{\partial \tau} + \frac{\partial R(q,\tau)}{\partial q}\frac{\partial q}{\partial \tau}$$

and

$$\frac{\partial R(P,\tau)}{\partial P} = \frac{\partial R(q,\tau)}{\partial q}\frac{\partial q}{\partial P} = \frac{\partial R(q,\tau)}{\partial q} \quad (\text{because } \frac{\partial q}{\partial P} = 1).$$

As a result, $\frac{\partial R(P,\tau)}{\partial \tau}=0$ and $\frac{\partial R(P,\tau)}{\partial P}=0$ are equivalent to $\frac{\partial R(q,\tau)}{\partial \tau}=0$ and $\frac{\partial R(q,\tau)}{\partial q}=0$. Therefore, the interior optimal solution of $R(P,\tau)$ is equivalent to the interior optimal solution of $R(q,\tau)$.

We solve the first order condition as follows.

$$\frac{\partial R(q,\tau)}{\partial q}=\frac{1}{l+\tau}-\frac{c_w\left[V(\tau)-(1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)\right]}{\left[V(\tau)-q-(1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)\right]^2}=0;$$

$$\frac{\partial R(q,\tau)}{\partial \tau}=q\left(-\frac{1}{(l+\tau)^2}+\frac{c_w\left(\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)\right)\frac{1}{\tau^2}}{\left[V(\tau)-q-(1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)\right]^2}\right)=0.$$

After some algebraic manipulations of the above two equations, we can obtain

$$\frac{l+\tau}{\tau^2}\left(\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)\right)=V(\tau)-(1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right).$$

Further, the equation is equivalent to

$$l\frac{1}{\tau^2}+2\frac{1}{\tau}-\left(\frac{V_b-(1-\delta)\left(K\phi+k(1-\phi)\right)}{\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)}+\frac{1}{\tau_b}\right)=0.$$

Solving the above equation, we obtain

$$\tau_1^{ns}=\frac{1+\sqrt{1+l\eta}}{\eta},$$

where $\eta=\frac{V_b-(1-\delta)\left(K\phi+k(1-\phi)\right)}{\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)}+\frac{1}{\tau_b}$. Substituting τ_1^{ns} into $\frac{\partial R(q,\tau)}{\partial \tau}=0$, and using the equation $q=P+\delta\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)$, we can get

$$P_1^{ns}=V_b+\frac{\alpha}{\tau_b}-\left(1-\frac{\beta}{\tau_b}\right)\left(K\phi+k(1-\phi)\right)-\left(\alpha+\beta\left(K\phi+k(1-\phi)\right)\right)\frac{\left(\sqrt{1+l\eta}-1\right)}{l}-$$

$$\sqrt{c_w(1+l\eta)}\left(\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)\right).$$

Further, the induced effective demand is

$$\lambda_1^*=\frac{\eta}{1+l\eta+\sqrt{1+l\eta}}-\sqrt{\frac{c_w}{\left(\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)\right)(1+l\eta)}}.$$

Next, we confirm that the above solution is unique for the optimization problem. For any given τ, we have

$\frac{\partial^2 R(q,\tau)}{\partial q^2}=-c_w\left[V(\tau)-(1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)\right]\left[V(\tau)-q-(1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)\right]^{-3}<0.$

That is, for any given τ, $R(q,\tau)$ is concave in q, and the optimal q is the solution of $\frac{\partial R(q,\tau)}{\partial q}=0$. In particular,

$$q(\tau) = V(\tau) - (1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi)) - \sqrt{c_w(l+\tau)\left[V(\tau) - (1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi))\right]}.$$

Substituting $q(\tau)$ into $R(q,\tau)$, we have

$$R(q(\tau),\tau) = \left[V(\tau) - (1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi)) - \right.$$

$$\left.\sqrt{c_w(l+\tau)\left[V(\tau) - (1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi))\right]}\right]\left(\frac{1}{\tau+l} - \sqrt{\frac{c_w}{(l+\tau)\left(V(\tau)-(1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))\right)}}\right).$$

The above equation can be further simplified as

$$R(q(\tau),\tau) = \left[\frac{l+\tau}{\tau^2}\left(\alpha + (1-\delta)\beta(K\phi + k(1-\phi))\right) - \frac{l+\tau}{\tau}\sqrt{c_w\left(\alpha + (1-\delta)\beta(K\phi + k(1-\phi))\right)}\right]\left(\frac{1}{\tau+l} - \right.$$

$$\left.\frac{\tau}{l+\tau}\sqrt{\frac{c_w}{(\alpha+(1-\delta)\beta(K\phi+k(1-\phi)))}}\right) = \left(\sqrt{\alpha + (1-\delta)\beta(K\phi + k(1-\phi))}\frac{1}{\tau} - \sqrt{c_w}\right)^2.$$

Clearly, $\frac{\partial^2 R(q(\tau),\tau)}{\partial \tau^2} < 0$. Therefore, $R(q(\tau),\tau)$ is concave in τ, and the above solution, τ_1^{ns} and P_1^{ns}, is the unique solution of the optimization problem.

Proof of Corollary 1

As $\tau_1^{ns} = \frac{1+\sqrt{1+l\eta}}{\eta}$ and $\eta = \frac{V_b-(1-\delta)(K\phi+k(1-\phi))}{\alpha+(1-\delta)\beta(K\phi+k(1-\phi))} + \frac{1}{\tau_b}$, it is obvious to obtain that τ_1^{ns} increases in l. The effective demand $\lambda_1^{ns} = \frac{\eta}{1+l\eta+\sqrt{1+l\eta}} - \sqrt{\frac{c_w}{\left(\alpha+(1-\delta)\beta(K\phi+k(1-\phi))\right)(1+l\eta)}}$, which can also be written as $\lambda_1^{ns} = \frac{1}{\sqrt{1+l\eta}}\left(\frac{\eta}{\sqrt{1+l\eta}+1} - \sqrt{\frac{c_w}{\left(\alpha+(1-\delta)\beta(K\phi+k(1-\phi))\right)}}\right)$. It is obvious that $\frac{\partial \lambda_1^{ns}}{\partial l} < 0$, which means that the effective demand decreases in l.

Proof of Corollary 2

Take K as an example to analyze the effects of parts price on the equilibrium. Taking the first order derivative of τ_1^{ns} with respect to K, we can obtain $\frac{\partial \tau_1^{ns}}{\partial K} = \frac{d\tau_1^{ns}}{d\eta}\frac{\partial \eta}{\partial K}$. With the expression of τ_1^{ns}, we can directly get $\tau_1^{ns} = \frac{1}{\eta} + \sqrt{\frac{1}{\eta^2} + l\frac{1}{\eta}}$. It is easy to obtain $\frac{d\tau_1^*}{d\eta} < 0$. In addition, according to the formulation of η, we can obtain

$$\frac{\partial \eta}{\partial K} = \frac{-(1-\delta)\left(\alpha+(1-\delta)\beta(K\phi+k(1-\phi))\right)-(1-\delta)\left(V_b-(1-\delta)(K\phi+k(1-\phi))\right)}{\left(\alpha+(1-\delta)\beta(K\phi+k(1-\phi))\right)^2}\varphi < 0.$$

Hence, $\frac{\partial \tau_1^{ns}}{\partial K} > 0$. The results regarding k can be obtained with the same process.

Similarly, we can get $\frac{\partial \tau_1^{ns}}{\partial \delta} = \frac{d\tau_1^{ns}}{d\eta}\frac{\partial \eta}{\partial \delta}$, and

$$\frac{\partial \eta}{\partial \delta}=\frac{(K\varphi+k(1-\varphi))\left(\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)\right)+(V_b-(1-\delta)K)\beta\left(K\phi+k(1-\phi)\right)}{\left(\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)\right)^2}>0.$$

Therefore, we can get $\frac{\partial \tau_1^{ns}}{\partial \delta}<0$.

Regarding the effective arrival rate $\lambda_1^{ns}=\frac{1}{\sqrt{1+l\eta}}\left(\frac{\eta}{\sqrt{1+l\eta}+1}-\sqrt{\frac{c_w}{\left(\alpha+(1-\delta)\beta(K\phi+k(1-\phi))\right)}}\right)$, we denote $(1-\delta)(K\phi+k(1-\phi))=x$. After some algebraic manipulations, the equilibrium demand is

$$\lambda_1^{ns}=\frac{1}{\sqrt{\left(1+\frac{l}{\tau_b}\right)(\alpha+\beta x)+l(V_b-x)}}\left[\frac{V_b-x+\frac{1}{\tau_b}(\alpha+\beta x)}{\sqrt{\left(1+\frac{l}{\tau_b}\right)(\alpha+\beta x)+l(V_b-x)}+\sqrt{(\alpha+\beta x)}}-\sqrt{c_w}\right].$$

To avoid the trivial case that the effective demand or spare parts consumption is negative, we assume $\lambda_1^{ns}>0$ and $\left(1+\beta\left(\frac{1}{\tau_1^{ns}}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))>0$. Thus, we know that $V_b-x+\frac{1}{\tau_b}(\alpha+\beta x)>0$. According to the expression of λ_1^{ns}, we note that if both the $\beta\left(1+\frac{l}{\tau_b}\right)-l>0$ and $\frac{\beta}{\tau_b}-1<0$ are satisfied, λ_1^{ns} decreases in x, i.e., $\frac{\partial \lambda_1^{ns}}{\partial x}<0$. Next, we prove that the two conditions hold under the assumption $\left(1+\beta\left(\frac{1}{\tau_1^{ns}}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))>0$. First, we can derive $\frac{1}{\tau_1^{ns}}\geq\frac{1}{\tau_b}-\frac{1}{\beta}$ from $\left(1+\beta\left(\frac{1}{\tau_1^{ns}}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))>0$. Substituting the expression of τ_1^{ns} into $\frac{1}{\tau_1^{ns}}\geq\frac{1}{\tau_b}-\frac{1}{\beta}$ and doing some algebraic manipulations, we can obtain $\beta\sqrt{1+l\eta}\leq\beta\left(1+\frac{l}{\tau_b}\right)-l$. Thus, $\beta\left(1+\frac{l}{\tau_b}\right)-l>0$ is always satisfied. Similarly, from $\frac{1}{\tau_1^{ns}}\geq\frac{1}{\tau_b}-\frac{1}{\beta}$, we can also derive that $\frac{\beta}{l}\left(\sqrt{1+l\eta}-1\right)\leq 1-\frac{\beta}{\tau_b}$. As a result, we can get $\frac{\beta}{\tau_b}-1<0$ is always satisfied. In summary, we get $\frac{\partial \lambda_1^{ns}}{\partial x}<0$.

Thus, taking the first order derivatives of λ_1^{ns} with respect to $K(k)$ and δ, respectively,

$$\frac{\partial \lambda_1^{ns}}{\partial K}=\frac{\partial \lambda_1^{ns}}{\partial x}\frac{\partial x}{\partial K}=\frac{\partial \lambda_1^{ns}}{\partial x}\varphi(1-\delta)<0;$$

$$\frac{\partial \lambda_1^{ns}}{\partial k}=\frac{\partial \lambda_1^{ns}}{\partial x}\frac{\partial x}{\partial k}=\frac{\partial \lambda_1^{ns}}{\partial x}(1-\varphi)(1-\delta)<0;$$

$$\frac{\partial \lambda_1^{ns}}{\partial \delta}=\frac{\partial \lambda_1^{ns}}{\partial x}\frac{\partial x}{\partial \delta}=\frac{\partial \lambda_1^{ns}}{\partial x}(-K)>0.$$

That is, the effective arrival rate decreases in $K(k)$ but increases in δ.

Proof of Proposition 3

We prove the proposition by contradiction. Assume that the optimal equilibrium arrival rate is $\lambda^e(N,\tau)<\Lambda$. Because we want to know the relationship between the number of servers

and equilibrium demand, we let the service price P and repair time τ be fixed.
The equilibrium demand is

$$\lambda(N,P,\tau)=\frac{N}{\tau+l}-\frac{c_w}{V(\tau)-P-\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)}.$$

Thus, $\frac{\partial\lambda^e}{\partial N}=\frac{1}{\tau+l}>0$.

Then, we define $v^e(N,\tau)=\frac{\lambda^e(N,\tau)}{N}$, taking the first order derivative with respect to N, we can get:

$$\frac{\partial v^e(N,\tau)}{\partial N}=\frac{N\frac{\partial\lambda^e}{\partial N}-\lambda^e}{N^2}=\frac{N\frac{1}{\tau+l}-\lambda^e}{N^2}=\frac{\frac{c_w}{V(\tau)-P-\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)}}{N^2}>0.$$

Now, we consider the objective function

$$\max_N R(N)=\left(P+\delta\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)\right)\lambda^e(N,P,\tau)-\theta N.$$

Taking the first order derivative with respect to N, we have

$$\frac{\partial R}{\partial N}=\frac{\partial\left(\left(P+\delta\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)\right)v^e-\theta\right)N}{\partial N}=\left[\left(P+\delta\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)\right)\frac{\partial v^e}{\partial N}\right]N+$$

$$\left(P+\delta\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)\right)v^e-\theta.$$

Because we consider the case that the service provider's profit $R>0$, we can obtain that $\left(P+\delta\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)\right)v^e-\theta>0$. Thus, as long as $\lambda^e(N,P,\tau)<\Lambda$, increasing N will strictly improve service provider's revenue. Since $\frac{\partial\lambda^e(N,P,\tau)}{\partial N}>0$, λ^e increases as N increases until $\lambda^e=\Lambda$. Once $\lambda^e=\Lambda$, increasing the number of servers only leads to an increase in the investment cost and the net revenue decreases. Therefore, it is optimal for the service provider to set the number of servers at the minimum value that can exactly attract all potential users into the system. Solving $V(\tau)-c_w\frac{1}{\frac{N}{\tau+l}-\Lambda}-P-\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)=0$, we have $N(P,\tau)=(\tau+l)\left[\Lambda+\frac{c_w}{V(\tau)-P-\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)\left(K\phi+k(1-\phi)\right)}\right]$.

Proof of Proposition 4

Substituting $N(P,\tau)$ into $R(N,P,\tau)$, the service provider's objective function becomes

$$\max_{P,\tau} R(P,\tau) = \left(P + \delta\left(1 + \beta\left(\frac{1}{\tau} - \frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi))\right)\Lambda - \theta(\tau + l)\left[\Lambda + \frac{c_w}{V(\tau) - P - \left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi))}\right].$$

We solve the service provider's revenue maximization problem in two steps. First, we find the optimal price $P(\tau)$ for a given repair time τ. Then, using $P(\tau)$, we find the revenue-maximizing repair time τ^*.

Step 1: Let $q = P + \delta\left(1 + \beta\left(\frac{1}{\tau} - \frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi))$, the objective function becomes

$$\max_{q,\tau} R(q,\tau) = q\Lambda - \theta(\tau + l)\left[\Lambda + \frac{c_w}{V(\tau) - q - (1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi))}\right].$$

According to the proof of Proposition 1, we can obtain that

$\frac{\partial R(P,\tau)}{\partial \tau} = 0$ and $\frac{\partial R(P,\tau)}{\partial P} = 0$ are equivalent to $\frac{\partial R(q,\tau)}{\partial \tau} = 0$ and $\frac{\partial R(q,\tau)}{\partial q} = 0$.

For given repair time τ, taking the first and second order derivatives of R with respect to q, we can get

$$\frac{\partial R}{\partial q} = \Lambda - \theta(\tau + l)\frac{c_w}{\left[V(\tau) - q - (1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi))\right]^2};$$

$$\frac{\partial^2 R}{\partial q^2} = -\theta(\tau + l)\frac{2c_w}{\left[V(\tau) - q - (1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi))\right]^3} < 0.$$

Thus, we can get the optimal solution $q(\tau)$ by solving the first order condition. Specifically,

$$q(\tau) = V(\tau) - (1-\delta)\left(1 + \beta\left(\frac{1}{\tau} - \frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi)) - \sqrt{\frac{c_w\theta(\tau+l)}{\Lambda}}.$$

Further,

$$P(\tau) = V(\tau) - \left(1 + \beta\left(\frac{1}{\tau} - \frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi)) - \sqrt{\frac{c_w\theta(\tau+l)}{\Lambda}}.$$

Then, substituting $P(\tau)$ into the objective function, we can get

$$\max_{\tau} R(\tau) = \left(V(\tau) - (1-\delta)\left(1 + \beta\left(\frac{1}{\tau} - \frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi))\right)\Lambda - 2\sqrt{c_w\Lambda\theta(\tau + l)} - \theta\Lambda(\tau + l).$$

Taking the first and second order derivatives of R with respect to τ, we have

$$\frac{\partial R}{\partial \tau} = \Lambda[\alpha + (1-\delta)\beta(K\phi + k(1-\phi))]\tau^{-2} - \sqrt{c_w\Lambda\theta}(\tau + l)^{-\frac{1}{2}} - \theta\Lambda;$$

$$\frac{\partial^2 R}{\partial \tau^2} = -2\Lambda[\alpha + (1-\delta)\beta(K\phi + k(1-\phi))]\tau^{-3} + \frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau + l)^{-\frac{3}{2}}.$$

Then, we confirm that $R(\tau)$ is maximized at τ_N^{ns}, where τ_N^{ns} can be solved by $\frac{\partial R}{\partial \tau} = 0$.

Note that $\frac{\partial^2 R}{\partial \tau^2} = -2\Lambda[\alpha + (1-\delta)\beta(K\phi + k(1-\phi))]\tau^{-3} + \frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau + l)^{-\frac{3}{2}} = -\frac{1}{\tau}\left\{2\Lambda[\alpha +\right.$

$(1-\delta)\beta(K\phi+k(1-\phi))]\tau^{-2}-\frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau+l)^{-\frac{1}{2}}\Big\}<-\frac{1}{\tau}\Big\{\Lambda[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))]\tau^{-2}-\sqrt{c_w\Lambda\theta}(\tau+l)^{-\frac{1}{2}}\Big\}$. When $\frac{\partial R}{\partial\tau}=0$, we have $\Lambda[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))]\tau^{-2}-\sqrt{c_w\Lambda\theta}(\tau+l)^{-\frac{1}{2}}=\theta\Lambda$. Thus, $\frac{\partial^2 R}{\partial\tau^2}<-\frac{1}{\tau}\theta\Lambda<0$. That is, the unique solution of τ_N^{ns} can be solved by the first order condition $\frac{\partial R}{\partial\tau}=0$. Substituting τ_N^{ns} into $P(\tau)$ and $N(P(\tau),\tau)$, we can obtain

$$P_N^{ns}=V_b+\alpha\left(\frac{1}{\tau_b}-\frac{1}{\tau_N^{ns}}\right)-\left(1+\beta\left(\frac{1}{\tau_N^{ns}}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))-\sqrt{\frac{c_w\theta(\tau_N^{ns}+l)}{\Lambda}};$$

$$N_N^{ns}=\Lambda(\tau_N^{ns}+l)+\sqrt{c_w\Lambda(\tau_N^{ns}+l)/\theta}.$$

Proof of Corollary 3

(i) We find the derivative of τ_N^{ns} with respect to l by utilizing the implicit function theory.

$$\Lambda[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))]\tau_N^{ns}\cdot(-2)\cdot\frac{\partial\tau_N^{ns}}{\partial l}=-\frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau+l)^{-\frac{3}{2}}\left(\frac{\partial\tau_N^{ns}}{\partial l}+1\right).$$

As a result,

$$\frac{\partial\tau_N^{ns}}{\partial l}=\frac{\frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau_N^{ns}+l)^{-\frac{3}{2}}}{2\Lambda[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))]\tau_N^{ns\,-3}-\frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau_N^{ns}+l)^{-\frac{3}{2}}}.$$

Moreover, we have

$$2\Lambda[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))]\tau_N^{ns\,-3}-\frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau_N^{ns}+l)^{-\frac{3}{2}}=-\frac{\partial^2 R}{\partial\tau^2}|_{\tau=\tau_N^{ns}}>0.$$

Therefore, we can obtain that $\frac{\partial\tau_N^{ns}}{\partial l}>0$. In addition, the upper bound of τ_N^{ns} can be obtained when $l\to\infty$, i.e., $\overline{\tau}=\sqrt{\alpha+(1-\delta)\beta(K\phi+k(1-\phi))/\theta}$. The lower bond $\overline{\tau}$ can be obtained when $l\to 0$, and can be obtained by solving

$$\Lambda\left[\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)\right]\tau^{-2}-\sqrt{c_w\Lambda\theta}\tau^{-\frac{1}{2}}-\theta\Lambda=0.$$

(ii) In equilibrium, the service price equals to

$$P_N^{ns}=V_b+\alpha\left(\frac{1}{\tau_b}-\frac{1}{\tau_N^{ns}}\right)-\left(1+\beta\left(\frac{1}{\tau_N^{ns}}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))-\sqrt{\frac{c_w\theta(\tau_N^{ns}+l)}{\Lambda}}.$$

Taking the first order derivative of P^* with respect to τ_0, we have

$$\frac{\partial P_N^{ns}}{\partial l}=\left(\alpha+\beta(K\phi+k(1-\phi))\right)\tau_N^{ns\,-2}\frac{\partial\tau_N^{ns}}{\partial l}-\frac{1}{2}\sqrt{\frac{c_w\theta}{\Lambda}}(\tau_N^{ns}+l)^{-\frac{1}{2}}\left(\frac{\partial\tau_N^{ns}}{\partial l}+1\right).$$

Substituting $\frac{\partial\tau_N^{ns}}{\partial l}$ into the above equation, we can finally get

$$\frac{\partial P_N^{ns}}{\partial l}=\frac{\frac{1}{2}\sqrt{c_w\Lambda\theta}\left(\alpha+\beta\left(K\phi+k(1-\phi)\right)\right)\tau^{-2}\left(\tau_N^{ns}+l\right)^{-\frac{3}{2}}-\frac{1}{2}\sqrt{c_w\Lambda\theta}\left(\tau_N^{ns}+l\right)^{-\frac{1}{2}}\left[\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)\right]\tau_N^{ns\,-3}}{2\Lambda\left[\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)\right]\tau_N^{ns\,-3}-\frac{1}{2}\sqrt{c_w\Lambda\theta}\left(\tau_N^{ns}+l\right)^{-\frac{3}{2}}}=$$

$$\frac{\frac{1}{2}\sqrt{c_w\Lambda\theta}\tau_N^{ns\,-3}\left(\tau_N^{ns}+l\right)^{-\frac{3}{2}}\left[\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)\right]\tau_N^{ns}\left[\frac{\left(\alpha+\beta\left(K\phi+k(1-\phi)\right)\right)}{\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)}-\frac{\tau_N^{ns}+l}{\tau_N^{ns}}\right]}{2\Lambda\left[\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)\right]\tau_N^{ns\,-3}-\frac{1}{2}\sqrt{c_w\Lambda\theta}\left(\tau_N^{ns}+l\right)^{-\frac{3}{2}}}.$$

Let $h(l)=\frac{\tau_N^{ns}+l}{\tau_N^{ns}}$, we can obtain $h'(l)=\frac{\tau_N^{ns}-l\frac{\partial\tau_N^{ns}}{\partial l}}{\tau_N^{ns\,2}}$. As τ_N^{ns} increases in l and has an upper threshold, we have $\frac{\partial^2\tau_N^{ns}}{\partial l^2}<0$. Because $\left(\tau_N^{ns}-l\frac{\partial\tau_N^{ns}}{\partial l}\right)'=\frac{\partial\tau_N^{ns}}{\partial l}-l\frac{\partial^2\tau_N^{ns}}{\partial l^2}>0$, we have $\left(\tau_N^{ns}-l\frac{\partial\tau_N^{ns}}{\partial l}\right)_{min}=\lim_{l\to0}\tau_N^{ns}-l\frac{\partial\tau_N^{ns}}{\partial l}>0$. Therefore, $h(l)$ increases in l, and $h(l)_{min}=\lim_{l\to0}h(l)=1$. Since $\frac{\left(\alpha+\beta\left(K\phi+k(1-\phi)\right)\right)}{\alpha+(1-\delta)\beta\left(K\phi+k(1-\phi)\right)}>1$, we can get that the value of $\frac{\partial P_N^{ns}}{\partial l}$ is first higher than 0 and then lower than zero, i.e., the optimal service price first increases and then decreases in l.

(iii) The result can be easily obtained from the expression of N_N^{ns}.

Proof of Corollary 4

(i) δ

Taking the first order derivative of τ_N^{ns} with respect to δ by utilizing the implicit function theory,

$$\Lambda\left\{-\beta(K\phi+k(1-\phi))\tau^{-2}+[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))](-2\tau^{-3})\frac{\partial\tau}{\partial\delta}\right\}=-\frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau+l)^{-\frac{3}{2}}\frac{\partial\tau_N^{ns}}{\partial\delta}.$$

As a result,

$$\frac{\partial\tau_N^{ns}}{\partial\delta}=\frac{-\Lambda\beta(K\phi+k(1-\phi))\tau^{-2}}{2\Lambda[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))]\tau^{-3}-\frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau+l)^{-\frac{3}{2}}}<0.$$

Then,

$$\frac{\partial P_N^{ns}}{\partial\delta}=\left(\alpha+\beta(K\phi+k(1-\phi))\right)\tau^{-2}\frac{\partial\tau_N^{ns}}{\partial\delta}-\frac{1}{2}\sqrt{\frac{c_w\theta}{\Lambda}}(\tau+l)^{-\frac{1}{2}}\frac{\partial\tau_N^{ns}}{\partial\delta}.$$

Since $\Lambda[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))]\tau^{-2}-\sqrt{c_w\Lambda\theta}(\tau+l)^{-\frac{1}{2}}-\theta\Lambda=0$ at equilibrium, we obtain:

$$\left(\alpha+\beta(K\phi+k(1-\phi))\right)\tau^{-2}-\frac{1}{2}\sqrt{\frac{c_w\theta}{\Lambda}}(\tau+l)^{-\frac{1}{2}}>[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))]\tau^{-2}-\frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau+l)^{-\frac{1}{2}}=\theta>0.$$

Therefore, we can get $\frac{\partial P_N^{ns}}{\partial\delta}<0$. In addition, $\frac{\partial N_N^{ns}}{\partial\delta}<0$ can also be obtained.

(ii) $K(k)$

We take K as an example to prove the results. Results regarding k can be obtained similarly. Taking the first order derivative of τ_N^{ns} with respect to K and utilizing the implicit function theory, we have

$$\Lambda\left\{(1-\delta)\beta\varphi\tau^{-2}+[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))](-2\tau^{-3})\frac{\partial\tau_N^{ns}}{\partial K}\right\}=-\frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau+l)^{-\frac{3}{2}}\frac{\partial\tau_N^{ns}}{\partial K}.$$

As a result,

$$\frac{\partial\tau_N^{ns}}{\partial K}=\frac{\Lambda(1-\delta)(K\phi+k(1-\phi))\tau^{-2}}{2\Lambda[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))]\tau^{-3}-\frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau+l)^{-\frac{3}{2}}}\varphi>0.$$

Then, we can obtain

$$\frac{\partial\tau_N^{ns}}{\partial k}=\frac{\Lambda(1-\delta)(K\phi+k(1-\phi))\tau^{-2}}{2\Lambda[\alpha+(1-\delta)\beta(K\phi+k(1-\phi))]\tau^{-3}-\frac{1}{2}\sqrt{c_w\Lambda\theta}(\tau+l)^{-\frac{3}{2}}}(1-\varphi)>0.$$

As a result, $\frac{\partial\tau_N^{ns}}{\partial K}>0$, $\frac{\partial\tau_N^{ns}}{\partial k}>0$; $\frac{\partial N_N^{ns}}{\partial K}>0$, $\frac{\partial N_N^{ns}}{\partial k}>0$.

Proof of Proposition 5

For given repair time τ and service price P, it is easy to know that the service provider's revenue decreases in the number of servers N. Therefore, the service provider will set N as small as possible. However, the service provider has to keep the expected waiting time $W(N,\tau)\le d$. In order to obtain the maximum revenue, the service provider will set N at the minimum value that can exactly make $W(N,\tau)=d$. Dealing with the above equation, we can get $N=\left[\Lambda\left(1-\frac{c_w(d-(\tau+l))}{V(\tau)-P-\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))-c_w(\tau+l)}\right)+1/d\right](\tau+l)$.

Proof of Proposition 6

Let $q=\left(P+\delta\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))\right)$. The optimization problem in (14) becomes

$$\max_{q,\tau}R(q,\tau)=q\left[1-\frac{c_w(d-(\tau+l))}{V(\tau)-q-(1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))-c_w(\tau+l)}\right]\Lambda-\theta\left[\Lambda\left(1-\frac{c_w(d-(\tau+l))}{V(\tau)-q-(1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))-c_w(\tau+l)}\right)+1/d\right](\tau+l).$$

In order to simplify the formulation, we let $B=V(\tau)-(1-\delta)\left(1+\beta\left(\frac{1}{\tau}-\frac{1}{\tau_b}\right)\right)(K\phi+k(1-\phi))-c_w(\tau+l)$. Taking the first order derivative of R with respect to q, we can obtain

$$\frac{\partial R}{\partial q}=\Lambda\left[1-\frac{c_w(d-(\tau+l))}{B-q}-q\frac{c_w(d-(\tau+l))}{(B-q)^2}\right]-\theta(\tau+l)\Lambda\frac{c_w(d-(\tau+l))}{(B-q)^2}=\Lambda\left(1-\frac{c_w(d-(\tau+l))}{B-q}\right)-$$

$$\Lambda[q-\theta(\tau+l)]\frac{c_w(d-(\tau+l))}{(B-q)^2}.$$

Then, taking the second order derivative of R with respect to q, we can obtain

$$\frac{\partial^2 R}{\partial q^2}=-\Lambda\frac{c_w(d-(\tau+l))}{(B-q)^2}-2\Lambda[q-\theta(\tau+l)]\frac{c_w(d-(\tau+l))}{(B-q)^3}<0.$$

Thus, the objective function R is concave in q for given repair time τ. The optimal solution $q^*(\tau)$ can be obtained by solving the first order condition. In particular,

$$\Lambda\left(1-\frac{c_w(d-(\tau+l))}{B-q}\right)-\Lambda[q-\theta(\tau+l)]\frac{c_w(d-(\tau+l))}{(B-q)^2}=0.$$

There are two solutions $B\pm\sqrt{c_w(d-(\tau+l))B}$. It is nonsense when $B<q$, because users will never obtain a positive utility and the effective demand always equals to 0. Therefore, we take the solution with the negative sign:

$$q^*(\tau)=B-\sqrt{c_w(d-(\tau+l))B}.$$

Substituting $q^*(\tau)$ into the objective function and letting $D=c_w(d-(\tau+l))$, the objective function becomes

$$\max_{\tau}R(\tau)=\Lambda(B-\sqrt{BD})\left(1-\sqrt{\frac{D}{B}}\right)-\theta(\tau+l)\left(\Lambda-\Lambda\sqrt{\frac{D}{B}}+\frac{1}{d}\right)=\Lambda\left(B^{\frac{1}{2}}-D^{\frac{1}{2}}\right)^2-$$

$$\theta(\tau+l)\left(\Lambda+\frac{1}{d}\right)+\theta\Lambda(\tau+l)D^{\frac{1}{2}}B^{-\frac{1}{2}}.$$

Taking the first order derivative of $R(\tau)$ with respect to τ, we can get

$$\frac{\partial R}{\partial\tau}=2\Lambda\left(B^{\frac{1}{2}}-D^{\frac{1}{2}}\right)\left(\frac{1}{2}B^{-\frac{1}{2}}\frac{\partial B}{\partial\tau}-\frac{1}{2}D^{-\frac{1}{2}}\frac{\partial D}{\partial\tau}\right)+\theta\Lambda\left[B^{-\frac{1}{2}}D^{\frac{1}{2}}+(\tau+l)\left(-\frac{1}{2}B^{-\frac{3}{2}}D^{\frac{1}{2}}\frac{\partial B}{\partial\tau}+\right.\right.$$

$$\left.\left.\frac{1}{2}B^{-\frac{1}{2}}D^{-\frac{1}{2}}\right)\frac{\partial D}{\partial\tau}\right]-\theta\left(\Lambda+\frac{1}{d}\right)=\frac{\partial B}{\partial\tau}\left[\left(1-B^{-\frac{1}{2}}D^{\frac{1}{2}}\right)\Lambda-\frac{1}{2}\theta\Lambda(\tau+l)B^{-\frac{3}{2}}D^{\frac{1}{2}}\right]-c_w\left[\left(1-\right.\right.$$

$$\left.\left.B^{\frac{1}{2}}D^{-\frac{1}{2}}\right)\Lambda+\frac{1}{2}\theta\Lambda(\tau+l)B^{-\frac{1}{2}}D^{-\frac{1}{2}}\right]+\theta\Lambda B^{-\frac{1}{2}}D^{\frac{1}{2}}-\theta\left(\Lambda+\frac{1}{d}\right).$$

As a result, the above equation can be written as

$$\frac{\partial R}{\partial\tau}=-\Lambda\left(B^{-\frac{1}{2}}D^{-\frac{1}{2}}+\frac{1}{2}\theta(\tau+\tau_0)B^{-\frac{3}{2}}D^{-\frac{1}{2}}\right)\left(\frac{\partial B}{\partial\tau}D-c_wB\right)+\frac{\partial B}{\partial\tau}+\theta\Lambda B^{-\frac{1}{2}}D^{\frac{1}{2}}-\theta\left(\Lambda+\frac{1}{d}\right)-$$

$$c_w\Lambda.$$

Next, we discuss that the above first order condition has a unique solution τ_N^S that maximizes the service provider's objective function. We first derive that $B^{-\frac{1}{2}}D^{\frac{1}{2}}$ decreases in τ.

$$\frac{\partial\left(B^{-\frac{1}{2}}D^{\frac{1}{2}}\right)}{\partial\tau}=\frac{\frac{1}{2}D^{-\frac{1}{2}}B^{\frac{1}{2}}\frac{\partial D}{\partial\tau}+\frac{1}{2}D^{\frac{1}{2}}B^{-\frac{3}{2}}\frac{\partial B}{\partial\tau}}{B}.$$

Since $\frac{\partial B}{\partial \tau} = \left[\alpha + (1-\delta)\beta\left(K\varphi + k(1-\varphi)\right)\right]\tau^{-2}$, $\frac{\partial D}{\partial \tau} = -c_w$, we can eventually get

$$\frac{\partial\left(B^{-\frac{1}{2}}D^{\frac{1}{2}}\right)}{\partial \tau} = \frac{-c_w(V(\tau)-c_w d)-D[\alpha+(1-\delta)\beta K]\tau^{-2}}{B}.$$

Because the utility function satisfies $V(\tau) - c_w d > 0$, we can prove that $B^{-\frac{1}{2}}D^{\frac{1}{2}}$ decreases in τ.

Then, we prove that $\left(\frac{\partial B}{\partial \tau}D - c_w B\right)$ increases in τ. Let $g(\tau) = \left(\frac{\partial B}{\partial \tau}D - c_w B\right)$, the first and second order derivatives are

$$g'(\tau) = -2[\alpha + (1-\delta)\beta(K\phi + k(1-\phi))]\tau^{-3}c_w\left(d - (\tau + l)\right) - 2c_w[\alpha + (1-\delta)\beta(K\phi + k(1-\phi))]\tau^{-2},$$

$$g''(\tau) = \frac{d\left(\frac{\partial^2 B}{\partial \tau^2}\right)}{d\tau}D - 2c_w\frac{\partial^2 B}{\partial \tau^2} < 0.$$

Thus, $g'(\tau)_{min} = g'(\tau)|_{\tau\to\infty} > 0$, and we can get $\left(\frac{\partial B}{\partial \tau}D - c_w B\right)$ increases in τ. As a result, $g(\tau)_{min} = g(\tau)|_{\tau=l} > 0$.

At last, we confirm $h(\tau) = \left(B^{-\frac{1}{2}}D^{-\frac{1}{2}} + \frac{1}{2}\theta(\tau + l)B^{-\frac{3}{2}}D^{-\frac{1}{2}}\right)$ also increases in τ.

Taking the first order derivative of $h(\tau)$, we obtain

$$h'(\tau) = -\frac{1}{2}B^{-\frac{3}{2}}D^{-\frac{1}{2}}\frac{\partial B}{\partial \tau} - \frac{1}{2}B^{-\frac{1}{2}}D^{-\frac{3}{2}}\frac{\partial D}{\partial \tau} + \frac{1}{2}\theta\left[B^{-\frac{3}{2}}D^{-\frac{1}{2}} + (\tau + l)\left(-\frac{3}{2}B^{-\frac{5}{2}}D^{-\frac{1}{2}}\frac{\partial B}{\partial \tau} - \frac{1}{2}B^{-\frac{3}{2}}D^{-\frac{3}{2}}\frac{\partial D}{\partial \tau}\right)\right].$$

As a result, we obtain $h'(\tau) = \left(\frac{\partial B}{\partial \tau}D - c_w B\right)\left[\frac{1}{2}B^{-\frac{3}{2}}D^{-\frac{3}{2}} + \frac{3}{4}\theta(\tau + l)B^{-\frac{5}{2}}D^{-\frac{3}{2}}\right] > 0$ (We have already obtained ($g(\tau) = \frac{\partial B}{\partial \tau}D - c_w B > 0$).

Because $h(\tau)$ and $g(\tau)$ are positive and increase in τ, it is easy to confirm that $h(\tau)\cdot g(\tau)$ increases in τ. Therefore, based on the above analysis, we can confirm that $\frac{\partial B}{\partial \tau}$ decreases in τ, and there exists a unique solution τ_3^* that maximizes the service provider's revenue. Correspondingly, the optimal service price, the equilibrium demand, and the number of servers are

$$P_N^S = B(\tau_N^S) - \delta\left(1 + \beta\left(\frac{1}{\tau_N^S} - \frac{1}{\tau_b}\right)\right)(K\phi + k(1-\phi)) - \sqrt{c_w\left(d - (\tau_N^S + l)\right)B(\tau_N^S)};$$

$$\lambda_N^S = \Lambda\left(1 - \sqrt{\frac{D(\tau_N^S)}{B(\tau_N^S)}}\right);$$

$$N_N^S = (\tau_N^S + l)\left(\Lambda - \Lambda\sqrt{\frac{D(\tau_N^S)}{B(\tau_N^S)}} + \frac{1}{d}\right).$$

Proof of Proposition 7

When considering reusable parts, the service provider's expected profit from each customer is $[r + (1-r)\delta]C(\tau)$. That is, the coefficient of $C(\tau)$ changes from δ to $r + (1-r)\delta$, which means that δ and $r + (1-r)\delta$ have the same effects on the equilibrium. In addition, we have $\frac{\partial[r+(1-r)\delta]}{\partial r} = 1 - \delta > 0$. Hence, we can obtain that the effects of r on the equilibrium is the same with δ.

Proof of Proposition 8

We prove by contradiction. For given P and τ, suppose that the optimal solution is (λ_1, N_1) with $\lambda_1 < \Lambda$. Then, if the service provider's profit is non-negative, we have $f(\lambda_1, N_1) = \big(P + \delta C(\tau)\big)\lambda_1 - \theta N_1 > 0$. Define $N_2 = mN_1(m > 1)$, the effective demand λ_2 can be obtained by $V(\tau) - P - C(\tau) = c_w W(\lambda_2, N_2)$. The exact expected waiting time in an M/M/N queue is $W(\lambda; N, \tau) = \frac{1}{1+(N!(1-\rho)/(N^N\rho^N))\sum_{i=0}^{N-1}(N^i\rho^i/i!)}\big(\rho/\lambda(1-\rho)\big)$, where $\rho = \lambda(\tau_0 + \tau)/N$. It is easy to show λ grows more than proportionately in N. Hence, in equilibrium, we have $\lambda_2 > m\lambda_1$. Thus, we have

$$f(\lambda_2, N_2) - f(\lambda_1, N_1) = \big(P + \delta C(\tau)\big)\lambda_2 - \theta N_2 - \big(P + \delta C(\tau)\big)\lambda_1 + \theta N_1 = \big(P +$$

$$\delta C(\tau)\big)(\lambda_2 - \lambda_1) - \theta(N_2 - N_1) > (m-1)\big[\big(P + \delta C(\tau)\big)\lambda_1 - \theta N_1\big] > 0.$$

Thus, increasing N can always improve $f(\lambda, N)$ until the effective demand λ reaches the potential demand rate Λ.

Proof of Proposition 9

Proof of Proposition 9 is similar with Proposition 5.