Joint optimization of component reliability and spare parts inventory for capital goods

Joint Optimization of Component Reliability and Spare Parts Inventory for Capital Goods

K.B. Öner*, G.P. Kiesmüller, G.J. van Houtum

Department of Technology Management, Eindhoven University of Technology,

P.O. Box 513, 5600 MB Eindhoven, the Netherlands

August 7, 2008

Abstract

We consider an OEM who is responsible for the availability of its systems in the field through service contracts. Upon a failure of a critical part in a system during the exploitation phase, the failed part is replaced by a ready-for-use part from a spare parts inventory. In an out-of-stock situation, a costly emergency procedure is applied. The reliability levels and spare parts inventory levels of the critical components are the two main factors that determine the downtime and corresponding costs of the systems. We introduce a quantitative model for the joint optimization of these two levels. We formulate the portions of Life Cycle Costs (LCC) which are affected by a component's reliability and its spare parts inventory level. These costs consist of design costs, production costs, and maintenance and downtime costs in the exploitation phase. We conduct exact analysis and provide an efficient optimization algorithm. In our numerical experiment which is based on real-life data, our method leads to significant cost reductions in comparison to a method that ignores costs in the exploitation phase when the reliability level is determined. We also show that the optimal reliability level also strongly depends on the component type (cheap or expensive), the size of installed base, the downtime penalty rate, and the lifetime of the system in our experiment.

Keywords: Capital goods, reliability, spare parts inventory, life cycle costs.

* *Corresponding author, e-mail: k.b.oner@tue.nl, tel: + 31 40 247 2505*

1. Introduction

Complex technical systems such as computer networks, defense systems, material handling systems, and medical systems are capital goods that are used in the primary processes of their users. Keeping the systems up in the field (availability of the systems) is crucial since operations of the users may halt due to failures of these systems. Operational interruptions lead to significant losses for the users and as the duration of the interruptions get longer, the losses may even grow more than linearly.

In general, the users ask for after-sales service from Original Equipment Manufacturers (OEMs) of the systems. The OEMs are willing to fulfill this demand because not only service provides advantage in sales, but also profit margins are higher in the service business. However, it is a challenge for a manufacturing oriented company to integrate service operations in their scope. Oliva and Kallenberg [23] propose a process model for the transition of orientation from manufacturing to service, which helps OEMs in changing their organization and processes.

The OEMs provide services to their customers (the users) through service contracts. The service contracts can be of two main types (see Kim *et al.* [12]): a material contract or a Performance Based Logistics (PBL) contract. Under a material contract, a customer pays the OEM for spare parts, labor, and other resources that are used during service activities, while a PBL contract contains a service level agreement with respect to the availability of the system(s) at the customer site. The material contracts have been the traditional means of service agreements. However, the main issue for the customers is the availability of their systems rather than the resources used for keeping them up. Therefore, PBL contracts are becoming more common and OEMs are increasingly forced to focus on availability management.

The availability of a system is determined by its reliability and the speed of system repair activities. The reliability of the system is primarily designated by the reliability of the critical parts; i.e., the parts that are critical for the functionality of the system. We make a distinction between a part and a component. We denote a physical item in a system

as a part and refer to its abstract representation (e.g. during design/development) as a component. Decisions on reliability levels of critical components and their realizations in corresponding parts compose one of the major design/development problems of the OEMs. From now on, we will use solely the term design to refer to both design and development.

During the exploitation phase of the systems, the speed of system repair activities is increased by the repair-by-replacement concept and keeping spare parts of the critical components on stock at a short distance of the installed systems. The activities to be executed upon a system failure due to a failure of a critical part depend on the status of the spare parts inventory of that component. If there is a ready-for-use part available from the inventory, the system is repaired by simply replacing the failed part with the ready-for-use one. In case of an out-of-stock situation, in general, more costly activities (emergency procedures) are carried out and the downtime will be longer.

The reliability level of a critical component and its spare parts inventory level play substitutional roles for the availability of the systems. The system availability can be improved by increasing the reliability level and/or by increasing the spare parts inventory level. In general, decisions on reliability level and the spare parts inventory are made by different departments of an OEM at different points in time. First the reliability level is decided, and later on the inventory level is determined. Design departments have the tendency to focus on design and production costs rather than all costs affected by the reliability decisions, such as maintenance and downtime costs. In general, there is a trade-off among these costs: Reliability improvement implies higher design and production costs and lower maintenance and downtime costs. The latter types of costs may be large, and hence ignoring them in design decisions may lead to suboptimal decisions; see Öner *et al.* [24] for a real-life case where downtime costs account for about 50% of Life Cycle Costs (LCC) of an engineer-to-order type of system.

In this paper, we present a quantitative model to support the integrated decision of the reliability level and the spare parts inventory level of a critical repairable component during the design of a system (capital good). We investigate a situation in which an OEM

will sell a number of units of the same system (capital good) together with a PBL contract which covers the life time of a system. The PBL contract specifies multiple service aspects among which a downtime penalty. That is, the OEM pays a certain amount of money to its customers per unit of downtime. The reliability level of a component is measured in terms of Mean Time Between Failures (MTBF). The systems are installed in one region that is served by a single spare parts inventory stock point which is at a sufficiently close distance from all systems. Our objective is the minimization of the portion of the system's LCC which is affected by the component's MTBF and the spare parts inventory level.

A service contract can be considered as a long-term warranty. Normally, a system (product) is sold together with a base warranty and a customer can obtain an additional warranty period against a premium payment. Warranties have different aspects in terms of management, marketing, engineering, logistics and accounting. As a consequence of these various aspects, warranties (in particular base warranties) have been investigated by researchers from different fields (see Blischke amd Murthy [3] for a compilation). Blischke and Murthy [2, 17, 18] provide an extensive review about the studies conducted on warranty till 1992. The review by Murthy and Djamaludin [19] covers the later period till 2002. A number of quantitative models has been developed for warranty cost analysis (see [18], [1]). These models have different aspects with respect to warranty policies (see [2] for a taxonomy for warranty policies), the viewpoint taken (OEM's or customer's), cost elements included, whether the items are repairable or not, etc. In general, the length of the warranty period is a decision variable in these models. A general lifetime distribution for items is given and failures throughout the warranty period are modeled as renewal processes. Costs are derived through cost parameters and formulations obtained from the renewal processes. These models were mostly developed for base warranties, however, they also have been used as a basis for long-term warranties (see Murthy and Djamaludin [19], and Rahman and Chattopadhyay [26]). Chattopadhyay and Rahman [4] examine lifetime warranty, which is a form of long-term warranty. Through a number of practical examples, they show that the definition of lifetime is not clear and changes from case to case. They define the lifetime as a

random variable and develop models for predicting failures and estimating costs throughout the whole lifetime.

Since warranty costs depend on the reliability of systems, reliability level decisions are also studied in warranty literature. The reliability of a system can be improved in two main ways:

(i) Through redundancy. Instead of a single critical part, a module with a number of parallel parts is used.

(ii) Through a reliability improvement process (a test-redesign cycle) during design.

In both cases, design and production costs increase while warranty costs decrease. In the first case, the reliability level of a part is fixed and the objective is to find the number of parallel parts in a module that minimizes total costs. Hussain and Murthy [7, 8, 20] present models for simplified cases in which a system is composed of only one critical part. Nguyen and Murthy [22] and Monga and Zuo [15] develop models for certain system designs. In the second case, the reliability level of a system is a decision variable and it is formulated as a function of time spent on reliability improvement during design. The objective is to find the length of time to be spent on reliability improvement to minimize the total cost (see Hussain and Murthy [9]). Huang *et al.* [6] propose the joint consideration of the reliability level of a system, warranty decisions, and price for new products. They develop a model in which the time spent on reliability improvement, the duration of the warranty period and the price of a system (product) are decision variables.

In a recent paper, Murthy *et al.* [21] highlight the current issues and challenges in product warranty logistics. They underline the need for linking the spare parts inventory levels to failures of parts, i.e., to component reliability. To the best of our knowledge, so far, this linkage has only been studied in recent work by Kim *et al.* [12, 11]. In these papers, game-theoretic models are introduced for the comparison of certain service contract types. In [12], the reliability level is incorporated into the model explicitly and the trade-off between investing in reliability and investing in spare parts is evaluated, while the reliability level is indirectly included in the model in [11].

The situations considered in [11, 12] do not involve any emergency procedures. That is, when a part in a system fails and there is no ready-for-use part available from a warehouse, the system is down till a part is available from a repair facility, meaning that downtime can be considerably long. Further, to simplify the analysis, they use the normal approximation for the leadtime demand. We develop a mathematical model, where we assume that a certain emergency procedure is applied when no spare part is available to replace a failed part. This aspect is important for capital goods as it has a large impact on downtimes and downtime costs. Incorporation of this aspect leads to a more complex model than the models in [11, 12]. We provide an exact analysis (without usage of the normal approximation) for the LCC function and we derive an exact optimization procedure.

The contribution of this paper can be stated as follows. First, we propose a new quantitative model for the joint optimization of the reliability level and spare parts inventory level of a critical component. In this model, we incorporate design costs, production costs and service costs (including downtime costs). Second, we perform an exact analysis on the LCC and we derive several analytical properties of it. Third, we provide an efficient optimization algorithm. Finally, we conduct a numerical experiment based on real-life data, and compare costs obtained under our joint optimization method to costs obtained via a non-integrated method. In our experiment, we show that the joint optimization leads to an average cost reduction of 52% and the optimal reliability level significantly depends on component type, the size of the installed base, the downtime penalty rate, and the lifetime of the system.

The outline of the paper is as follows. We present our model in Section 2. In Section 3, we derive the LCC function and provide a number of analytical properties and an optimization procedure. We give the setting and the results of our numerical experiment in Section 4. In Section 5, we draw conclusions and give directions for future research.

2. Model

An OEM is designing/developing a critical repairable component for a capital good (system). After the design of the capital good (system), the OEM expects to produce and sell

N units of the same system ($N \in \mathbb{N} = \{1, 2, 3, ...\}$). Each unit of the system will have a single part which is an embodied unit of this component. The OEM will support these systems throughout their exploitation phase which is considerably longer than the design and production phases (e.g. at least ten times longer).

During the exploitation phase of the systems, when the part in a system fails, the system will be repaired by replacement of the failed part with a ready-for-use one. Thus, a number of parts will be stocked as spare parts.

We assume that the N systems will be sold at the same time. We further assume that all the systems have an exploitation phase of length T. The exploitation phase is denoted by the time interval $[0, T]$.

In general, failures of the repairable parts in the systems will depend on the parts' quality, how the systems will be used and the usage conditions (e.g. environmental conditions). We assume that no systematic failure will occur due to deficiencies in production and the systems, and, so the parts, will satisfy all quality specifications at the beginning of their lives. We assume that the systems (so the parts) will be used in a convenient manner under convenient conditions as well. We denote the MTBF of the component by τ. We assume that failures of each part will occur according to a Poisson process with rate $1/\tau$ during the exploitation phase and τ is fixed throughout $[0, T]$. τ is a decision variable to be fixed during the design of the component. We assume that there is a baseline MTBF value $\underline{\tau} > 0$ that the OEM will provide anyway. We also assume that there is a limit $\bar{\tau}$ that the MTBF can be improved up to. So, τ is bounded from below and above by $\underline{\tau}$ and $\bar{\tau}$, respectively.

The systems will be supported by a single warehouse where all spare parts are stocked. There is a single repair facility where defective parts will be repaired and it is co-located with the warehouse.

As we assume that failures of a part will occur according to a Poisson process with rate $1/\tau$, the total stream of system failures due to failures of the parts in them will follow a Poisson process with rate N/τ. When a system at a customer site fails, the failure will be (remotely) diagnosed with 100% accuracy. In such a case, one of the following two repair

procedures will be applied depending on the availability of a part from the spare parts inventory:

1. Ordinary Repair Procedure: If a ready-for-use part is available from the inventory, it will be transported to the customer site. A service engineer will visit the customer site and repair the system by replacing the defective part with the ready-for-use one. The defective part will be transported to the repair facility for a *normal repair*. After repair, the part will be added to the spare parts inventory.
2. Emergency Repair Procedure: If there is an out-of-stock situation, a service engineer will visit the customer site and take out the defective part. The defective part will be transported to the repair facility for a *fast repair*. When the repair is over, the repaired part will be transported back to the customer site and a service engineer will visit the customer site again for the installation of the component.

We assume that after any type of repair, the repaired component will attain its quality and reliability level that it had when it was new.

A system suffers from a random downtime with mean t_1 or t_2 when it undergoes an ordinary repair procedure or an emergency repair procedure, respectively. The downtime includes the time elapsed from the instant the failure occurs to the instant the system is up again. Thus it is equivalent to the system repair time. We assume that $0 < t_1 \leq t_2$.

The result of the repair procedures is that the spare parts inventory position (the summation of pipeline stock and stock on hand) will be constant and will be equal to the initial amount of spare parts that one stocks in the warehouse. We denote the inventory position by s. We assume that the time to diagnose that the part in a system has failed upon a system failure will be negligible. Then, a demand for a component from the inventory will occur when a component fails and the inventory level will be affected only when the failure leads to an ordinary system repair. When an ordinary system repair procedure is initiated, the failed part will be added to the inventory after a random lead time which we call the replenishment time. The replenishment time includes the time form failure till the failed part is replaced with a ready-for-use one (i.e., the downtime), the time to transport the

failed part to the repair facility and the normal repair time. We assume that replenishment times are independent and identically distributed with mean $L > 0$ which is constant over time. Then, the situation for the spare parts inventory is equivalent to an inventory system with lost sales in which a base stock policy with a base stock level of s is followed (base-stock policies are commonly used in inventory management for spare parts; see Muckstadt [16] and Sherbrooke [27]). We assume that the expected time spent for an emergency repair (t_2) is much shorter than L.

The inventory position s is a decision variable which is to be fixed concurrently with τ during design. τ and s have certain effects on certain costs that contribute to the Life Cycle Costs (LCC) of the N systems:

- The reliability level τ affects design costs of the repairable component, production costs of the repairable parts that will be installed in the N systems or kept in spare parts stock, costs of system repairs and downtime costs that stem from the failures of the repairable components.
- The inventory position s affects production costs of the repairable parts that will be kept in spare parts stock, costs of system repairs and downtime costs that stem from the failures of the repairable parts, and spare parts storage costs.

τ and s do not affect any other costs that contribute to the LCC of the N systems. We assume that design costs and production costs are incurred at time 0 and formulate the Net Present Values (NPVs) of the other costs, which occur throughout $[0, T]$, at time 0. We denote the discount rate by $\alpha > 0$; a cost of 1 at time t contributes $e^{-\alpha t}$ to the NPV (notice that $\alpha = 0$ would correspond to no discounting). We assume that the stock-on-hand process is in steady state from the beginning.

We use the following notation to refer to cost parameters and the costs affected by τ and s:

h: The storage cost rate per part ($h > 0$).

r_1: Expected cost of an ordinary repair.

r_2: Expected cost of an emergency repair.

p: Downtime penalty rate ($p > 0$).

d_1: Expected downtime penalty incurred because of a failure leading to an ordinary repair.

d_2: Expected downtime penalty incurred because of a failure leading to an emergency repair.

$\pi(\tau, s)$: The expected NPV of the LCC affected by the of a part, τ, and the spare parts stock level of the part s.

$K(\tau)$: The expected NPV of the design costs of the component.

$P(\tau)$: The expected NPV of the production costs of the parts that will be installed in the N systems.

$S(\tau, s)$: The expected NPV of the spare parts costs

$R(\tau, s)$: The expected NPV of the system repair costs that stem from the failures of the repairable parts.

The factors r_1 and r_2 include all costs originating from the corresponding repair procedure, such as administrative costs, costs of one or more visits of a service engineer, transportation costs, repair costs of a failed part, and storage costs during the repair lead time at the repair facility. We assume that $hL \leq r_1 \leq r_2$, as emergency repairs require more expensive activities than ordinary repairs do and r_1 and r_2 include storage costs at the repair facility. We also assume that r_1 and r_2 are immediately incurred when a failure occurs.

Upon a system failure, an average downtime cost of $d_1 = pt_1$ is incurred if an ordinary repair is performed and an average downtime cost of $d_2 = pt_2$ is incurred if an emergency repair is performed. We assume that downtime costs are immediately incurred when a failure occurs. As $t_1 \leq t_2$, $d_1 \leq d_2$.

Since $\underline{\tau}$ is the baseline reliability that the manufacturer has to provide, we define the function $K(\tau)$ for the extra design cost that would be incurred to improve reliability to τ, $\underline{\tau} \leqslant \tau \leqslant \bar{\tau}$. Thus, $K(\underline{\tau}) = 0$. Design costs of a component can be derived by analyzing data of previous versions of the component or data of a similar component. In general, design costs are assumed to be an increasing convex function of the reliability level (see Mettas *et*

al. [14] and Kim *et al.* [12]). We also assume that $K(\tau)$ is an increasing convex function of τ.

Production costs include all the costs incurred for the production of these N components that will be installed in the systems. The production cost per part is $c(\tau)$ which is an increasing convex function of τ. Then $c(\underline{\tau})N$ is the baseline production cost. This fixed amount will be invested by the manufacturer regardless of the choice τ. Thus, we include $P(\tau) = \Big[c(\tau) - c(\underline{\tau})\Big] N$ in our model. This is the extra production costs that is incurred when N parts are produced with an MTBF of τ.

The LCC function is

$$\pi(\tau, s) = K(\tau) + P(\tau) + S(\tau, s) + R(\tau, s) + D(\tau, s) \tag{1}$$

and our problem formulation is

$$\begin{aligned} \text{(P)} \qquad & \min \quad \pi(\tau, s) \\ & \text{s.t.} \quad \underline{\tau} \leq \tau \leq \bar{\tau} \\ & \qquad\quad s \in \mathbb{N}_0 = \{0, 1, 2, ...\}. \end{aligned}$$

Remark 1. *In our model, if we drop the assumption that the failures of each part follow a Poisson process with rate $1/\tau$, we do not get a Poisson process for the total stream of system failures. However, relaxing Poisson assumption for the parts and assuming that the total stream of system failures follow a Poisson process with rate N/τ will not change our model. The Poisson assumption of the total stream of system failures, independent of the process followed by the part failures, is justified when the number of the systems (N) is large.*

3. Analysis

In this section, we first give the formulations of the spare parts costs, the repair costs, and the downtime costs. Next, we derive a number of analytical properties of the LCC function. We finalize the section by providing an optimization algorithm based on those analytical properties.

The average spare parts inventory, the number of ordinary and emergency repairs and the downtime throughout $[0, T]$ depend on the out-of-stock probability of the spare parts inventory. This out-of-stock probability is denoted by $g(\tau, s)$ and we start by determining this function.

It follows from our assumptions that the stock-on-hand process of the spare parts inventory is identical to the process for the number of free servers in an Erlang loss system (also denoted as the $M/G/s/s$ queueing system) with an arrival rate N/τ, mean service time L, and s servers. As a result, $g(\tau, s)$ is equal to the Erlang loss probability, and we obtain

$$g(\tau, s) = \frac{\frac{(NL/\tau)^s}{s!}}{\sum_{i=0}^{s} \frac{(NL/\tau)^i}{i!}}. \tag{2}$$

Later on we will exploit the following property of $g(\tau, s)$.

Property 1. *$g(\tau, s)$ is strictly decreasing and strictly convex in τ.*

Proof. Our formulation of the Erlang loss probability $g(\tau, s)$ is mathematically equivalent to the Erlang loss probability given by Harel [5] with $\lambda = NL$ and $\mu = \tau$, where λ is the arrival rate and μ is the service rate. In [5], Harel shows that the Erlang loss probability is strictly decreasing and strictly convex in μ. This is equivalent to $g(\tau, s)$ being strictly decreasing and strictly convex in τ. □

The spare parts costs $S(\tau, s)$ are the sum of spare parts investment costs, $S_1(\tau, s) = c(\tau)s$, and spare parts storage costs, $S_2(\tau, s)$. The formulations of $S_2(\tau, s)$, repair costs $R(\tau, s)$, and downtime costs $D(\tau, s)$ are given in Lemma 2 below. After that, in Lemma 3 and Lemma 4 we provide the monotonicity properties of $S(\tau, s)$, $R(\tau, s)$, and $D(\tau, s)$. In the proof of Lemma 2, we will exploit the property stated in Lemma 1, which therefore is presented first.

Lemma 1. *The numbers of ordinary repairs and emergency repairs performed throughout $[0, T]$ have Poisson distributions with means $[1 - g(\tau, s)](N/\tau)T$ and $g(\tau, s)(N/\tau)T$, respectively.*

Proof. When a part fails, the probabilities that an ordinary repair or an emergency repair is performed are constant, and are $[1-g(\tau,s)]$ and $g(\tau,s)$, respectively. Then, as the total stream of system failures is a Poisson process with rate $(N/\tau)T$, the total stream of ordinary repairs and that of emergency repairs are Poisson processes with rates $[1-g(\tau,s)]N$ and $g(\tau,s)N$, respectively. □

Lemma 2. *It holds that:*

(i)

$$S_2(\tau,s) = \frac{h}{\alpha}\left(1-e^{-\alpha T}\right)\left[s - \frac{NL}{\tau} + \frac{NL}{\tau}g(\tau,s)\right]. \tag{3}$$

(ii)

$$R(\tau,s) = [1-g(\tau,s)]\frac{N}{\tau}\frac{r_1}{\alpha}\left(1-e^{-\alpha T}\right) + g(\tau,s)\frac{N}{\tau}\frac{r_2}{\alpha}\left(1-e^{-\alpha T}\right).$$

(iii)

$$D(\tau,s) = [1-g(\tau,s)]\frac{N}{\tau}\frac{d_1}{\alpha}(1-e^{-\alpha T}) + g(\tau,s)\frac{N}{\tau}\frac{d_2}{\alpha}(1-e^{-\alpha T}).$$

Proof. See Appendix. □

By Lemma 2, we can write equation (1) as

$$\begin{aligned}\pi(\tau,s) &= K(\tau) + \left[c(\tau) - c(\underline{\tau})\right]N + c(\tau)s + \frac{h}{\alpha}\left(1-e^{-\alpha T}\right)\left[s - \frac{NL}{\tau} + \frac{NL}{\tau}g(\tau,s)\right] \\ &\quad + [1-g(\tau,s)]\frac{N}{\tau}\frac{r_1+d_1}{\alpha}\left(1-e^{-\alpha T}\right) + g(\tau,s)\frac{N}{\tau}\frac{r_2+d_2}{\alpha}\left(1-e^{-\alpha T}\right). \end{aligned} \tag{4}$$

Lemma 3. *For a fixed value of s, the following monotonicity properties hold:*

(i) $S(\tau,s)$ is strictly increasing in τ for $s>0$; $S(\tau,s)=0$ for $s=0$.

(ii) $R(\tau,s)$ and $D(\tau,s)$ are strictly decreasing in τ.

Proof. See Appendix. □

Recall the assumption that $K(\tau)$ and $c(\tau)$ are increasing in τ. Together with this assumption, Lemma 3 reflects the conflicting behavior of the costs included in $\pi(\tau,s)$ for varying reliability levels.

Lemma 4. *For a fixed value of τ, the following monotonicity properties hold:*

(i) *$S(\tau, s)$ is strictly increasing in s.*

(ii) *$R(\tau, s)$ and $D(\tau, s)$ are decreasing in s.*

Proof. See Appendix. □

Lemma 4 shows that there is also a trade-off among costs for varying spare parts level. In Lemma 5, we give the properties of $\pi(\tau, s)$ that we exploit for it optimization

Lemma 5. *$\pi(\tau, s)$ has the following properties:*

(i) *For a fixed $\tau \in [\underline{\tau}, \bar{\tau}]$, $\pi(\tau, s)$ is strictly convex in s.*

(ii) *For all $\tau \in [\underline{\tau}, \bar{\tau}]$, $\lim_{s\to\infty} \pi(\tau, s) = \infty$.*

(iii) *For a fixed $s \in \mathbb{N}_0$, $\pi(\tau, s)$ is strictly convex in τ.*

(iv) *For a fixed $\tau \in [\underline{\tau}, \bar{\tau}]$, define $s^*(\tau) = \min\{\arg\min_{s\in\mathbb{N}_0} \pi(\tau, s)\}$, i.e., $s^*(\tau)$ is the smallest value of s under which $\pi(\tau, s)$ is minimized. Then, $s^*(\tau)$ is decreasing in τ.*

Proof. See Appendix. □

(i) and (ii) in Lemma 5 imply that $s^*(\tau)$ is finite for a fixed value of τ and can be found by standard procedures for optimization in one variable. Let (τ^*, s^*) be a minimizer of $\pi(\tau, s)$. By (iv), $s^*(\bar{\tau}) \leq s^* \leq s^*(\underline{\tau})$. (iii) implies that we can also find the optimal value of τ for a fixed value of s, which we denote by $\tau^*(s)$. Then, an optimal solution (τ^*, s^*) can be found by enumerating all solutions $(\tau^*(s), s)$ for $s^*(\bar{\tau}) \leq s \leq s^*(\underline{\tau})$.

Theorem 1. *The following procedure determines an optimal solution of of problem (P):*

1. *Find $s^*(\bar{\tau})$ and $s^*(\underline{\tau})$.*

2. *For each $s = s^*(\bar{\tau}), s^*(\bar{\tau})+1, ..., s^*(\underline{\tau})$, solve the problem $\{min\, \pi(\tau, s),\ s.t.\ \underline{\tau} \leq \tau \leq \bar{\tau}\}$. Let $\tau^*(s)$ be an optimal τ for a given s.*

3. *$(\tau^*, s^*) = \arg\min_{(\tau^*(s),s)} \{\pi(\tau^*(s), s),\ s = s^*(\bar{\tau}), s^*(\bar{\tau}) + 1, ..., s^*(\underline{\tau})\}$ is an optimal solution and $\pi^* = \pi(\tau^*, s^*)$ is the corresponding minimum LCC.*

4. Numerical Experiment

In this section, we present a numerical experiment which is based on healthcare systems data. We use the following modified version of the design cost function introduced by Mettas [14] (see Huang *et al.* [6] as well) in our numerical experiments:

$$K(\tau) = B_1 \left[\exp\left(k \frac{\tau - \underline{\tau}}{\tau_\infty - \tau} \right) - 1 \right], \qquad \underline{\tau} \leq \tau \leq \bar{\tau},$$

where B_1 and k are strictly positive factors and τ_∞ is a given reliability level that exceeds $\bar{\tau}$ (i.e. $\bar{\tau} < \tau_\infty$). Notice that under this definition, reliability improvement becomes infeasible already before the costs become infinite. k is a parameter that represents the difficulty in increasing MTBF due to complexity, limited resources and technology, etc. Larger values of k correspond to more difficulties in increasing MTBF, so, higher design costs.

We formulate the unit production cost function as

$$c(\tau) = A + B_2(\tau^m - \underline{\tau}^m), \qquad \underline{\tau} \leq \tau \leq \bar{\tau},$$

where $A \geq 0$, $B_2 > 0$, and $m \geq 1$. This is a modified version of the unit production cost function used by Huang *et al.* [6]. In their paper, Huang *et al.* consider a situation in which production is carried out for a considerable duration and incorporate a learning effect on top of an initial unit cost in their unit production cost formulation. Since the production phase is negligibly short in our case, we omit the learning effect and our formulation is similar to their initial unit cost function. Note that $\lim_{\tau \to \underline{\tau}} c(\tau) = A$.

We investigate the effect of four factors on optimal decisions: *component type* (explained below), number of systems N, downtime penalty rate p, and length of the exploitation phase T. We created 81 instances by all combinations of three choices of the four factors. The choices of the factors are given in Table 1.

Table 1: Choices of the factors

Component type	N	T (months)	p ($ per hour)
cheap, medium, expensive	100, 500, 2500	60, 120, 240	100, 500, 2500

Component type reflects the value of a component in monetary terms. At the baseline MTBF, $\underline{\tau}$, the unit production cost of an expensive component is larger than that of a cheaper one. Furthermore, a certain amount of improvement in MTBF leads to a higher increase in both the unit production cost and design cost of an expensive component compared to a cheaper component. We realize the choices of component type mainly through the parameter B_1 in $K(\tau)$ and the parameters A and B_2 in $c(\tau)$. h (the storage cost rate per part), r_1, and r_2 are also varied for different choices of component type since each include a variable part which is positively correlated with the design costs and unit production cost. We use $k = 1$ and $m = 1$ in $K(\tau)$ and $c(\tau)$, respectively, for all three types of components. The other parameter values for different component types are given in Table 2. In the table, comp., mt., pt., and rp. stands for component, month, part, and repair, respectively.

Table 3 shows the values of the other parameters, which are fixed.

Table 2: Parameter values for component types

	Parameters					
Comp. Type	B_1 ($)	A ($/mt.)	B_2 ($)	h ($/mt. per pt.)	r_1 ($ per rp.)	r_2 ($ per rp.)
1 (ch)	200000	1000	10	20	600	1200
2 (med)	2000000	10000	100	200	1500	3000
3 (exp)	20000000	100000	1000	2000	10500	21000

Table 3: Values of fixed parameters

$\underline{\tau}$ (mt.)	$\bar{\tau}$ (mt.)	τ_∞ (mt.)	t_1 (hours)	t_2 (hours)	L (mt.)	α (per year)
24	120	240	10	50	3	0.05

We call the method where decisions on reliability level and the spare parts inventory level are made separately the *non-integrated method*. In the non-integrated method, while deciding on the reliability level, the focus is only on the design costs and the production costs. Then, for any values of the parameters of the cost functions, the OEM sets MTBF at $\underline{\tau}$

since the sum of design costs and production costs has its minimum value when $\tau = \underline{\tau}$. Next, the summation of the other cost terms in $\pi(\tau, s)$, which belong to the exploitation phase, are optimized by the optimal inventory level s for $\tau = \underline{\tau}$. Thus, the LCC found by the non-integrated method is $\pi_n = \pi(\underline{\tau}, s^*(\underline{\tau}))$. Denoting the optimal LCC in the joint optimization case as π^*, we define the *relative cost reduction* achieved by the joint optimization as

$$\Delta = \frac{\pi_n - \pi^*}{\pi_n}(100\%).$$

We present the results of the experiment in Table 4. The numbers given in the columns named Avg. τ^* and Avg. Δ are the average optimal MTBF values and the average relative cost reduction values, respectively. For example, the values 115.10 and 0.73 in the first row are the average optimal MTBF and the average relative cost reduction values found in the 27 instances with the cheap component. We also depict the minimum and maximum values of the optimal MTBF and relative cost reduction. u is the number of experiments in which optimal MTBF is found to be the upper bound $\bar{\tau}$ ($\tau^* = \bar{\tau}$).

In our experiment, we observe the following behaviour for Avg. τ^*:

- Avg. τ^* is much higher for the cheap component than for the expensive component. For the cheap component, reliability improvement is favoured by the less cost of reliability improvement and reductions in the repair costs and downtime costs achieved by reliability improvement.
- Avg. τ^* increases as the number of the systems increases. When the number of the systems increases, the frequency of system failures increases under the same τ, and, thus, repair costs and downtime costs increase. This constitutes an incentive to choose a higher reliability level.
- Avg. τ^* increases as the downtime penalty rate increases. The only effect of an increase in downtime penalty rate is that it increases the downtime costs. Higher reliability level compensates this increase.
- Avg. τ^* increases as the exploitation phase gets longer. A longer exploitation phase implies that the OEM has to deal with a larger number of failures and suffer from

Table 4: Results of the experiment

		Avg. τ^*	min τ^*	max τ^*	u	Avg.Δ	min Δ	max Δ
Comp.	1 (ch)	108.36	59.57	120.00	15	68%	37%	79%
	2 (med)	71.34	29.27	120.00	3	40%	3%	73%
	3 (exp)	40.63	24.00	73.04	0	15%	0%	44%
N	100	63.27	24.00	120.00	3	34%	0%	78%
	500	76.17	27.50	120.00	6	43%	1%	79%
	2500	80.88	28.81	120.00	9	46%	3%	79%
p	100	56.44	24.00	120.00	1	28%	0%	69%
	500	72.30	25.17	120.00	5	40%	0%	77%
	2500	91.58	33.32	120.00	12	56%	8%	79%
T	60	64.27	24.00	120.00	4	33%	0%	78%
	120	73.90	27.89	120.00	6	42%	2%	79%
	240	82.14	33.18	120.00	8	48%	7%	79%
All		73.44	24.00	120.00	18	41%	0%	79%

higher repair costs and downtime costs. This provokes choosing a higher reliability level.

Avg. Δ follows the same pattern as Avg. τ^*. Generally, the larger the distance between τ^* and $\underline{\tau}$, the larger the difference between π_n and π^*, and the larger Δ.

We should remark that the joint optimization leads to an average cost reduction of 41% in our experiment and it can even go up to 79%. These reductions correspond to large savings in absolute terms.

5. Conclusion

In this paper, we introduced a model for the joint optimization of reliability level and spare parts inventory level of a critical repairable component. We formulated the costs that are affected by the reliability level of a component and its spare parts inventory level throughout

the life time of a number of systems (LCC). We showed certain analytical properties of the cost function and provided an optimization procedure based on these properties. We conducted a numerical study based on real-life data and showed that our method leads to significant cost reductions compared to a non-integrated optimization method. The results of the experiment revealed that the optimal value of MTBF of a component depends on whether the component is cheap or expensive, the number of systems to be installed, downtime penalty rate and the length of exploitation phase.

In practice, reliability and spare parts inventory problems are generally dealt with on system level (multi-component). Further, there may exist multiple warehouses and repair facilities. We plan to extens the current work in this direction.

Acknowledgement

The authors gratefully acknowledge the support of the Innovation-Oriented Research Programme 'Integrated Product Creation and Realization (IOP IPCR)' of the Netherlands Ministry of Economic Affairs.

References

[1] W.R. Blischke. Mathematical models for analysis of warranty policies. *Mathematical and computer modelling*, 13:1–16, 1990.

[2] W.R. Blischke and D.N.P. Murthy. Product warranty management - i: A taxonomy for warranty policies. *European Journal of Operational Research*, 62:127–148, 1992.

[3] W.R. Blischke and D.N.P. Murthy. *Product Warranty Handbook*. Dekker, New York, 1996.

[4] G. Chattopadhyay and A. Rahman. Development of lifetime warranty policies and models for estimating costs. *Reliability Engineering and System Safety*, 93:522–529, 2008.

[5] A. Harel. Convexity properties of the Erlang loss formula. *Operations Research*, 38:499–505, 1990.

[6] H.Z. Huang, H.J. Liu, and D.N.P. Murthy. Optimal reliability, warranty and price for new products. *IIE Transactions*, 39:819–827, 2007.

[7] A.Z.M.O. Hussain and D.N.P. Murthy. Warranty and redundancy design with uncertain quality. *IIE Transactions*, 30:1191–1199, 1998.

[8] A.Z.M.O. Hussain and D.N.P. Murthy. Warranty and optimal redundancy with uncertain quality. *Mathematical and Computer Modelling*, 31:175–182, 2000.

[9] A.Z.M.O. Hussain and D.N.P. Murthy. Warranty and optimal reliability improvement through product development. *Mathematical and Computer Modelling*, 38:1211–1217, 2003.

[10] W. Karush. A queueing model for an inventory problem. *Operations Research*, 5:693–703, 1957.

[11] S.H. Kim, M.A. Cohen, and S. Netessine. Performance contracting in after-sales service supply chains. *Management Science*, 53:1843–1858, 2007.

[12] S.H. Kim, M.A. Cohen, and S. Netessine. Reliability or inventory? Contracting strategies for after-sales product support. *Working paper, The Wharton School*, 2007.

[13] A.A. Kranenburg and G.J. van Houtum. Cost optimization in the (S-1,S) lost sales inventory model with multiple demand classes. *Operations Research Letters*, 35:493–502, 2007.

[14] A. Mettas and R. Kallenberg. Reliability allocation and optimization for complex systems. *Proceedings of the Annual Reliability and Maintainability Symposium*, pages 216–221, 2000.

[15] A. Monga and M.J. Zuo. Optimal system design considering maintenance and warranty. *Computers and Operations Research*, 25:691–705, 1998.

[16] J.A. Muckstadt. *Analysis and Algorithms for Service Parts Supply Chains*. Springer, USA, 2005.

[17] D.N.P. Murthy and W.R. Blischke. Product warranty management - ii: An integrated framework for study. *European Journal of Operational Research*, 62:261–281, 1992.

[18] D.N.P. Murthy and W.R. Blischke. Product warranty management - iii: A review of mathematical models. *European Journal of Operational Research*, 62:1–34, 1992.

[19] D.N.P. Murthy and I. Djamaludin. New product warranty: A literature review. *International Journal of Production Economics*, 79:231–260, 2002.

[20] D.N.P. Murthy and A.Z.M.O. Hussain. Warranty and optimal redundancy design. *Engineering Optimization*, 23:301–314, 1995.

[21] D.N.P. Murthy, O. Solem, and T. Roren. Product warranty logistics: Issues and challanges. *European Journal of Operational Research*, 156:110–126, 2004.

[22] D.G. Nguyen and D.N.P. Murthy. Optimal reliability allocation for products sold under warranty. *Engineering Optimization*, 13:34–45, 1988.

[23] R. Oliva and R. Kallenberg. Managing the transition from products to services. *International Journal of Service Industry Management*, 14:160–172, 2003.

[24] K.B. Öner, R. Franssen, G.P. Kiesmüller, and G.J. Van Houtum. Life cycle costs measurement of complex systems manufactured by an engineer-to-order company. *In R.G. Qui, D.W. Russell, W.G. Sullivan, Ahmad, M. (Eds.), The 17th International Conference on Flexible Automation and Intelligent Manufacturing*, pages 569–589, 2007.

[25] K.B. Öner, G.P. Kiesmüller, and G.J. Van Houtum. A monotonicity result for the load carried by the last server in the erlang loss system. *Working Paper - BETA Research School*, 2008.

[26] A. Rahman and G. Chattopadhyay. Review of long term warranty policies. *Asia-Pacific Journal of Operational Research*, 23:453–472, 2006.

[27] C.C. Sherbrooke. *Optimal Inventory Modeling of Systems: Multi-Echelon Techniques - INTL Series in Operations Research & Management Science*. Kluwer Academic, 2004.

Appendix

Proof of Lemma 2. *(i)* Let $I(t)$ be the random variable denoting inventory on hand at time t, $f_{I(t)}(x)$ be its probability mass function and $h'(t)$ be the expected rate at which storage cost is incurred at time t. Then

$$h'(t) = h\sum_{x=0}^{\infty} x f_{I(t)}(x).$$

The expected NPV of storage cost throughout $[0,T]$, $S_2(\tau,s)$, is

$$S_2(\tau,s) = \int_0^T h'(t)e^{-\alpha t}dt = \int_0^T h\left(\sum_{x=0}^{\infty} x f_{I(t)}(x)\right)e^{-\alpha t}dt. \tag{A-1}$$

Because the stochastic process $\mathbf{I}=\{I(t): t \geqslant 0\}$ is assumed to be in steady-state throughout $[0,T]$, (A-1) may be further rewritten as

$$\begin{aligned} S_2(\tau,s) &= \int_0^T h\left(\sum_{x=0}^{\infty} x f_I(x)\right)e^{-\alpha t}dt = \left(\int_0^T he^{-\alpha t}dt\right)\left(\sum_{x=0}^{\infty} x f_I(x)dx\right) \\ &= \frac{h}{\alpha}\left(1-e^{-\alpha T}\right)\left(\sum_{x=0}^{\infty} x f_I(x)dx\right), \end{aligned} \tag{A-2}$$

where by $f_I(x)$ denotes the steady-state distribution of $\mathbf{I}$. Let $\bar{I}(\tau,s)$ denote the expected steady-state inventory level for a given τ and s. Then $\bar{I}(\tau,s)$ is equal to the the average number of idle servers in Erlang Loss System:

$$\bar{I}(\tau,s) = \sum_{x=0}^{\infty} x f_I(x)dx = s - \frac{NL}{\tau} + \frac{NL}{\tau}g(\tau,s) \tag{A-3}$$

where $g(\tau, s)$ is the probability of being out of stock (see (2) in Section 3). By substitution of this result into (A-2), we obtain

$$S_2(\tau, s) = \frac{h}{\alpha}\left(1 - e^{-\alpha T}\right)\left[s - \frac{NL}{\tau} + \frac{NL}{\tau}g(\tau, s)\right]$$

(ii) Let M_1 and M_2 be the random variables representing the numbers of ordinary repairs and emergency repairs performed throughout $[0, T]$, respectively. By Lemma 1, M_1 and M_2 have Poisson distributions with means $[1 - g(\tau, s)](N/\tau)T$ and $g(\tau, s)(N/\tau)T$, respectively.

Let W_1,..., W_{M_1} be the random variables representing the times of failures leading to instances of ordinary system repair throughout $[0, T]$ and V_1,..., V_{M_1} be the NPVs of the costs of the respective instances. W_i's are unordered times, that is, W_1 does not necessarily represent the time of the first failure, W_2 does not necessarily represent the time of the second failure, and so on. Since failures follow a Poisson process, W_i's are independent and uniformly distributed. That is, letting $f_{W_i}(t)$ denote the probability density function of W_i, $f_{W_i}(t) = 1/T$, $0 \le t \le T$. Then we can derive $E[V_i]$, the expectation of V_i, by conditioning on W_i.

$$E[V_i] = \int_0^T E[V_i|W_i = t] f_{W_i}(t)dt = \int_0^T r_1 e^{-\alpha t}\frac{1}{T}dt = \frac{r_1}{\alpha T}\left(1 - e^{-\alpha T}\right) \qquad i = 1, ..., M_1.$$

Let P_1 be the NPV of the costs of instances of ordinary system repair that are performed throughout $[0, T]$.

$$E[P_1|M_1 = m] = E\left[\sum_{i=0}^{m} V_i|M_1 = m\right] = E\left[\sum_{i=0}^{m} V_i\right] = \sum_{i=0}^{m} E[V_i] = mE[V_1].$$

since $E[V_i]$'s are the same for all i. Let $f_{M_1}(m)$ denote the probability mass function of M_1. Then

$$\begin{aligned} E[P_1] &= \sum_{m=0}^{\infty} E[P_1|M_1 = m] f_{M_1}(m) = \sum_{m=0}^{\infty} mE[V_1] f_{M_1}(m) \\ &= E[M_1] E[V_1] = [1 - g(\tau, s)] \frac{N}{\tau}\frac{r_1}{\alpha}\left(1 - e^{-\alpha T}\right). \end{aligned}$$

We denote the NPV of the costs of instances of emergency system repair that are performed throughout $[0, T]$ by P_2. Using similar arguments, $E[P_2]$ can be derived as

$$E[P_2] = g(\tau, s)\frac{N}{\tau}\frac{r_2}{\alpha}\left(1 - e^{-\alpha T}\right).$$

Then

$$R(\tau, s) = [1 - g(\tau, s)]\frac{N}{\tau}\frac{r_1}{\alpha}\left(1 - e^{-\alpha T}\right) + g(\tau, s)\frac{N}{\tau}\frac{r_2}{\alpha}\left(1 - e^{-\alpha T}\right).$$

(iii) Follows from the same arguments used in the proof for $R(\tau, s)$ in part(ii). □

Proof of Lemma 3. *(i)* $S(\tau,s)=S_1(\tau,s)+S_2(\tau,s)$ and it is trivial to show that $S(\tau,s)=0$ when $s=0$. For a fixed $s>0$, $S_1(\tau,s)=c(\tau)s$ is increasing in τ as $c(\tau)$ is increasing in τ.

The Erlang loss probability $g(\tau,s)$ can be expressed as $B(s,a)=\frac{a^s}{s!}/\sum_{i=0}^{s}\frac{a^i}{i!}$, where $a=(N/\tau)L$; a represents the arriving workload in the corresponding Erlang loss system. Then the average inventory given in equation $(A-3)$ can be rewritten as

$$\bar{I}\,(\tau,s)=\bar{I}\left(\frac{NL}{a},s\right)=s-a[1-B(s,a)].$$

$a[1-B(s,a)]$ is known as the carried load in the Erlang loss system and is strictly increasing in a for a fixed value of $s>0$ (see Corollary 1 in Öner *et al.* [25]). Thus, $\bar{I}\left(\frac{NL}{a},s\right)$ is strictly decreasing in a, which implies that $\bar{I}(\tau,s)$ is strictly increasing in τ. As $\frac{h}{\alpha}\left(1-e^{-\alpha T}\right)>0$, this monotonicity property of $\bar{I}(\tau,s)$ implies that $S_2(\tau,s)$ is strictly increasing in τ for $s>0$ (see equation (3)). Hence, $S(\tau,s)$ is strictly increasing in τ for $s>0$.

(ii) $R(\tau,s)$ can be rewritten as

$$R(\tau,s)=\frac{r_1}{\alpha}\frac{N}{\tau}\left(1-e^{-\alpha T}\right)+g(\tau,s)\frac{r_2-r_1}{\alpha}\frac{N}{\tau}\left(1-e^{-\alpha T}\right). \tag{A-4}$$

For its derivative with respect to τ, we find:

$$\frac{\partial R(\tau,s)}{\partial\tau}=-\frac{r_1}{\alpha}\frac{N}{\tau^2}\left(1-e^{-\alpha T}\right)+N\frac{r_2-r_1}{\alpha}\left(1-e^{-\alpha T}\right)\left[\frac{\partial g(\tau,s)}{\partial\tau}\frac{1}{\tau}-\frac{1}{\tau^2}g(\tau,s)\right].$$

$\frac{\partial g(\tau,s)}{\partial\tau}<0$ as $g(\tau,s)$ is strictly decreasing in τ (see Property 1). Thus, $\frac{\partial R(\tau,s)}{\partial\tau}<0$.

The proof for $D(\tau,s)$ follows from the same arguments as used for $R(\tau,s)$. □

Proof of Lemma 4. *(i)* It is trivial that $S_1(\tau,s)=c(\tau)s$ is increasing in s (note that $S_1(\tau,s)=0$ when $c(\tau)=0$). Let $\Delta S_2(\tau,s)=S_2(\tau,s+1)-S_2(\tau,s)$.

$$\Delta S_2(\tau,s)=\frac{h}{\alpha}\left(1-e^{-\alpha T}\right)\left\{1-[g(\tau,s)-g(\tau,s+1)]\frac{NL}{\tau}\right\}. \tag{A-5}$$

The Erlang loss probability $g(\tau,s)$ is strictly decreasing and strictly convex in s (see Karush [10]; see also Remark 2 in Kranenburg and van Houtum [13]). This implies that $\Delta g(\tau,s)=g(\tau,s)-g(\tau,s+1)$ in equation (A-5) is strictly decreasing in s. The maximum value of $\Delta g(\tau,s)$ for a fixed τ is attained when $s=0$, and

$$\Delta g(\tau,0)=1-\frac{\frac{NL}{\tau}}{1+\frac{NL}{\tau}}=\frac{\tau}{\tau+NL}. \tag{A-6}$$

Then

$$\Delta g(\tau,s)\frac{NL}{\tau}\leqslant\Delta g(\tau,0)\frac{NL}{\tau}=\frac{NL}{\tau+NL}<1, \tag{A-7}$$

which implies that

$$\Delta S_2(\tau,s)=\frac{h}{\alpha}\left(1-e^{-\alpha T}\right)\left[1-\Delta g(\tau,s)\frac{NL}{\tau}\right]>0. \tag{A-8}$$

That is, $S_2(\tau, s)$ is strictly increasing in s. Hence, $S(\tau, s)$ is strictly increasing in s.

(ii) The first term of $R(\tau, s)$ in equation (A-4) is constant. As $g(\tau, s)$ is strictly decreasing in s, $(1 - e^{\alpha T}) > 0$, and $r_1 \leq r_2$, $R(\tau, s)$ is decreasing in s. The proof for $D(\tau, s)$ follows from the same arguments. □

Proof of Lemma 5. *(i)* We rewrite equation (4) as

$$\begin{aligned}\pi(\tau, s) &= K(\tau) + \left[c(\tau) - c(\underline{\tau})\right] N + \frac{N}{\alpha}(r_1 + d_1)\left(1 - e^{-\alpha T}\right)\frac{1}{\tau} + c(\tau)s \\ &\quad + \frac{h}{\alpha}\left(1 - e^{-\alpha T}\right)\left[s - \frac{NL}{\tau}\right] \\ &\quad + \frac{N}{\alpha}\left(1 - e^{-\alpha T}\right)(hL + r_2 + d_2 - r_1 - d_1)\frac{1}{\tau}g(\tau, s). \end{aligned} \tag{A-9}$$

For a fixed value of τ:

- The first three terms of (A-9) are constant.
- The fourth and fifth term are linear in s.
- The last term in (A-9) is strictly convex in s because the Erlang loss probability $g(\tau, s)$ is strictly convex in s, $\left(1 - e^{-\alpha T}\right) > 0$, $r_1 \leq r_2$, and $d_1 \leq d_2$.

Hence $\pi(\tau, s)$ is strictly convex in s.

(ii) For a fixed value of τ:

- The first three terms of (A-9) are constant.
- $$\lim_{s \to \infty} \left\{ c(\tau)s + \frac{h}{\alpha}\left(1 - e^{-\alpha T}\right)\left[s - \frac{NL}{\tau}\right] \right\} = \infty$$
- The last term depends on s via $g(\tau, s)$, but it is bounded from below by 0.

Hence $\lim_{s \to \infty} \pi(\tau, s) = \infty$.

(iii) We rewrite equation (4) for $\pi(\tau, s)$ as

$$\begin{aligned}\pi(\tau, s) &= -c(\underline{\tau})N + \frac{h}{\alpha}\left(1 - e^{-\alpha T}\right)s + K(\tau) + c(\tau)N + c(\tau)s \\ &\quad + \frac{N}{\alpha}\left(1 - e^{-\alpha T}\right)(r_1 + d_1 - hL)\frac{1}{\tau} \\ &\quad + \frac{N}{\alpha}\left(1 - e^{-\alpha T}\right)(hL + r_2 + d_2 - r_1 - d_1)\, g(\tau, s)\frac{1}{\tau} \end{aligned} \tag{A-10}$$

Recall that $hL \leq r_1 \leq r_2$, $h > 0$, $L > 0$, and $d_1 \leq d_2$. For a fixed value of s:

- The first two terms of (A-10) are constant.
- The following three terms are convex in τ.
- The sixth term is strictly convex in τ.

- The last term is also strictly convex in τ. Let

$$f(\tau, s) = g(\tau, s)\frac{1}{\tau}.$$

Then

$$\frac{\partial^2 f(\tau,s)}{\partial \tau^2} = . \left[\frac{\partial^2 g(\tau,s)}{\partial \tau^2}\frac{1}{\tau} - 2\frac{1}{\tau^2}\frac{\partial g(\tau,s)}{\partial \tau} + \frac{2}{\tau^3}g(\tau,s)\right].$$

As $g(\tau, s)$ is strictly decreasing and strictly convex in τ, $\frac{\partial g(\tau,s)}{\partial \tau} < 0$ and $\frac{\partial^2 g(\tau,s)}{\partial \tau^2} > 0$. So, $\frac{\partial^2 f(\tau,s)}{\partial \tau^2} > 0$, which implies the strict convexity of the last term in τ.

Hence, $\pi(\tau, s)$ is strictly convex in τ.

(iv) Let $\Delta\pi(\tau, s) = \pi(\tau, s+1) - \pi(\tau, s)$. As we define $s^*(\tau) = \min\{\arg\min_{s\in\mathbb{N}_0} \pi(\tau, s)\}$, by parts (i) and (ii) imply that $s^*(\tau)$ is the smallest value of s satisfying $\Delta\pi(\tau, s) \geqslant 0$ and $s^*(\tau)$ is finite. The inequality $\Delta\pi(\tau, s) \geqslant 0$ may be shown to be equivalent to (use equation (A-9))

$$\frac{\Delta g(\tau,s)}{\tau} \quad \geqslant \quad \frac{h(1-e^{-\alpha T}) + \alpha c(\tau)}{N(1-e^{-\alpha T})(hL + r_2 + d_2 - r_1 - d_1)}, \tag{A-11}$$

where that $\Delta g(\tau, s) = g(\tau, s) - g(\tau, s+1)$. Let $a = (N/\tau)L$ represents the arriving workload in the corresponding Erlang loss system, as in the proof of Lemma 3. Let $F_B(s+1, a) = a\,[B(s, a) - B(s+1, a)]$. $F_B(s+1, a)$ is known as the load carried by the last server in the Erlang loss system (with $s+1$ servers). Then, equation (A-11) may be rewritten as

$$\frac{hL(1-e^{-\alpha T}) + \alpha c(\frac{NL}{a})}{(1-e^{-\alpha T})(hL + r_2 + d_2 - r_1 - d_1)} - F_B(s+1, a) \quad \geq \quad 0. \tag{A-12}$$

$F_B(s+1, a)$ is known to be increasing in a (see Öner *et al.* [25]). Hence, the lefthand side of inequality (A-12) is decreasing in a for each s. This implies that the first s for which inequality (A-12) is satisfied is increasing in a, and $s^*(\tau)$ (the first s for which inequality (A-11) is satisfied) is decreasing in τ.

□