Front. Eng. Manag.
https://doi.org/10.1007/s42524-024-0145-3

RESEARCH ARTICLE

Tongdan JIN, Shubin SI, Wenjin ZHU

Allocating redundancy, maintenance and spare parts for minimizing system cost under decentralized repairs

Abstract Reliability-redundancy allocation, preventive maintenance, and spare parts logistics are crucial for achieving system reliability and availability goal. Existing methods often concentrate on specific scopes of the system's lifetime. This paper proposes a joint redundancy-maintenance-inventory allocation model that simultaneously optimizes redundant component, replacement time, spares stocking, and repair capacity. Under reliability and availability criteria, our objective is to minimize the system's lifetime cost, including design, manufacturing, and operational phases. We develop a unified system availability model based on ten performance drivers, serving as the foundation for the establishment of the lifetime-based resource allocation model. Superimposed renewal theory is employed to estimate spare part demand from proactive and corrective replacements. A bisection algorithm, enhanced by neighborhood exploration, solves the complex mixed-integer, nonlinear optimization problem. The numerical experiments show that component redundancy is preferred and necessary if one of the following situations occurs: extremely high system availability is required, the fleet size is small, the system reliability is immature, the inventory holding is too costly, or the hands-on replacement time is prolonged. The joint allocation model also reveals that there exists no monotonic relation between spares stocking level and system availability.

Keywords system availability, installed base, decentralized repair, redundancy-maintenance-inventory model, superimposed renewal process

Received Dec. 22, 2023; revised Feb. 26, 2024; accepted Mar. 13, 2024

Tongdan JIN
Ingram School of Engineering, Texas State University, San Marcos, TX 78666, USA

Shubin SI, Wenjin ZHU (✉)
School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China; Key Laboratory of Industrial Engineering and Intelligent Manufacturing (Ministry of Industry and Information Technology), Xi'an 710072, China
E-mail: wenjin.zhu@nwpu.edu.cn

The research of the first author is supported by the US National Science Foundation (Grant No. 1704933). The research of the second and the third authors are supported by the National Natural Science Foundation of China (Grant No. 72231008), and by the Natural Science Basic Research Program of Shaanxi (Program No. 2022JQ-734).

1 Introduction

In the integrated product-service paradigm, many original equipment manufacturers (OEMs) strive for delivering high-reliability products along with responsive repair and maintenance services. However, achieving these goals at the same time can be challenging due to resource, time, and cost constraints. Therefore, it is important to develop a holistic framework that can effectively coordinate reliability design, maintenance policy, repair capacity, and spares provisioning throughout the entire product lifetime.

Various models have been proposed to achieve high system reliability at a low cost, including reliability-redundancy allocation (RRA), preventive (or predictive) (PM), and spare parts logistics (SPL). These models often focus on specific phases of the product lifetime. RRA primarily addresses product design and manufacturing, while PM and SPL are concerned with the aftermarket period. However, since these models are often implemented independently, they often lead to suboptimal solutions. To gain a better understanding of RRA, references such as Coit and Zio (2019) and Si et al. (2020) can be consulted. For comprehensive reviews on PM, including condition-based maintenance (CBM), refer to Alaswad and Xiang (2017) and Hu et al. (2022). Basten and van Houtum (2014) and Zhang et al. (2021) provide insights into SPL models. Recently, there has been a growing research stream on the coordination of RRA and PM, RRA and SPL, and PM and SPL, which will be discussed in Section 2. Despite the aforementioned studies, there is a lack of a holistic framework in which RRA, PM (or

CBM), and SPL are jointly optimized over the product lifetime (Jin, 2023). A holistic approach can guide firms in maintaining market competitiveness and achieving a win-win result between the OEM and customers. With emerging technologies such as digital twin and Internet of Things, the integration of all product phases, including design, manufacturing, and aftermarket, is the basis for minimizing product lifetime cost without compromising reliability and availability performance (Wang, 2021).

This paper aims to fill this gap by proposing a joint RRA, PM, and SPL optimization model to minimize costs across system design, manufacturing, and aftermarket. To that end, we present a mixed-integer, redundancy-maintenance-inventory allocation model that optimizes redundancy level, replacement time, spares inventory, and repair and renewing capacity. The goal is to minimize annualized system cost while satisfying reliability and availability criteria. The proposed model is applied in the semiconductor equipment industry, where zero system downtime is desirable for high production throughput. Our study shows that the OEM opts to adopt a redundancy strategy if: 1) extraordinary system availability, such as 0.999, is required; 2) the system fleet size is small; 3) parts holding costs are extremely high; 4) system reliability is immature; or 5) a prolonged replacement time occurs. The joint allocation model also reveals that the correlation between spares inventory and system availability is not necessarily monotonic.

The remainder of the article is organized as follows: Section 2 reviews the related literature. Section 3 characterizes Erlang-C repair and renewal queues under superimposed renewal processes. Section 4 presents a unified system availability model incorporating redundancy, maintenance, spares, and repair capacity. In Section 5, a joint redundancy-maintenance-inventory allocation model is formulated, and the bisection search algorithm is also elaborated. In Sections 6 and 7, the proposed model is demonstrated on semiconductor test equipment comprised of single and multiple redundant subsystems, respectively. Section 8 concludes the paper.

2 Literature review

This section reviews the works pertaining to three research streams: 1) joint allocation of RRA and SPL; 2) joint decision on RRA and PM; and 3) joint optimization of PM and SPL.

2.1 Joint allocation of reliability-redundancy and spares inventory

Much effort has been dedicated to managing spare parts inventory through the consideration of component reliability and installed base data (Louit et al., 2011; Dekker et al., 2013; Selviaridis and Wynstra, 2015). For example, Jin and Tian (2012) treat component reliability as an endogenous variable and combine it with an adaptive (Q,r) inventory policy to minimize the overall cost of the growing installed base throughout its lifecycle. This model has been further expanded by Jin et al. (2017) to integrate redundancy, along with reliability and spares stocking, in order to minimize system lifetime cost. Selçuk and Agrali (2013) study the trade-off between reliability investment and parts base-stock level to minimize the cost of a multi-item system fleet. Öner et al. (2013) propose an on-site, cold-standby redundancy strategy to mitigate equipment downtime, utilizing performance measures such as parts availability, expected backorders, and inventory cost. In our paper, we aim to optimize maintenance time and repair capacity, along with component redundancy and spares inventory for attaining the system availability goal.

Xie et al. (2014) present a continuous-time Markov chain model to maximize the system availability by jointly optimizing active redundancy and the base-stock level. Sleptchenko and van der Heijden (2016) jointly allocate redundancy and spare parts for a k-out-of-n system with different standby modes and part types. They find that high redundancy levels are only beneficial when components are relatively inexpensive and part replacement times are long. The latter also echoes our finding. Zhao et al. (2019) concurrently allocate repairmen, cold standby redundancy, and spares inventory to maximize system availability. A common assumption in these RRA-SPL models is that component lifetimes follow an exponential distribution with a constant failure rate. In our paper, we relax the constant failure rate assumption, and consider time-varying failure rates to generalize component lifetime distribution.

2.2 Joint decision on reliability-redundancy and maintenance

Some researchers argue that it is necessary to combine RRA and PM decisions because these decisions influence each other and collectively impact the total cost of a system's lifetime. For instance, Levitin and Lisnianski (1999) jointly optimize component redundancy and replacement schedules for multi-state systems to achieve the desired reliability objectives. They employ genetic algorithms to minimize system costs, which include capital, maintenance, and random failures. Nourelfath et al. (2012) and Liu et al. (2013) address the redundancy-maintenance optimization problem for multi-state systems under imperfect repair. The focus of both studies is to achieve the desired system availability while minimizing investments in redundant units and maintenance activities.

Moghaddass et al. (2012) conduct a study comparing the trade-off between component redundancy and its maintenance frequency to maximize the profitability in a

multi-state system, rather than solely focusing on cost reduction. They use a continuous-time Markov process model to estimate system availability and determine maintenance initiation criteria. Bei et al. (2017) formulate a two-stage stochastic optimization method assuming constant stress and perfect repair to determine component choice, redundancy level, and maintenance time for a series-parallel system. Later, Zhu et al. (2018) extend the redundancy-maintenance optimization model by incorporating time-varying usage stress and minimal repair. Bei et al. (2019) solve a similar problem by considering worst-case scenarios for future system usage. They minimize the conditional value-at-risk of the cost rate to obtain the risk-averse decision.

One common assumption in existing RRA-PM allocation models is the availability of spare parts is guaranteed. However, our paper acknowledges the backorder situation when spares inventory runs out. We aim to mitigate parts supply uncertainty and make a robust redundancy-maintenance decision by optimizing redundant components and replacement time.

2.3 Joint optimization of maintenance and spares inventory

This research stream is also known as maintenance service logistics (Vaughan, 2005; Van Horenbeek et al. 2013). The objective is to achieve high system availability by coordinating part replacement time with spares provisioning. For instance, de Smidt-Destombes et al. (2009) conduct a joint optimization of maintenance initiation, spares quantity, and repair capacity to minimize the ownership cost in a k-out-of-n system. Bjarnason and Taghipour (2016) coordinate inspection time, periodic reorders, and emergency order-up-to level using an (s, S) replenishment policy to minimize the system cost rate. Zhu et al. (2020) utilize maintenance schedules and advance demand information to forecast intermittent spares demand and develop a dynamic inventory control mechanism to minimize costs. Wang and Zhu (2021) jointly coordinate condition-based replacement and spares stocking policies for a multi-state k-out-of-n system. Zhang et al. (2022) address a condition-based maintenance service logistics problem for a series-parallel system with both hard and soft failures. These studies assume a pre-defined component redundancy level. However, in our model, redundancy is treated as an endogenous decision variable that is optimized alongside replacement time and spares stocking level.

Jin et al. (2015) present a principal-agent game model to minimize the annualized cost of repairable systems through the coordination of maintenance, spares inventory, and repair and renewing times in the aftermarket. Our study expands their model in two aspects. First, in addition to PM and SPL, we adopt component redundancy as an alternative approach to enhancing system reliability and availability. Secondly, we consider the limited capacity of repair and renewing shops, which are operated in a decentralized mode to accommodate different levels of skills and resources.

For further research on PM-SPL, we refer readers to the works of Wang et al. (2009), Chen et al. (2013), Bjarnason et al. (2014), Olde Keizer et al. (2017), Basten and Ryan (2019), and Zhu et al. (2022). It is common for maintenance service logistics models to assume unlimited repair capacity. However, our paper distinguishes itself from existing PM-SPL works by considering a repairable inventory with limited repair capacity. Díaz and Fu (1997) and Sleptchenko et al. (2002) demonstrate that capacitated repair is more realistic due to constraints in facilities and manning hours.

2.4 Summary of the research gap

The literature review reveals a lack of joint optimization framework of RRA, PM, and SPL. Our paper contributes to the literature in three key ways. First, our proposed redundancy-maintenance-inventory allocation model is the first of its kind to drive system reliability and availability performance throughout the design, manufacturing, and field use stages. Secondly, we introduce two parallel Erlang-C queues to handle parts repair and renewing tasks, respectively. Both queues can effectively accommodate the distinctions in processing time, manning skills, and reasons for return. Thirdly, we derive a unified system availability model that captures ten performance drivers, including redundancy level, maintenance time, spares stocking, and repair and renewing capacity.

3 An integrated product-service supply chain

3.1 The network setting

As depicted in Fig. 1, the system consists of multiple k_i-out-of-n_i active redundant subsystems (for $i = 1, 2, \ldots, N$) connected in series. The components within each subsystem are identical, but they differ among subsystems. Therefore, the system is made of N different part types. For the i^{th} subsystem, k_i represents the minimum required

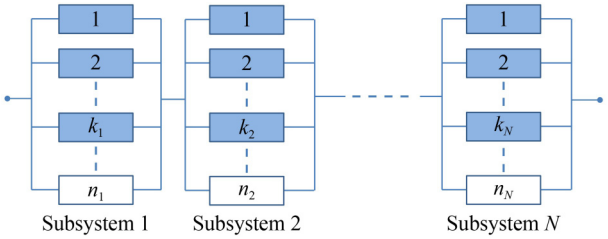

Fig. 1 A system comprised of N redundant subsystems in series.

working units, with $k_i \leqslant n_i$. As components are removable, they are also referred to as line replaceable units (LRUs). In this study, we use the terms component, part, and item interchangeably to refer to a repairable LRU.

The OEM implements an integrated product-service offering program to support m systems at the customer site shown in Fig. 2. Since the demand for spare parts is intermittent, a spares inventory is placed in proximity to these systems to facilitate replacement (Hekimoğlu et al., 2018). In the industry, age-based replacement is widely used due to its technical maturity and scheduling flexibility (El-Ferik, 2008; Huynh et al., 2012). For a part type i, where $i = 1,2,\ldots,N$, it is inspected at a predefined time interval τ_i. If the item survives through τ_i, it is proactively replaced with a spare item. If the item fails prior to τ_i, a corrective replacement is performed immediately. As a result, two types of spares demands are generated from the fleet: one for proactive replacement and the other for failure replacement. Upon renewal or repair, the part is put back into the inventory for future use.

Since repairing a failed part requires more time, resources, and skills than renewing an aging item, the OEM decides to decentralize the renewal and repair shops. Poisson process is commonly used to estimate spare part demands in repairable inventory literature (Lee, 1987; Kim et al., 2007; Öner et al., 2013). We adopt a similar approach to model the renewal and repair shops, respectively. Particularly, the $M_i/M_i/p_i/\infty$ model represents the renewal process, and the $M_i/M_i/q_i/\infty$ model represents the repair process, where p_i and q_i are the numbers of servers, respectively.

Table 1 lists the decision variables that the OEM attempts to optimize, including component redundancy, base stock level, replacement age, and renewing and repair servers. Table A. in Appendix A summarizes the notation of the model parameters of this paper. The objective is to minimize the annualized system cost subject to system reliability and availability criteria which will be elaborated in Section 5.

3.2 Superimposed parts renewal process

We begin the analysis of parts renewal process from single-item system that contains only one LRU. Reliability of a single-item system under age-based maintenance is often characterized by the mean-time-between-replacements (MTBRs). Let $R(t)$ be the component reliability, and $F(t)$ be the cumulative distribution function. Its MTBR can be estimated as:

$$MTBR = \int_0^\tau R(t)\,\mathrm{d}t = \tau - \int_0^\tau F(t)\,\mathrm{d}t. \tag{1}$$

In age-based maintenance, the spare parts demand process can be treated as the superposition of two renewal processes: a proactive replacement stream and a failure (i.e., corrective) replacement stream (Jin et al., 2015). For a single-item system with one LRU, let $\lambda_p(\tau)$ and $\lambda_q(\tau)$ be the spare parts demand rate for proactive replacement and failure replacement, respectively. Based on Eq. (1), we have:

$$\lambda_p(\tau) = \frac{R(\tau)}{\int_0^\tau R(t)\,\mathrm{d}t}, \tag{2}$$

$$\lambda_q(\tau) = \frac{F(\tau)}{\int_0^\tau R(t)\,\mathrm{d}t}. \tag{3}$$

Here $R(\tau)$ is the probability of a proactive replacement, and $F(\tau)$ is the probability of a failure placement. Given a fleet with m single-item systems, each system indepen-

Table 1 Decision variables

Notation	Definition
x_i	Redundancy level for part type i, for $i = 1,2,\ldots,N$
s_i	Base-stock level for part type i, for $i = 1,2,\ldots,N$
τ_i	Replacement age or interval for part type i, for $i = 1,2,\ldots,N$
p_i	Number of renewing servers for part type i, for $i = 1,2,\ldots,N$
q_i	Number of repair servers for part type i, for $i = 1,2,\ldots,N$

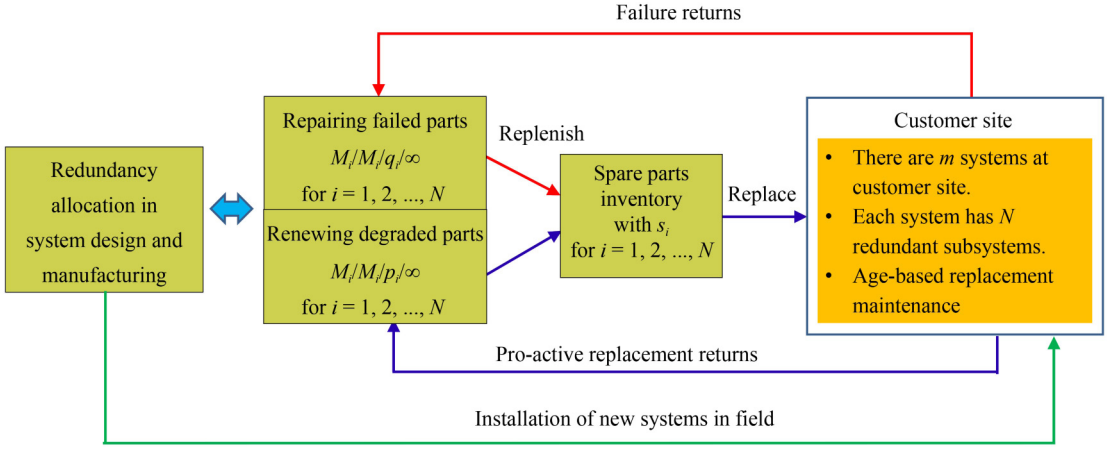

Fig. 2 Product-service integration with decentralized repair services.

dently generates proactive replacement and failure replacement streams, respectively. Hence the aggregate spare part demand rate of a single-item system fleet, denoted as $\lambda_m(\tau)$, can be estimated as:

$$\lambda_m(\tau) = m \times (\lambda_p(\tau) + \lambda_q(\tau)) = \frac{m}{\int_0^\tau R(t)\,\mathrm{d}t}. \tag{4}$$

The process formed by the union of fleet replacements is called a *superimposed renewal process* (SRP). Cox and Smith (1954) have demonstrated that as the fleet size (m) approaches infinity and the operating time is sufficiently large, the SRP becomes a homogeneous Poisson process, regardless of the lifetime distribution of each system. Wang (2012) further proves that the occurrence times between two successive replacements can be approximated as exponential as long as $m \geqslant 10$. The simulation done by Jin et al. (2021) also supports this statement, specifically in the context of age-based replacement. Wu (2019, 2021) has expanded the SRP theory to investigate systems under imperfect repair, incorporating non-exponential failures such as the arithmetic reduction of failure intensity and the arithmetic reduction of age. In our study, since a failed part is replaced with a spare part, the replacement is equivalent to a perfect repair.

3.3 Parts repair queueing model

Since SRP can be approximated as a homogeneous Poisson process, the Erlang-C queueing model can be used to characterize the performance of the repair shop. The Erlang-C queue accommodates a waiting line, which is commonly found in a capacitated repair shop. Let q denote the number of repair servers, and $\lambda_{F,q}$ denote the arrival rate of failed parts to the repair shop. If a fleet consists of m single-item systems, then $\lambda_{F,q} = m\lambda_q(\tau)$ where $\lambda_q(\tau)$ is given in Eq. (3). The transition diagram of the $M/M/q/\infty$ queue is provided in Fig. 3.

The state in the transition diagram represents the number of failed parts in the repair shop, and μ_q is the repair rate per server. Let $B(q)$ denote the probability that an incoming part needs to wait in the queue. According to Winston (2004), we have:

$$\begin{aligned} B(q) &= \frac{\dfrac{(\lambda_{F,q}/\mu_q)^q}{q!(1-\lambda_{F,q}/(q\mu_q))}}{\displaystyle\sum_{j=0}^{q-1}\frac{(\lambda_{F,q}/\mu_q)^j}{j!} + \frac{(\lambda_{F,q}/\mu_q)^q}{q!(1-\lambda_{F,q}/(q\mu_q))}} \\ &= \frac{\dfrac{(q\rho_q)^q}{q!(1-\rho_q)}}{\displaystyle\sum_{j=0}^{q-1}\frac{(q\rho_q)^j}{j!} + \frac{(q\rho_q)^q}{q!(1-\rho_q)}}, \end{aligned} \tag{5}$$

where $\rho_q = \dfrac{\lambda_{F,q}}{q\mu_q}$ is called the traffic intensity rate. The queue is stable if and only if $\rho_q < 1$. The repair turn-around time, denoted as t_q, measures the duration from when the part enters the repair shop to when it is fixed and put back to the spares inventory. If the part transportation time is small or can be ignored, t_q can be obtained as:

$$t_q = \frac{B(q)}{q\mu_q - \lambda_{F,q}} + \frac{1}{\mu_q}. \tag{6}$$

3.4 Parts renewing queueing model

A separate Erlang-C queue denoted as $M/M/p/\infty$ is used to characterize the renewing shop. The probability that an incoming part needs to wait before being renewed can be estimated as:

$$C(p) = \frac{\dfrac{(p\rho_p)^p}{p!(1-\rho_p)}}{\displaystyle\sum_{j=0}^{p-1}\frac{(p\rho_p)^j}{j!} + \frac{(p\rho_p)^p}{p!(1-\rho_p)}}, \tag{7}$$

where $\rho_p = \dfrac{\lambda_{F,p}}{(p\mu_p)}$ is called the renewing traffic intensity rate. Note that $\lambda_{F,p}$ is the parts arrival rate to the renewing shop with $\lambda_{F,p} = m\lambda_p(\tau)$, and μ_p is the renewing rate per server. The renewing queue is stable if and only if $\rho_p < 1$. The renewing turn-around time, denoted as t_p, can be estimated as:

$$t_p = \frac{C(p)}{p\mu_p - \lambda_{F,p}} + \frac{1}{\mu_p}. \tag{8}$$

By combining Eqs. (6) and (8), the average part turn-around time (ATT), denoted as t_{ATT}, is obtained as follows:

$$\begin{aligned} t_{ATT} &= t_q F(\tau) + t_p R(\tau) \\ &= F(\tau)\left(\frac{B(q)}{q\mu_q - \lambda_{F,q}} + \frac{1}{\mu_q}\right) + R(\tau)\left(\frac{C(p)}{p\mu_p - \lambda_{F,p}} + \frac{1}{\mu_p}\right). \end{aligned} \tag{9}$$

4 Availability of repairable system

4.1 Availability of single-item system

The availability of a single-item system is frequently used to manage preventive maintenance and spare parts logistics when the system's unitization remains relatively stable (Louit et al., 2011; de Smidt-Destombes et al., 2009). It can be calculated using the following expression:

$$A = \frac{MTBR}{MTBR + MDT}, \tag{10}$$

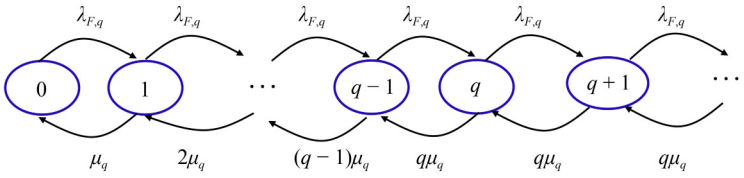

Fig. 3 The $M/M/q/\infty$ queueing model for parts repair process.

where MTBR is given in Eq. (1), and MDT stands for the system mean downtime either due to a planned or a failure replacement. System downtime under a planned replacement comprises of hands-on replacement time and delay if the inventory is out of stock. Let O be a random variable representing the spare part demand of the inventory, and s be the base-stock level with one-for-one replenishment. The downtime under a planned replacement, denoted as T_p, can be expressed as:

$$T_p = t_s + t_p \Pr\{O > s\}, \tag{11}$$

where t_s is the hands-on replacement time, and $\Pr\{O > s\}$ is the stockout probability. Similarly, the downtime of a failure replacement, denoted as T_q, can be expressed as:

$$T_q = t_s + t_q \Pr\{O > s\}. \tag{12}$$

By combining both scenarios, the actual MDT of a single-item system is given as

$$\begin{aligned} MDT &= T_p R(\tau) + T_q F(\tau) \\ &= t_s + (t_p R(\tau) + t_q F(\tau)) \Pr\{O > s\} \\ &= t_s + t_{ATT} \Pr\{O > s\}, \end{aligned} \tag{13}$$

where t_{ATT} is the average part turn-around time in Eq. (9). Since O is the fleet spare parts demand that follows the Poisson process, the stockout probability can be obtained as:

$$\Pr\{O > s\} = \sum_{j=s+1}^{\infty} \frac{\mu^j e^{-\mu}}{j!} = 1 - \sum_{j=0}^{s} \frac{\mu^j e^{-\mu}}{j!}, \text{ for } s = 0, 1, 2, \ldots \tag{14}$$

with

$$\mu = \lambda_F (t_p R(\tau) + t_q F(\tau)), \tag{15}$$

where μ is the mean spare parts demand during ATT, and λ_F is the parts demand rate of the fleet. If a fleet comprises m single-item systems, we have $\lambda_F = \lambda_m(\tau)$ as shown in Eq. (4). Now the single-item system availability, denoted as A, is obtained by substituting Eqs. (1), (13) and (14) into (10) as follows:

$$A = \frac{\int_0^{\tau} R(t)\,\mathrm{d}t}{\int_0^{\tau} R(t)\,\mathrm{d}t + t_s + (t_p R(\tau) + t_q F(\tau)) \left(1 - \sum_{j=0}^{s} \mu^j e^{-\mu} (j!)^{-1}\right)}. \tag{16}$$

Note that Eq. (16) incorporates nine performance drivers. These are the part reliability $R(t)$, the maintenance interval τ, the base stock level s, the fleet size m, the hands-on replacement time t_s, the number of renewing and repair servers p and q, and the parts renewing rate μ_p and repair rate μ_q that are embedded in μ through Eqs. (6), (8) and (15).

4.2 Availability of k-out-of-n redundant system

For a k-out-of-n system with active redundancy, the system is functional provided that at least k components are good at any point in time. Hence the system availability, denoted as A_R, is estimated by

$$\begin{aligned} A_R(x, s, \tau, p, q) &= \sum_{j=k}^{n} \binom{n}{j} A^j (1-A)^{n-j} \\ &= \sum_{j=k}^{k+x} \binom{k+x}{j} A^j (1-A)^{k+x-j}, \end{aligned} \tag{17}$$

where x is the number of redundant units with $x + k = n$. Note that A is the single-item system availability in Eq. (16). Together with x, there are ten performance drivers in A_R. For a fleet with m redundant systems, the spare parts demand rate of the fleet is $\lambda_F = (x + k)\lambda_m(\tau)$. Two assumptions are made in Eqs. (16) and (17). First, the system is repairable with random up and down cycles. Second, the utilization of each system may vary, but the average utilization shall remain stable over time.

5 Redundancy-maintenance-inventory allocation model

5.1 Minimizing annualized system cost

Based on the integrated product-service supply chain depicted in Fig. 2, we propose a redundancy-maintenance-inventory allocation (RMIA) model with the objective of minimizing the annualized system cost of the fleet. RMIA represents a lifetime approach to attaining system reliability and availability goal by integrating design, manufacturing, and maintenance logistics activities. The system cost is comprised of: 1) initial capital, 2) overhead costs for repairing and renewing parts, 3) inventory expenses for spare parts, and 4) operating costs for repair

and renewal shops. Table 1 lists the decision variables, which include redundancy level, spares stocking level, replacement age, renewing servers, and repair servers. We denote this model as RMIA, and it is formulated as follows:

Model RMIA

Min:

$$f(\boldsymbol{x},\boldsymbol{s},\boldsymbol{\tau},\boldsymbol{p},\boldsymbol{q}) = \sum_{i=1}^{N}(k_i + x_i)(\varphi_1 c_{LRU,i} + \lambda_{p,i}c_{u,i} + \lambda_{q,i}c_{v,i}) + \frac{1}{m}\sum_{i=1}^{N}\left(s_i(\varphi_2 c_{LRU,i} + c_{h,i}) + (p_i c_{p,i} + q_i c_{q,i})\right), \tag{18}$$

subject to:

$$\prod_{i=1}^{N} A_{R,i}(x_i, s_i, \tau_i, p_i, q_i) \geqslant A_{\min}, \tag{19}$$

$$\tau_i \geqslant \theta \overline{T}_i, \text{ for } i = 1,2,\ldots,N, \tag{20}$$

$$x_i + k_i \leqslant n_{\max,i}, \text{ for } i = 1,2,\ldots,N, \tag{21}$$

$$s_i \text{ and } x_i \in \{0,1,2,\ldots\}, \text{ for } i = 1,2,\ldots,N, \tag{22}$$

$$p_i \text{ and } q_i \in \{1,2,3,\ldots\}, \text{ for } i = 1,2,\ldots,N. \tag{23}$$

The objective Eq. (18) captures the annualized system cost associated with initial capital, preventive maintenance, spares inventory, and parts repair and renewal activities. Note that $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]$, $\boldsymbol{s} = [s_1, s_2, \ldots, s_N]$, $\boldsymbol{\tau} = [\tau_1, \tau_2, \ldots, \tau_N]$, $\boldsymbol{p} = [p_1, p_2, \ldots, p_N]$, and $\boldsymbol{q} = [q_1, q_2, \ldots, q_N]$ represent the decision variables. Model RMIA is also applicable to new product introduction phase, when the cost, reliability, repair skillset, and technology maturity differ significantly among different LRU types. To meet the time-to-market goal, the OEM utilizes both in-house and global resources to perform decentralized repairs in different locations with distinct repair crews. For instance, in the high-speed rail industry, maintenance tasks are assigned to different repair crews based on their individual skillsets to enhance accountability and categorize labor skills. Both φ_1 and φ_2 are the capital recovery factors for the system and spare parts, respectively. $\lambda_{p,i}(\tau)$ and $\lambda_{q,i}(\tau)$ are the spare parts demand rate for planned and failure replacements of part type i. Additionally, $c_{u,i}$ and $c_{v,i}$ represent the renewal and repair costs of part type i, respectively, while $c_{h,i}$ signifies the unit annual holding cost. Finally, $c_{p,i}$ represents the annual cost per repair server, and $c_{q,i}$ denotes the annual cost per renewing server.

Constraint (19) defines the system availability target, where $A_{R,i}$ stands for the availability of redundant subsystem i, as given in Eq. (17). Constraint (20) defines the reliability criterion for each LRU type. That is, τ_i should exceeds certain percentage of component's MTBF, and typically $\theta \geqslant 50\%$. Constraint (21) defines the physical limitations of each subsystem. Constraints (22) and (23) simply stipulate that $\boldsymbol{x}$, $\boldsymbol{s}$, $\boldsymbol{p}$, and $\boldsymbol{q}$ are nonnegative integers.

5.2 Bisection search algorithm

The bisection algorithm is a highly effective method for solving non-convex optimization models that arise in a variety of fields including reliability, inventory, power systems, and space-trajectory problems. For example, Mouatasim (2018) proposes a reduced gradient and bisection method for optimizing a non-convex differentiable objective function, with results confirming the global convergence of the algorithm. Reddy and Bijwe (2018) combine the bisection method with simulation to efficiently solve a large-scale optimal power flow model involving non-convex and discrete variables. Jin et al. (2017) demonstrate the use of the bisection search to address a joint RRA and SPL allocation problem. More recently, Barnett and Gosselin (2021) have developed a bisection algorithm to minimize the time required to follow a path defined in space by dividing the global problem into a series of simpler subproblems. In this paper, we propose the use of bisection search coupled with neighborhood exploration to solve the RMIA model. Specifically, we utilize Algorithms 1 and 2 for solving the case of single k-out-of-n systems (i.e., when $N = 1$), while Algorithm 3 becomes necessary when $N \geqslant 2$.

Algorithm 1: (Minimizing system cost)

Step 1: Initialization: estimate q_L, q_U, p_L, and p_U using Eqs. (B3), (B4), (B7), and (B9), respectively. Set $x = 0$, $s = 0$, $\tau_{\min} = \gamma_{\min}\overline{T}$, $\tau_{\max} = \gamma_{\max}\overline{T}$, $p = p_L$, $q = q_L$, and $f_{\min} = 10^9$ (an arbitrarily large value).

Step 2: Compute system availability using Eq. (17) based on current $\{x, s, \tau, p, q\}$.

Step 3: If $A < A_{\min}$, let $s = s + 1$, and go to Step 2. Else, compute $f(x, s, \tau, p, q)$ using Algorithm 2. If $f(x, s, \tau, p, q) < f_{\min}$, let $f_{\min} = f(x, s, \tau, p, q)$, $x^* = x$, $s^* = s$, $\tau^* = \tau$, $p^* = p$, and $q^* = q$.

Step 4: If $p < p_U$, let $p = p + 1$, and $s = 0$, and go to Step 2.

Step 5: If $q < q_U$, let $q = q + 1$, $p = p_L$, and $s = 0$, go to Step 2.

Step 6: If $x < n_{\max} - k$, let $x = x + 1$, $q = q_L$, $p = p_L$, and $s = 0$, go to Step 2.

Step 7: Output $f_{\min}$, and $\{x^*, s^*, \tau^*, p^*, q^*\}$.

Algorithm 2: (Bisection search)

Let τ_L and τ_U be the lower and upper bounds of τ, and f_L and f_U are the corresponding objective function values for given x, s, p, and q. Figure 4 illustrates the working principle of the bisection search. The detailed procedures are given below.

Step 1: Let $\tau_1 = (\tau_L + \tau_U)/2$, and use Algorithm 1 to find f_1.

Step 2: Let $\tau_2 = (\tau_1 + \tau_L)/2$, and use Algorithm 1 to find f_2.

Step 3: Let $\tau_3 = (\tau_1 + \tau_U)/2$, and use Algorithm 1 to find f_3.

Step 4: If $f_2 > f_1$, and $f_3 > f_1$, let $\tau_L = \tau_2$, $\tau_U = \tau_3$, and $f_{\min} = f_1$, go to step 2.

Step 5: If $f_2 > f_1 > f_3$, let $\tau_L = \tau_2$, and $f_{\min} = f_3$, or if $f_3 > f_1 > f_2$, let $\tau_U = \tau_3$, and $f_{\min} = f_2$, go to step 2.

Step 6: The algorithm terminates if $|f_{\min} - f(x^*, s^*, \tau^*, p^*, q^*)| < \varepsilon$, where $f_{\min}$ and $f(x^*, s^*, \tau^*, p^*, q^*)$ are the previous and the current values, and ε is a small threshold. Finally, the optimal solution is $\{x^*, s^*, \tau^*, p^*, q^*\}$.

Algorithm 3: (Neighborhood exploration)

This algorithm solves Model RMIA for systems comprised of multiple k_i-out-of-n_i redundant subsystems for $i = 1, 2, \ldots, N$. First, Algorithms 1 and 2 are used to find the optimal solution for each subsystem. Next, a neighborhood search is employed to further reduce the cost by refining all the decision variables. The detailed procedures are as follows:

Step 1: Set $A_{\min,i} = (A_{\min})^{1/N}$ where $A_{\min,i}$ is the subsystem availability for $i = 1, 2, \ldots, N$. Find the optimal solution of subsystem i using Algorithms 1 and 2. The results are kept as f_i, A_i^*, and $z_i = \{x_i, s_i, \tau_i, p_i, q_i\}$. Note that $A_i^* \geqslant A_{\min,i}$.

Step 2: For subsystem i, perform neighborhood exploration by increasing or decreasing $z_i = \{x_i, s_i, \tau_i, p_i, q_i\}$ by one step size, i.e., $\{\Delta x_i, \Delta s_i, \Delta \tau_i, \Delta p_i, \Delta q_i\}$, and compute the new cost and subsystem availability for $i = 1, 2, \ldots, N$. The results are kept at $\{z_i^+, f_i^+, A_i^+\}$ and $\{z_i^-, f_i^-, A_i^-\}$. Note that "+" stands for the increment, and "−" stands for the decrease.

Step 3: Among N subsystems, choose the subsystem with the maximum cost saving and the smallest availability reduction, say subsystem j. Also choose the subsystem with the minimum cost increase and the largest availability growth, say subsystem l.

Step 4: If the cost saving of subsystem j is less than the cost increase of subsystem l, using the current solutions for subsystems j and l. Compute the new system availability A_{sys}.

Step 5: If $A_{\text{sys}} \geqslant A_{\min}$, let $z_j = z_j^-$, $f_j = f_j^-$, and $A_j = A_j^-$ for subsystem j. Let $z_l = z_l^+$, $f_l = f_l^-$, and $A_l = A_l^+$ for subsystem l. Also update objective function $f_{\min}$, and $\{\boldsymbol{x}, \boldsymbol{s}, \boldsymbol{\tau}, \boldsymbol{p}, \boldsymbol{q}\}$. Go back to Step 2.

Step 6: The algorithm terminates if $|f_{\min} - f(\boldsymbol{x}^*, \boldsymbol{s}^*, \boldsymbol{\tau}^*, \boldsymbol{p}^*, \boldsymbol{q}^*)| < \varepsilon$, where $f_{\min}$ and $f(\boldsymbol{x}^*, \boldsymbol{s}^*, \boldsymbol{\tau}^*, \boldsymbol{p}^*, \boldsymbol{q}^*)$ are the previous and the current costs, and ε is a small threshold. The final solution is $\{\boldsymbol{x}^*, \boldsymbol{s}^*, \boldsymbol{\tau}^*, \boldsymbol{p}^*, \boldsymbol{q}^*\}$.

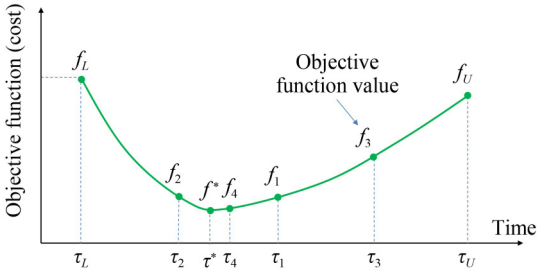

Fig. 4 A graphical illustration of the bisection search.

6 Applications to systems with single redundant subsystem

6.1 System description

Automated Test Equipment (ATE) is widely used for micro-device testing in the semiconductor manufacturing industry. ATE belongs to k-out-of-n redundant system, where k is the number of primary working units, and n is the total number of LRU items. Each LRU is made of a printed circuit board that is repairable. Without loss of generality, the lifetime of an LRU follows the Weibull distribution with shape and scale parameters $\alpha > 0$ and $\beta > 1$, respectively. Table 2 lists the reliability and cost data associated with ATE design, manufacturing, and after-sales support. The second column data are for the benchmark study, and those in the third column are for sensitivity analysis. Both φ_1 and φ_2 are estimated assuming a 5% discount rate with a 10-year and 5-year payoff period for systems and LRU, respectively.

6.2 Result and discussion of benchmark study

The benchmark data in Table 2 are used to solve Model RMIA for a fleet of k-out-of-n redundant systems. Algorithms 1 and 2 are used to search for the optimal solution. For $m = 50$ systems, the optimal decisions are $x^* = 0$, $s^* = 15$, $\tau^* = 3.706$, $p^* = 2$, and $q^* = 2$. The annualized system cost is $f_{\min} = \$126{,}407.18$. The achieved system availability is $A_{\text{sys}} = 0.9901$, larger than $A_{\min} = 0.99$.

Now we examine how the fleet size m influences the system cost, and the results are shown in Fig. 5. Initially, the system cost decreases with m due to economies of scale. However, it tends to level off as m further increases. For instance, the system cost drops to \$138,598.13 for $m = 20$, compared to \$191,147.79 for $m = 10$, resulting in a decrease of 27.4%. However, the cost tends to remain relatively flat with an average of \$120,415.45 for $m \geqslant 110$. The achieved system availability A_{sys} fluctuates between 0.99 and 0.992 as m increases from 10 to 200.

Figure 6 shows the solutions for $\{x, s, \tau, p, q\}$ as m increases from 10 to 200. Two observations can be made. First, it is not cost-effective to employ redundant units in order to achieve a system availability of 0.99. Secondly, the spares stock level does not increase monotonically with m. For instance, $s = 38$ for $m = 100$, but it decreases to 25 for $m = 130$. While $q = 4$ for $m = 100$ and 130, the OEM chooses to increase p from 3 to 6 in exchange for s.

Table 2 Parameter values of ATE system (n/a=not applicable, item=LRU)

Notation	Benchmark	Sensitivity Analysis	Unit
α	0.2	[0.2, 1.5]	failure/year
β	3	[1.5, 5]	n/a
k	10	10	item
$n_{\max}$	13	13	item
m	50	[10, 200]	system
t_s	8	[4, 48]	hour
$\frac{1}{\mu_p}$	6	[3, 18]	day
$\frac{1}{\mu_q}$	12	[6, 24]	day
c_{LRU}	50,000	[25000, 200000]	$/item
c_u	3,000	n/a	$/item
c_v	4,500	n/a	$/item
c_h	10,000	[5000, 50000]	$/item/year
c_p	480,000	$[0.5c_p, 1.5c_p]$	$/server
c_q	640,000	$[0.5c_q, 1.5c_q]$	$/server
$A_{\min}$	0.99	[0.9, 0.999]	n/a
θ	0.7	[0.5, 1.1]	n/a
$\gamma_{\min}, \gamma_{\max}$	0.5, 2	0.5, 2	n/a
φ_1, φ_2	0.1295, 0.2310	0.1295, 0.2310	n/a

This contradicts the intuition that more spare parts are needed as the fleet size increases under the ample repair capacity assumption.

Figure 7 illustrates the relationship between s and parts availability for $m \in [10, 200]$. Firstly, the parts availability remains relatively stable between 0.924 and 0.981 regardless of s. Secondly, the parts availability consistently falls below $A_{\min}$. Lastly, for $m \leqslant 100$, increasing s proves to be an effective method of meeting $A_{\min}$. However, when $m \geqslant 110$, the inventory levels off or even decreases. Hence, expanding repair and renewing servers becomes more cost-effective in order to achieve $A_{\min}$. Regardless of m, the utilization rate of renewing and repair servers is 0.84 and 0.86, respectively. This result aligns with the study by Sleptchenko et al. (2003), which demonstrates that capacitated repair shops typically have a utilization rate ranging from 0.8 to 0.95.

6.3 Comparison between redundancy and sparing

In this section, a sensitivity analysis is conducted by comparing redundancy allocation and spares stocking. First, five cases corresponding to parameters $A_{\min}$, θ, α, t_s and c_h are examined. For each parameter, Model RMIA is solved with three different values considering redundancy and non-redundancy, respectively. The results are presented in Table 3, where the optimal solutions are indicated by underscores.

In Case 1, we analyze the influence of $A_{\min}$ on the decisions regarding $\{x, s, \tau, p, q\}$. To achieve $A_{\min} = 0.999$, both redundancy and a larger spares inventory are required, with $x = 1$ and $s = 23$. The cost of the system is \$137,348. It should be noted that there is no feasible solution for $x = 0$, indicating that spares inventory alone cannot guarantee an availability of 0.999. As $A_{\min}$ decreases to 0.99, the optimal values are $x = 0$ and $s = 15$, resulting in a system cost of \$126,407. If $A_{\min}$ further decreases to 0.9, $x = 0$ and $s = 9$ are sufficient to achieve the target availability with a lower cost of \$123,757. Case 1 also demonstrates that as $A_{\min}$ is relaxed, the system cost is reduced, but the values of τ, p and q remain relatively stable.

In Case 2, we increase θ from 0.5 to 1.1 and examine its impact on the decision variables. For $\theta = 0.5$, the optimal solution is the same as that of $\theta = 0.7$, indicating that a high replacement frequency is not necessarily optimal. For $\theta = 1.1$, the optimal $\tau = 5.068$, which is 1.17 times of the MTBF. A larger τ results in a lower proactive replacement frequency, but an increased corrective maintenance. As a result, the system cost increases to \$129,623, compared to \$126,407 for $\theta = 0.5$ and 0.7.

In Case 3, we decrease the LRU reliability by increasing

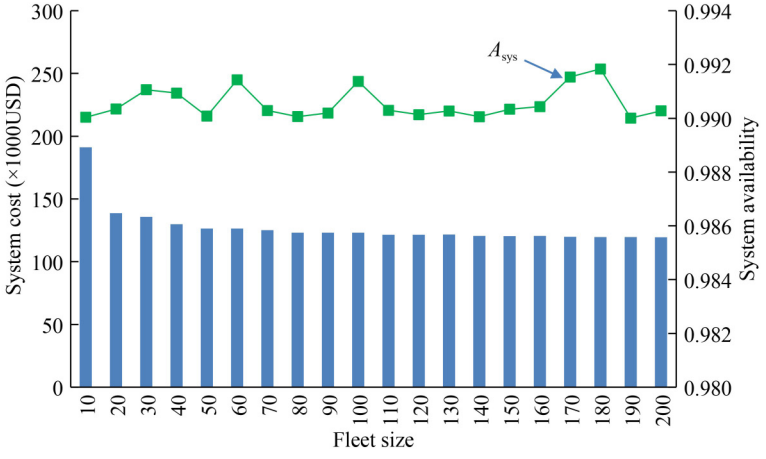

Fig. 5 System cost and availability for different fleet sizes.

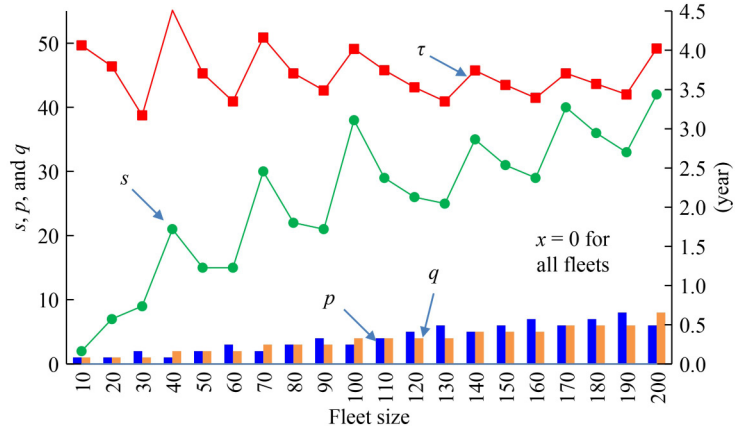

Fig. 6 Optimal solutions for different fleet sizes.

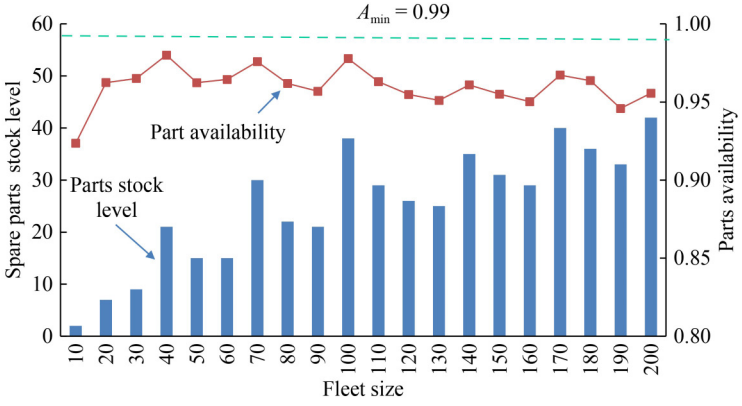

Fig. 7 Spare part stock level and parts availability for benchmark study.

α from 0.2 to 1. The system can achieve the target availability $A_{\min} = 0.99$ by using spare parts alone for $\alpha = 0.2$ and 0.5. However, redundancy with $x = 1$ must be adopted for $\alpha = 1$ alone with $s = 33$. It is also observed that s, p, and q increase with α, which is expected due to the growing number of field returns.

Case 4 examines the impact of the hands-on replacement time t_s on the decision making. It demonstrates that t_s has no direct effect on τ, p and q. In addition, a smaller s is sufficient to achieve $A_{\min} = 0.99$ if $t_s = 8$ or 24 h. However, if $t_s = 48$ h, $x = 1$ must be adopted to attain the desired system availability.

In Case 5, we increase the inventory holding cost and exmine its impact on the decision variables. Redundancy is not necessary when the part's annual holding cost is relatively low, with $c_h = 10{,}000$ and 20,000. However, if c_h reaches the item cost, $x = 1$ and $s = 0$ result in a lower system cost compared to the alternative solution of $x = 0$ and $s = 15$.

Next, we compare five additional cases pertaining to β, μ_p, μ_q, c_p, and c_q. Model RMIA is solved by varying one parameter, and the results are summarized in Table 4. The solutions marked with an underscore represent the optimal decisions. A common observation from Cases 6 to 10 is that spares inventory is more cost-effective than redundancy in achieving $A_{\min} = 0.99$.

Case 6 examines the influence of the shape parameter β on the decision variables and system cost. When β increases from 1.5 to 4.5, there is a preference for more proactive replacements as evidenced by the increased value of $R(\tau)$ from 0.21 to 0.86. This is because the life distribution with a higher β becomes more concentrated, thereby benefiting proactive replacements. Consequently, p increases from 1 to 3, and q decreases from 4 to 1. Additionally, it is observed that the system cost decreases with β due to the benefit of proactive replacements.

Case 7 explores the effects of μ_p on the decision variables and system cost. As $1/\mu_p$ increases from 3 to 12 days, s increases from 12 to 17, and the cost rises from \$115,023 to \$139,453. This result is expected, as a slower renewing process requires more spare parts to ensure the system availability. Furthermore, the OEM chooses to extend τ from 2.5 to 5.09 years to tolerate more failure replacements.

In Case 8, the repair time $1/\mu_q$ increases from 6 to 24 days. Similar to Case 7, the increase in $1/\mu_q$ leads to an increase in the inventory level from $s = 7$ to 32, and the cost from \$112,741 to \$145,399. Additionally, the OEM opts to adopt more renewing servers for proactive replacements. For instance, when $1/\mu_q = 6$, we have $p = 1$

Table 3 Comparison between redundancy and spares inventory for Cases 1 to 5

Case	Parameter	x	s	τ	p	q	A_{sys}	A_{part}	$R(\tau)$	Cost ($)
1	$A_{min} = 0.999$	1	23	3.728	2	2	0.9992	0.89	0.661	137,348
		0								No Solution
	$A_{min} = 0.99$	0	15	3.710	2	2	0.9901	0.962	0.661	126,407
		1	19	3.751	2	2	0.9901	0.590	0.656	135,599
	$A_{min} = 0.9$	0	9	3.773	2	2	0.9014	0.482	0.651	123,757
		1	0	3.840	2	2	0.9131	0	0.636	127,316
2	$\theta = 0.5$	0	15	3.710	2	2	0.9901	0.962	0.661	126,407
		1	19	3.751	2	2	0.9901	0.590	0.656	135,599
	$\theta = 0.7$	0	15	3.710	2	2	0.9901	0.962	0.661	126,407
		1	19	3.751	2	2	0.9901	0.590	0.656	135,599
	$\theta = 1.1$	0	16	5.068	1	3	0.9902	0.963	0.353	129,623
		1	0	5.068	1	4	0.9902	0.001	0.322	142,776
3	$\alpha = 0.2$	0	15	3.728	2	2	0.9901	0.962	0.661	126,388
		1	19	3.751	2	2	0.9901	0.590	0.656	135,599
	$\alpha = 0.5$	0	24	1.331	6	4	0.9902	0.991	0.745	211,168
		1	24	1.500	5	5	0.9906	0.678	0.656	222,028
	$\alpha = 1.0$	1	33	0.746	10	10	0.9920	0.787	0.661	366,491
		0								No Solution
4	$t_s = 8$	0	15	3.728	2	2	0.9901	0.962	0.661	126,388
		1	19	3.751	2	2	0.9901	0.590	0.656	135,599
	$t_s = 24$	0	18	3.817	2	2	0.9903	0.991	0.641	127,592
		1	19	3.728	2	2	0.9917	0.633	0.661	135,624
	$t_s = 48$	1	19	3.728	2	2	0.9907	0.633	0.661	135,624
		0								No Solution
5	$c_h = 10,000$	0	15	3.706	2	2	0.9901	0.962	0.666	126,412
		1	19	3.751	2	2	0.9901	0.590	0.656	135,599
	$c_h = 20,000$	0	15	3.728	2	2	0.9901	0.962	0.661	129,388
		1	0	3.483	3	2	0.9904	0.001	0.713	137,365
	$c_h = 50,000$	1	0	3.483	3	2	0.9904	0.001	0.713	137,365
		0	15	3.728	2	2	0.9901	0.962	0.661	138,388

and $\tau = 5.202$. If $1/\mu_q = 24$, p becomes 3 while τ drops to 2.679.

In Case 9, the cost of c_p is decreased from \$480k to \$240k. The value of p increases from 2 to 4, while q decreases from 2 to 1. Approximately 89% of field returns are proactive replacements. Conversely, if c_p is increased by 50%, the opposite conclusion can be drawn.

Case 10 investigates how c_q influences the decision variables. When c_q is reduced from \$640k to \$320k, the OEM opts to use more repair servers rather than renewing servers, as expected. Consequently, τ increases from 3.728 to 5.068, a 36% increase. In fact, only 35.3% of field returns are proactive replacements. Conversely, if c_q is increased by 50% from the benchmark cost, the opposite observation can be made.

6.4 Discussion of heuristic solution quality

Particle swarm optimization (PSO) and non-dominated genetic algorithm (GA) are also employed to solve Model RMIA using the benchmark data. The objective is to compare the solution quality of different heuristic algorithms. Both GA and PSO are frequently used to solve reliability, availability, and maintainability problems (Zaretalab et al., 2022), as well as PM planning (Alaswad and Xiang, 2017), and SPL models (Yan et al., 2023). The PSO and GA algorithms are implemented in Matlab and executed on a PC with an Intel(R) Core (TM) i5-7200U CPU @ 2.5GHz, 4 Core(s), 24 GB memory, and 4 Logical Processors.

Table 5 summarizes the optimization results obtained from three algorithms as the fleet size $m$increases from

Table 4 Comparison between redundancy and spares inventory for Cases 6 to 10

Case	Parameter	x	s	τ	p	q	A_{sys}	A_{part}	$R(\tau)$	Cost ($)
6	$\beta = 1.5$	0	11	6.771	1	4	0.9919	0.957	0.207	140,773
		1	7	6.048	1	4	0.9901	0.226	0.264	146,835
	$\beta = 33$	0	15	3.728	2	2	0.9901	0.962	0.661	126,388
		1	19	3.751	2	2	0.9901	0.590	0.656	135,599
	$\beta = 4.5$	0	10	3.262	3	1	0.9905	0.946	0.864	120,742
		1	6	3.308	3	1	0.9901	0.180	0.856	126,407
7	$\frac{1}{\mu_p} = 3$	0	12	2.500	2	1	0.9910	0.964	0.882	115,023
		1	9	2.433	2	1	0.9902	0.324	0.891	121,816
	$\frac{1}{\mu_p} = 6$	0	15	3.728	2	2	0.9901	0.962	0.661	126,388
		1	19	3.751	2	2	0.9901	0.590	0.656	135,599
	$\frac{1}{\mu_p} = 12$	0	17	5.090	2	3	0.9908	0.968	0.348	139,453
		1	9	5.961	1	4	0.9906	0.327	0.184	146,719
8	$\frac{1}{\mu_q} = 6$	0	7	5.202	1	2	0.9914	0.926	0.324	112,741
		1	0	5.202	1	2	0.9941	0.003	0.324	117,176
	$\frac{1}{\mu_q} = 12$	0	15	3.728	2	2	0.9901	0.962	0.661	126,388
		1	19	3.751	2	2	0.9901	0.590	0.656	135,599
	$\frac{1}{\mu_q} = 24$	0	32	2.679	3	2	0.9907	0.987	0.857	145,399
		1	12	2.478	4	2	0.9903	0.434	0.885	154,906
9	$c_p = 240$K	0	13	2.433	4	1	0.9902	0.963	0.891	115,728
		1	11	2.456	4	1	0.9906	0.408	0.888	122,575
	$c_p = 480$K	0	15	3.728	2	2	0.9901	0.962	0.661	126,388
		1	19	3.751	2	2	0.9901	0.590	0.656	135,599
	$c_p = 720$K	0	16	5.068	1	3	0.9902	0.963	0.353	134,223
		1	21	4.934	1	3	0.9910	0.636	0.383	143,841
10	$c_q = 320$K	0	16	5.068	1	3	0.9902	0.963	0.353	110,223
		1	0	5.202	1	4	0.9919	0.001	0.324	117,176
	$c_q = 640K$	0	15	3.728	2	2	0.9901	0.962	0.661	126,388
		1	19	3.751	2	2	0.9901	0.590	0.656	135,599
	$c_q = 960$K	0	32	2.679	3	1	0.9923	0.990	0.857	138,999
		1	11	2.456	4	1	0.9906	0.408	0.888	148,175

10 to 200. In comparison to the PSO and GA, the BS algorithm yields the lowest cost in 15 out of 20 cases. However, for m = 60, PSO proves to be the best option, with a cost lower than BS by $38.94. On the other hand, for m = 90, 110, 130, and 150, GA outperforms both BS and PSO. Nevertheless, the cost difference between GA and BS is relatively small, ranging from $1.01 to $2.69. It is worth noting that in these cases, the values of x, s, p, and q are identical between GA and BS, and the only difference is τ. Similarly, the values of x, s, p, and q are identical between PSO and BS, and the only difference is τ. Furthermore, Table 5 demonstrates that both GA and PSO tend to overestimate the system cost under a small fleet, such as m = 10, 20, 30 and 40. For $m \geqslant 110$, the cost difference among all three algorithms is less than 0.9%, suggesting that BS, GA, and PSO are capable of converging to the lowest cost under a large fleet.

7 Applications to systems with multiple redundant subsystems

In this section we extend the application of Model RMIA to series-parallel systems each comprised of four k -out-of-n subsystems (i.e., $N = 4$). Table 6 provides the parameter values of individual subsystems. The target system availability is set at 0.99, indicating that the availability of each subsystem should be approximately 0.997. First, Algorithms 1 and 2 are used to find the optimal solutions for each subsystem with A_{min} = 0.997. Then,

Table 5 Comparisons of BS, GA and PSO results (underscores are the lowest)

m	x	s	p	q	τ	A_{sys}	Cost	Cost Diff (%)	Algorithm
10	0	2	1	1	4.063	0.9900	**191,147.79**	0.000	**BS**
	0	2	2	2	5.210	0.9952	302,831.52	58.428	GA
	0	2	2	2	5.209	0.9952	302,831.52	36.880	PSO
20	0	7	1	1	3.795	0.9903	**138,598.13**	0.000	**BS**
	0	3	2	2	4.959	0.9900	189,764.39	36.917	GA
	0	3	2	2	4.957	0.9900	189,764.54	36.917	PSO
30	0	9	2	1	3.170	0.9911	**135,742.37**	0.000	**BS**
	0	5	2	2	4.386	0.9900	152,922.17	12.656	GA
	0	5	2	2	4.374	0.9902	152,927.02	12.660	PSO
40	0	21	1	2	4.510	0.9909	**129,934.56**	0.000	**BS**
	0	8	2	2	3.998	0.9900	135,191.86	4.046	GA
	0	8	2	2	3.974	0.9906	135,209.85	4.060	PSO
50	0	15	2	2	3.706	0.9901	**126,407.38**	0.000	**BS**
	0	16	2	2	3.829	0.9900	126,717.73	0.246	GA
	0	16	2	2	3.774	0.9926	126,770.23	0.287	PSO
60	0	15	3	2	3.349	0.9914	126,344.98	0.031	BS
	0	13	2	3	4.270	0.9900	127,384.86	0.854	GA
	0	15	3	2	3.372	0.9905	**126,306.04**	0.000	**PSO**
70	0	30	2	3	4.161	0.9903	**125,154.35**	0.000	**BS**
	0	13	3	3	3.865	0.9900	126,991.05	1.468	GA
	0	32	2	3	4.141	0.9948	125,780.55	0.500	PSO
80	0	22	3	3	3.706	0.9901	**123,073.06**	0.000	**BS**
	0	22	2	4	4.523	0.9900	124,541.17	1.193	GA
	0	23	3	3	3.775	0.9901	123,269.10	0.159	PSO
90	0	21	4	3	3.488	0.9906	123,109.26	0.001	BS
	0	21	4	3	3.488	0.9900	**123,107.83**	0.000	**GA**
	0	21	4	3	3.477	0.9909	123,123.63	0.013	PSO
100	0	38	3	4	4.018	0.9914	**123,057.65**	0.000	**BS**
	0	17	3	5	4.376	0.9900	124,731.38	1.360	GA
	0	39	3	4	3.980	0.9907	123,299.45	0.196	PSO
110	0	29	4	4	4.465	0.9903	121,513.23	0.001	BS
	0	29	4	4	3.745	0.9900	**121,512.23**	0.000	**GA**
	0	29	4	4	3.713	0.9917	121,545.98	0.028	PSO
120	0	26	5	4	3.527	0.9901	**121,363.67**	0.000	**BS**
	0	24	4	5	4.025	0.9900	121,838.51	0.391	GA
	0	26	5	4	3.492	0.9917	121,408.94	0.037	PSO
130	0	25	6	4	3.349	0.9903	121,614.60	0.002	BS
	0	25	6	4	3.349	0.9900	**121,611.91**	0.000	**GA**
	0	22	5	5	3.760	0.9907	121,811.97	0.165	PSO
140	0	35	5	5	3.743	0.9901	**120,495.55**	0.000	**BS**
	0	35	5	5	3.740	0.9900	120,496.74	0.001	GA
	0	37	5	5	3.744	0.9937	120,800.53	0.253	PSO

(*Continued*)

m	x	s	p	q	τ	A_{sys}	Cost	Cost Diff (%)	Algorithm
150	0	31	6	5	3.559	0.9903	120,307.38	0.001	BS
	0	31	6	5	3.558	0.9900	**120,305.60**	0.000	**GA**
	0	28	5	6	3.930	0.9902	120,556.70	0.209	PSO
160	0	29	7	5	3.393	0.9904	**120,459.42**	0.000	**BS**
	0	25	5	7	4.104	0.9900	121,176.29	0.595	GA
	0	26	6	6	3.716	0.9907	120,636.36	0.147	PSO
170	0	40	6	6	3.706	0.9915	**119,746.86**	0.000	**BS**
	0	29	8	5	3.288	0.9900	120,812.72	0.890	GA
	0	42	6	6	3.669	0.9915	120,039.51	0.244	PSO
180	0	36	7	6	3.572	0.9918	**119,613.35**	0.000	**BS**
	0	44	8	5	3.258	0.9900	120,155.51	0.453	GA
	0	39	7	6	3.502	0.9943	120,061.78	0.375	PSO
190	0	33	8	6	3.438	0.9900	**119,650.11**	0.000	**BS**
	0	30	7	7	3.732	0.9900	119,782.91	0.111	GA
	0	34	8	6	3.444	0.9916	119,752.30	0.085	PSO
200	0	42	6	8	4.024	0.9903	**119,390.37**	0.000	**BS**
	0	42	6	8	4.019	0.9900	119,391.56	0.001	GA
	0	29	8	7	3.600	0.9901	119,991.31	0.503	PSO

Table 6 Reliability and cost data for series-parallel system (n/a=not applicable)

Subsystem	$i=1$	$i=2$	$i=3$	$i=4$	Unit
α	0.25	0.3	0.35	0.4	failure/year
β	1.5	2.5	3	3.5	n/a
k	10	7	5	3	item
n_{max}	13	10	7	5	item
t_s	12	18	24	30	hour
$\frac{1}{\mu_p}$	6	9	12	15	day
$\frac{1}{\mu_q}$	12	14	18	21	day
c_{LRU}	60,000	90,000	110,000	130,000	$/item
c_u	3,000	5,000	6,000	7,500	$/item
c_v	4,500	7,500	9,500	11,250	$/item
c_h	12,000	18,000	22,000	26,000	$/item/year
c_p	250,000	320,000	375,000	420,000	$/server
c_q	350,000	384,000	494,000	630,000	$/server
$\gamma_{min}, \gamma_{max}$	0.5, 2	0.5, 2	0.5, 2	0.5, 2	n/a

Algorithm 3 utilizes neighborhood exploration to optimize the overall problem with $A_{min} = 0.99$.

The results under different fleet sizes are summarized in Table 7. The following observations can be made: first, as expected, the system cost decreases as m increases from 10 to 100. Specifically, the cost is \$776,664 for $m = 10$, and \$610,007 for $m = 100$, down by 21.5%. Secondly, Subsystem 1 opts for redundancy for $m = 10$ and 20. However, as m becomes larger, redundancy is no longer the preferred option. Subsystems 2 and 3, on the other hand, prefer to install one redundant component regardless of the fleet size. For Subsystem 4, redundancy is never the option regardless of the fleet size. This is because the unit cost of the LRU for Subsystem 4 is the highest among the four subsystems, and only 3 working units are required, compared to 5, 7 and 10 for the other subsystems. Consequently, the marginal cost of using one redundant unit is considerably higher for Subsystem 4.

8 Conclusions

This study proposes a joint redundancy-maintenance-inventory allocation model to minimize the annualized system cost while achieving the desired reliability and availability targets during the lifetime period. This model is the first of its kind in bringing together the decisions of reliability-redundancy, preventive maintenance, and spare parts logistics. Two parallel Erlang-C queues are utilized to characterize the decentralized repair and renewal shops, respectively. The demand for fleet spare parts is modeled as a superimposed renewal process, consisting of proactive and failure placement streams. To solve the redundancy-maintenance-inventory allocation model, a bisection

Table 7 The solution for systems with four redundant subsystems

m	10	20	30	40	50	60	70	80	90	100
Cost	776,664	701,735	666,538	645,222	634,393	627,222	627,411	619,270	612,960	610,007
A_{sys}	0.99001	0.99001	0.99001	0.99000	0.99001	0.99000	0.99000	0.99000	0.99001	0.99000
x_1	1	1	0	0	0	0	0	0	0	0
x_2	1	1	1	1	1	1	1	1	1	1
x_3	1	1	1	1	1	1	1	1	1	1
x_4	0	0	0	0	0	0	0	0	0	0
s_1	5	7	15	17	17	19	25	29	32	34
s_2	3	6	10	13	16	19	14	16	18	20
s_3	4	8	8	12	16	20	23	20	23	25
s_4	4	8	16	15	15	18	19	25	26	26
τ_1	3.033	4.044	5.416	5.416	5.416	5.416	5.416	5.416	5.416	5.416
τ_2	2.810	3.372	3.667	3.845	4.022	4.141	3.224	3.342	3.460	3.519
τ_3	2.398	2.781	2.679	2.322	2.526	2.653	2.730	2.424	2.526	2.602
τ_4	2.272	2.564	2.497	2.699	2.834	2.497	3.104	2.969	2.654	2.744
p_1	1	1	1	1	1	1	1	1	2	2
p_2	1	1	1	1	1	1	3	3	3	3
p_3	1	1	2	3	3	3	3	5	5	5
p_4	1	1	1	1	1	2	1	1	2	2
q_1	1	2	3	4	5	6	7	8	8	9
q_2	1	2	3	4	5	6	6	7	8	9
q_3	1	2	3	3	4	5	6	6	7	8
q_4	1	2	2	3	4	4	6	6	6	7

search algorithm that combines neighborhood exploration is developed. The numerical experiments provide several important insights. First, redundancy is preferred over spare parts when the fleet size is small, inventory holding costs are high, replacement time is extended, or extremely high system availability, such as 0.999, is required. Second, there is no monotonic correlation between spares inventory level, parts availability, and system availability in the joint allocation model. Third, both the spares inventory and the system cost decrease as the Weibull shape parameter increase, suggesting that age-based replacement becomes more cost-effective for items with a concentrated lifetime distribution.

In the future, the redundancy-maintenance-inventory model can be expanded in several directions. For example, with the increasing use of prognostics and health management systems, condition-based maintenance can be integrated into the joint allocation model. This will help prevent and reduce random failures, thereby improving spares provisioning efficiency. Additionally, multi-class queues can be employed to model repair and renewal tasks in a centralized facility. However, this may require theoretical advancements as current multi-class queueing models become computationally burdensome when dealing with multiple servers.

Competing Interests The authors declare that they have no competing interests.

Appendix A: Notation of model parameters

Table A1 Model parameters

Notation	Definition
α, β	Weibull scale and shape parameters, respectively
λ_p	Part demand rate of a single-item system in planned replacement
λ_q	Part demand rate of a single-item system in failure replacement
λ_m	Aggregate part demand rate of a fleet of single-item systems
$\lambda_{F,p}$	Aggregate part demand rate of a fleet in planned replacement
$\lambda_{F,q}$	Aggregate part demand rate of a fleet in failure replacement
λ_F	Aggregate part demand rate of a fleet, and $\lambda_F = \lambda_{F,p} + \lambda_{F,q}$
ρ_p, ρ_q	Part renewing and repair traffic intensity rate, respectively
φ_1	Capital recovery factor of system
φ_2	Capital recovery factor of spare part or LRU
θ	Percentage of mean time between failures of LRU
μ	Number of returned items during parts turn-around time
μ_p, μ_q	Part renewing rate and repair rate, respectively

(*Continued*)

Notation	Definition
$\tau_{\min}, \tau_{\max}$	The lower and upper limit of maintenance intervals
$\gamma_{\min}, \gamma_{\max}$	Minimum and maximum percentage of MTBF for LRU
c_{LRU}	Unit cost for a spare part or LRU
c_h	Unit holding cost per year
c_u, c_v	Cost for renewing and repairing a part, respectively
c_p, c_q	Cost for operating a renewing and repairing server, respectively
k	The minimum number of required working item in a system
m	System fleet size or installed base
$n_{\max}$	Maximum number of components a system can install
t_{ATT}	Average part turn-around time
t_p	Part renewing turn-around time
t_q	Part repair turn-around time
t_s	Hands-on time for replacing a part
$B(q)$	Probability for a part waiting in a repair shop
$C(p)$	Probability for a part waiting in a renewing shop
A	Availability of a single-item system
A_R	Availability of a k-out-of-n redundant system
$A_{\min}$	Target or desired system availability
A_{sys}	Actual system availability
A_{part}	Actual parts availability
$f(t)$	Probability density functions of the LRU lifetime
$R(t)$	Reliability function of the LRU
$F(t)$	Cumulative distribution function of the LRU lifetime
O	Steady-state inventory on-order, a random variable
T_p	Mean downtime in a planned replacement
T_q	Mean downtime in a failure replacement
$\overline{T}$	Mean-time-between-failures of the LRU
N	Number of redundant subsystems in a system

Appendix B: The range of decision variables

Th range of the decision variables is analyzed to reduce the search space of Model RMIA. For a given subsystem, the upper limit of x is governed by Constraint (21), namely $x \leqslant n_{\max} - k$. Hence the effort below is focused on τ, s, p and q.

B.1. The range of τ

The maintenance interval τ is correlated with MTBF denoted as $\overline{T}$. For Weibull distribution, $\overline{T} = (1/\alpha)\Gamma(1+1/\beta)$. Figure B1 plots the Weibull reliability $R(\tau)$ in three cases: $\tau = 0.5\overline{T}$, $\overline{T}$, and $2\overline{T}$. When β increases from 1 to 6, we find that $R(0.5\overline{T})$ increases from 0.61 to 1, and $R(2\overline{T})$ decreases from 0.14 to 0. If $\tau = 2\overline{T}$, the chance of making a proactive replacement is only 0.09. Hence τ should not exceed $2\overline{T}$. Otherwise, over 91% of replacements are due to failures. Thus the range of τ shall fall in

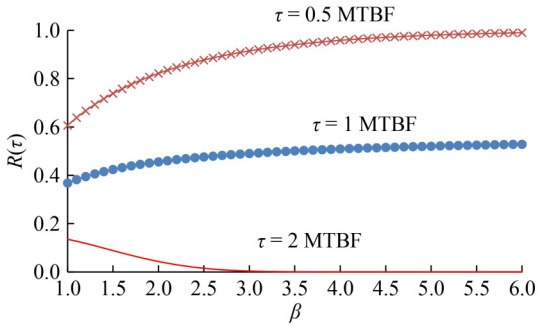

Fig. B1 Weibull reliability for τ = 0.5, 1 and 2MTBF

$(0, 2\overline{T}]$. Note that the reliability curves in Fig. B1 are independent to the scale parameter α.

B.2. The range of q and p

The Erlang-C queue is stable if and only if the repair traffic intensity rate $\rho_q = \lambda_{F,q}/(q\mu_q) < 1$. This implies that:

$$q > \frac{\lambda_{F,q}}{\mu_q}. \tag{B1}$$

The value of $\lambda_{F,q}$ depends on the maintenance interval $\tau \in [\tau_{\min}, \tau_{\max}]$. The lower limit of $\lambda_{F,q}$ occurs at $\tau = \tau_{\min}$. For m systems each with k-out-of-n configuration, the aggregate failure replacement rate is given as

$$\lambda_{F,q}(\tau_{\min}) = \frac{m(k+x)F(\tau_{\min})}{\int_0^{\tau_{\min}} R(t)\,\mathrm{d}t}. \tag{B2}$$

Substituting Eq. (B2) into Eq. (B1) yields the lower limit for q as follows:

$$q_L = \left\lceil \frac{\lambda_{F,q}(\tau_{\min})}{\mu_q} \right\rceil, \tag{B3}$$

where $\lceil a \rceil$ represents the smallest integer greater than a. Similarly, the upper limit of q is found at $\tau = \tau_{\max}$. That is:

$$q_U = \left\lceil \frac{\lambda_{F,q}(\tau_{\max})}{\mu_q} \right\rceil, \tag{B4}$$

where

$$\lambda_{F,q}(\tau_{\max}) = \frac{m(k+x)F(\tau_{\max})}{\int_0^{\tau_{\max}} R(t)\,\mathrm{d}t}. \tag{B5}$$

The derivation of lower limit of p is similar to q, and the results are given below. The fleet generates the smallest proactive replacements when $\tau = \tau_{\max}$, and its rate is

$$\lambda_{F,p}(\tau_{\max}) = \frac{m(k+x)R(\tau_{\max})}{\int_0^{\tau_{\max}} R(t)\,\mathrm{d}t}. \tag{B6}$$

Hence the lower limit of p is obtained by

$$p_L = \left\lceil \frac{\lambda_{F,p}(\tau_{\max})}{\mu_p} \right\rceil. \tag{B7}$$

Similarly, when $\tau = \tau_{\min}$, the fleet generates the largest proactive replacements, and the rate is

$$\lambda_{F,p}(\tau_{\min}) = \frac{(k+x)mR\tau_{\min}}{\int_0^{\tau_{\min}} R(t)\,\mathrm{d}t}. \tag{B8}$$

Thus the upper limit of p is given as

$$p_U = \left\lceil \frac{\lambda_{F,p}(\tau_{\min})}{\mu_p} \right\rceil. \tag{B9}$$

B.3. The range of *s*

Since the lower limit of s is zero, we just need to find its upper limit. Given x, τ, p, and q, the value of s increases with $A_{\min}$. According to Eq. (17), the redundant system availability must satisfy the following condition

$$\sum_{j=k}^{k+x} \binom{k+x}{j} \tilde{A}^j \left(1-\tilde{A}\right)^{k+x-j} \geqslant A_{\min}, \tag{B10}$$

where $\tilde{A}$ is the smallest component availability given in Eq. (16). After the re-arrangement, Eq. (16) becomes:

$$\sum_{j=0}^{s} \frac{\mu^j e^{-\mu}}{j!} \geqslant 1 + \frac{t_s + (1-\tilde{A}^{-1})\int_0^{\tau} R(t)\,\mathrm{d}t}{t_{ATT}}, \tag{B11}$$

where

$$\mu = \lambda_F \left(t_p R(\tau) + t_q F(\tau)\right) = \frac{m(k+x)t_{ATT}}{\int_0^{\tau} R(t)\,\mathrm{d}t}. \tag{B12}$$

Based on Eq. (B11) the upper limit for s can be derived using the procedure as follows.

Step 1: For given $\tau \in [\tau_{\min}, \tau_{\max}]$, estimate the q_L, q_U, p_L and p_U according to Eqs. (B3), (B4), (B7), and (B9), respectively.

Step 2: Compute the values of $B(q_L)$, $B(q_U)$, $C(p_L)$, and $C(p_U)$ based on Eqs. (5) and (7), respectively.

Step 3: Based on Eq. (9), compute t_{ATT} for $\tau = \tau_{\min}$ and $\tau = \tau_{\max}$, respectively.

Step 4: Based on Eq. (B12), compute μ for $\tau = \tau_{\min}$ and $\tau = \tau_{\max}$, respectively.

Step 5: Find $s(\tau_{\min})$ and $s(\tau_{\max})$ that satisfy Eq. (B11) with respect to $\tau = \tau_{\min}$ and $\tau = \tau_{\max}$.

Step 6: choose $s_{\max} = \max\{s(\tau_{\min}), s(\tau_{\max})\}$ as the upper limit of s.

The rationality of this 6-step procedure is that the upper limit of s occurs when p and q are in their lower limit either at $\tau = \tau_{\min}$ or $\tau = \tau_{\max}$. If p or q is above their lower limit, t_{ATT} deceases and μ becomes smaller. Hence less amounts of spare parts are needed to meet $A_{\min}$.

References

Alaswad S, Xiang Y (2017). A review on condition-based maintenance optimization models for stochastically deteriorating system. Reliability Engineering & System Safety, 157: 54–63

Barnett E, Gosselin C (2021). A bisection algorithm for time-optimal trajectory planning along fully specified paths. IEEE Transactions on Robotics, 37(1): 131–145

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

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

Bei X, Chatwattanasiri N, Coit D W, Zhu X (2017). Combined redundancy allocation and maintenance planning using a two-stage stochastic programming model for multiple component systems. IEEE Transactions on Reliability, 66(3): 950–962

Bei X, Zhu X, Coit D W (2019). A risk-averse stochastic program for integrated system design and preventive maintenance planning. European Journal of Operational Research, 276(2): 536–548

Bjarnason E T S, Taghipour S (2016). Periodic inspection frequency and inventory policies for a *k*-out-of-*n* system. IIE Transactions, 48(7): 638–650

Bjarnason E T S, Taghipour S, Banjevic D (2014). Joint optimal inspection and inventory for a *k*-out-of-*n* system. Reliability Engineering & System Safety, 131: 203–215

Chen L, Ye Z S, Xie M (2013). Joint maintenance and spare component provisioning policy for *k*-out-of-*n* systems. Asia-Pacific Journal of Operational Research, 30(6): 1350023

Coit D W, Zio E (2019). The evolution of system reliability optimization. Reliability Engineering & System Safety, 192: 106259

Cox D R, Smith W L (1954). On the superposition of renewal processes. Biometrika, 41(1–2): 91–99

de Smidt-Destombes K S, van der Heijden M C, van Harten A (2009). Joint optimisation of spare part inventory, maintenance frequency and repair capacity for *k*-out-of-*n* systems. International Journal of Production Economics, 118(1): 260–268

Dekker R, Pinçe Ç, Zuidwijk R, Jalil M N (2013). On the use of installed base information for spare parts logistics: A review of ideas and industry practice. International Journal of Production Economics, 143(2): 536–545

Díaz A, Fu M (1997). Models for multi-echelon repairable item inventory systems with limited repair capacity. European Journal of Operational Research, 97(3): 480–492

El-Ferik S (2008). Economic production lot-sizing for an unreliable machine under imperfect age-based maintenance policy. European Journal of Operational Research, 186(1): 150–163

Hekimoğlu M, van der Laan E, Dekker R (2018). Markov-modulated analysis of a spare parts system with random lead times and disruption risks. European Journal of Operational Research, 269(3): 909–922

Hu Y, Miao X, Si Y, Pan E, Zio E (2022). Prognostics and health management: A review from the perspectives of design, development and decision. Reliability Engineering & System Safety, 217: 108063

Huynh K T, Castro I T, Barros A, Bérenguer C (2012). Modeling age-based maintenance strategies with minimal repairs for systems subject to competing failure modes due to degradation and shocks. European Journal of Operational Research, 218(1): 140–151

Jin T (2023). Bridging reliability and operations management for superior system availability: Challenges and opportunities. Frontiers of Engineering Management, 10(3): 391–405

Jin T, Li H, Sun F (2021). System availability considering redundancy, maintenance and spare parts with dual repair processes. In: Proceedings of Industrial and Systems Engineer Conference, Montreal, Canada, 1–6

Jin T, Taboada H, Espiritu J, Liao H (2017). Allocation of reliability-redundancy and spares inventory under Poisson fleet expansion. IISE Transactions, 49(7): 737–751

Jin T, Tian Y (2012). Optimizing reliability and service parts logistics for a time-varying installed base. European Journal of Operational Research, 218(1): 152–162

Jin T, Tian Z, Xie M (2015). A game-theoretical approach for optimizing maintenance, spares and service capacity in performance contracting. International Journal of Production Economics, 161: 31–43

Kim S H, Cohen M A, Netessine S (2007). Performance contracting in after-sales service supply chains. Management Science, 53(12): 1843–1858

Lee H L (1987). A multi-echelon inventory model for repairable items with emergency lateral transshipments. Management Science, 33(10): 1302–1316

Levitin G, Lisnianski A (1999). Joint redundancy and maintenance optimization for multistate series–parallel systems. Reliability Engineering & System Safety, 64(1): 33–42

Liu Y, Huang H Z, Wang Z, Li Y, Yang Y (2013). A joint redundancy and imperfect maintenance strategy optimization for multi-state systems. IEEE Transactions on Reliability, 62(2): 368–378

Louit D, Pascual R, Banjevic D, Jardine A K S (2011). Optimization models for critical spare parts inventories—a reliability approach. Journal of the Operational Research Society, 62(6): 992–1004

Moghaddass R, Zuo M J, Pandey M (2012). Optimal design and maintenance of a repairable multi-state system with standby components. Journal of Statistical Planning and Inference, 142(8): 2409–2420

Mouatasim A E (2018). Implementation of reduced gradient with bisection algorithms for non-convex optimization problem via stochastic perturbation. Numerical Algorithms, 78(1): 41–62

Nourelfath M, Châtelet E, Nahas N (2012). Joint redundancy and imperfect preventive maintenance optimization for series–parallel multi-state degraded systems. Reliability Engineering & System Safety, 103: 51–60

Olde Keizer M C A, Teunter R H, Veldman J (2017). Joint condition-based maintenance and inventory optimization for systems with multiple components. European Journal of Operational Research, 257(1): 209–222

Öner K B, Scheller-Wolf A, van Houtum G J (2013). Redundancy optimization for critical components in high-availability technical systems. Operations Research, 61(1): 244–264

Reddy S S, Bijwe P R (2018). An efficient optimal power flow using bisection method. Electrical Engineering, 100(4): 2217–2229

Selçuk B, Agrali S (2013). Joint spare parts inventory and reliability decisions under a service constraint. Journal of the Operational Research Society, 64(3): 446–458

Selviaridis K, Wynstra F (2015). Performance-based contracting: a literature review and future research directions. International Journal of Production Research, 53(12): 3505–3540

Si S, Zhao J, Cai Z, Dui H (2020). Recent advancement in system reliability optimization driven by importance measures. Frontiers of Engineering Management, 7(3): 335–358

Sleptchenko A, van der Heijden M C (2016). Joint optimization of redundancy level and spare part inventories. Reliability Engineering & System Safety, 153: 64–74

Sleptchenko A, van der Heijden M C, van Harten A (2002). Effects of finite repair capacity in multi-echelon, multi-indenture service part supply systems. International Journal of Production Economics, 79(3): 209–230

Sleptchenko A, van der Heijden M C, van Harten A (2003). Trade-off between inventory and repair capacity in spare part networks. Journal of the Operational Research Society, 54(3): 263–272

Van Horenbeek A, Scarf P, Cavalcante C, Pintelon L (2013). The effect of maintenance quality on spare parts inventory for a fleet of assets. IEEE Transactions on Reliability, 62(3): 596–607

Vaughan T S (2005). Failure replacement and preventive maintenance spare parts ordering policy. European Journal of Operational Research, 161(1): 183–190

Wang J, Zhu X (2021). Joint optimization of condition-based maintenance and inventory control for a *k*-out-of-*n*: F system of multi-state degrading components. European Journal of Operational Research, 290(2): 514–529

Wang L, Chu J, Mao W (2009). A condition-based replacement and spare provisioning policy for deteriorating systems with uncertain deterioration to failure. European Journal of Operational Research, 194(1): 184–205

Wang W (2012). A stochastic model for joint spare parts inventory and planned maintenance optimization. European Journal of Operational Research, 216(1): 127–139

Wang Z (2021). Current status and prospects of reliability systems engineering in China. Frontiers of Engineering Management, 8(4): 492–502

Winston W (2004). Operations Research: Applications and Algorithms, 4th ed., Chapter 20, pp. 1051–1131, Brooke/Cole Cengage Learning, Belmont, CA, USA

Wu S (2019). Superimposed Renewal Processes in Reliability. Wiley Stats Ref: Statistics Reference

Wu S (2021). Two methods to approximate the superposition of imperfect failure processes. Reliability Engineering & System Safety, 207: 107332

Xie W, Liao H, Jin T (2014). Maximizing system availability through joint decision on redundancy allocation and spares inventory. European Journal of Operational Research, 237(1): 164–176

Yan B, Zhou Y, Zhang M, Li Z (2023). Reliability-driven multiechelon inventory optimization with applications to service spare parts for wind turbines. IEEE Transactions on Reliability, 72(2): 748–758

Zaretalab A, Sharifi M, Guilani P P, Taghipour S, Niaki S T A (2022). A multi-objective model for optimizing the redundancy allocation, component supplier selection, and reliable activities for multi-state systems. Reliability Engineering & System Safety, 222: 108394

Zhang J, Zhao X, Song Y, Qiu Q (2022). Joint optimization of condition-based maintenance and spares inventory for a series–parallel system with two failure modes. Computers & Industrial Engineering, 168: 108094

Zhang S, Huang K, Yuan Y (2021). Spare parts inventory management:

A literature review. Sustainability, 13(5): 2460

Zhao X, Zhang J, Wang X (2019). Joint optimization of components redundancy, spares inventory and repairmen allocation for a standby series system. Proceedings of the Institution of Mechanical Engineers. Part O, Journal of Risk and Reliability, 233(4): 623–638

Zhu S, Jaarsveld W, Dekker R (2020). Spare parts inventory control based on maintenance planning. Reliability Engineering & System Safety, 193: 106600

Zhu X, Bei X, Chatwattanasiri N, Coit D W (2018). Optimal system design and sequential preventive maintenance under uncertain aperiodic-changing stresses. IEEE Transactions on Reliability, 67(3): 907–919

Zhu X, Wang J, Coit D W (2022). Joint optimization of spare part supply and opportunistic condition-based maintenance for onshore wind farms considering maintenance route. IEEE Transactions on Engineering Management, 71: 1086–1102