Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Reliability-based measure of a retrial machine repair problem with working breakdowns under the F-policy

Tseng-Chang Yen a, Kuo-Hsiung Wang b,*, Chia-Huang Wu c

a Department of Applied Mathematics, National Chung-Hsing University, Taichung 40227, Taiwan
b Department of Business Administration, Asia University, Wufeng, Taichung 41354, Taiwan
c Department of Industrial Engineering and Management, National Chiao-Tung University, Hsinchu 30010, Taiwan

ARTICLE INFO

Keywords:
- F-policy
- Mean time to system failure
- Retrial machine repair problem
- Sensitivity analysis
- System reliability
- Working breakdowns

ABSTRACT

We study reliability and sensitivity analysis of a retrial machine repair problem with working breakdowns operating under the F-policy. The F-policy studies the most common issue of controlling arrival to a queueing problem and it requires startup time before allowing failed machines to enter the orbit. The server is subject to working breakdowns only when there is at least one failed machine in the system. When the server is busy it works at a fast rate, but when it is subject to working breakdowns, it works at a slow rate. Failure and repair times of the server are exponentially distributed. The Laplace transform technique is utilized to develop two system performance measures such as system reliability and mean time to system failure (MTTF). Extensive numerical results are carried out to reveal how performance measures are affected by changing various values of each system parameter.

1. Introduction

Cloud computing is a burgeoning computing paradigm and has gained tremendous popularity in recent years. Cloud computing paradigm shifts the deployment of computing infrastructure (such as CPU, network, and storage) from end users to the cloud data center. The provision of cloud computing services is delivered over the Internet via virtual machines (VMs) and virtual computing resources. Cloud services provide scalable and flexible infrastructure for Internet of things (IoT), machine learning, blockchain, computer-assisted instruction (CAI), and etc. In the COVID-19 pandemic era, many countries have decided to temporarily close schools to reduce contact and save lives. To avoid disrupting learning and upending lives, teaching is moving online on an untested and unprecedented scale. Online-learning is an Internet-based CAI system, which uses Internet technology to improve the efficiency of education. By using the benefits of cloud computing, the online-learning cloud system supports the ability to dynamically allocate computing and storage resources for teaching material. To keep the quality of online-learning service and the stability of real-time data transition, the reliability of the cloud system should be investigated. The maintenance of online-learning cloud system can be modelled as a retrial machine repair problem.

The retrial machine repair problem has been widely adopted to study many real-world systems, such as call centers, telecommunication systems, cloud computing networks, etc. In this paper, we consider a retrial machine repair problem (RMRP) with working breakdowns that combines F-policy and exponentially startup time before allowing failed machines to enter the retrial system. The server may be subject to working breakdown while serving a failed machine. When a server is busy or subject to working breakdowns, failed machines join the retrial orbit. Arriving failed machines which cannot receive service immediately join the retrial orbit for a certain period of time and retry to receive service. Maintaining a high level of system reliability is generally a requisite for many real-world repairable systems.

For the most comprehensive concepts in retrial queues, we referred the reader to Artalejo (1999a, 1999b), Artalejo and Gomez-Corral (2008), and Phung-Duc (2019). The literature on retrial systems was very vast and rich. Artalejo and Falin (2002) provided extensive comparisons of the standard and retrial queueing systems. Sherman and Kharoufeh (2006) studied an unreliable M/M/1 retrial queue with infinite-capacity orbit and normal queue. The authors derived stability conditions and some important stochastic decomposability results. Wang (2006) extended Sherman and Kharoufeh’s work to study an M/G/1/K retrial queue with server breakdowns. Efrosinin and Winkler

https://doi.org/10.1016/j.cie.2020.106885
Received 3 February 2020; Received in revised form 15 August 2020; Accepted 2 October 2020
Available online 6 October 2020
0360-8352/© 2020 Elsevier Ltd. All rights reserved.
(2011) proposed a Markovian retrial queue with constant retrial rate and an unreliable server that combines the threshold recovery policy. The authors developed the important results for the waiting time distribution and the optimal threshold level. The authors provided the extensive results of the arrival distribution, the busy period, and the waiting time process. Reliability analysis of the retrial queue with server breakdown and repairs was studied by Wang, Cao, and Li (2001). Gharbi and Ioualalen (2006) considered finite-source retrial systems with multiple server breakdowns and repairs using generalized stochastic petri nets model. The authors developed steady-state performance and reliability indices. Moreover, they also employed sensitivity analysis with numerical illustration. Ke, Yang, Sheu, and Kuo (2013) presented the availability analysis of a repairable retrial system with warm standbys. Kuo, Sheu, Ke, and Zhang (2014) investigated reliability analysis of a retrial system with mixed standbys. The authors developed steady-state availability as well as the steady-state availability. Chen (2016) proposed reliability analysis of repairable machine repair system with working breakdowns and a single repair server with recovery policy. Both sensitivity analyses and relative sensitivity analyses of system reliability and mean time to system failure were performed. Chen and Wang (2018) analyzed the system reliability of the N-policy retrial machine repair problem with a single controllable server. Numerical experiments were provided to reveal how performance measures are affected by the change of each system parameter.

For F-policy queues, Wang, Kuo, and Pearn (2008) considered the optimal control of a G/M/1/K system combining F-policy and exponentially startup time before starting to allow clients in the system. The authors applied the supplementary variable technique to develop the steady-state probability distribution of the number of clients in the system. Wang and Yang (2009) examined the F-policy M/M/1/K system with server breakdown. Quasi-Newton methods and the direct search were adopted to determine the optimal threshold F, the optimal service rate, and the optimal startup rate at minimum cost. An F-policy M/M/1/K system with working vacations and exponentially startup time was explored by Yang, Wang, and Wu (2010). Extensive numerical results were carried out to reveal the influence of system parameters on the cost function. For the systematic developments of the control F-policy systems, one can refer to the research papers by Jain, Sanga, & Meena (2016), Jain, Shekhar, & Shukla (2016) and Chang, Lee, Chang, and Yeh (2018). The fault tolerance problem of repairable redundant system with general retrial times operating under F-policy was analyzed by Jain and Sanga (2017). Numerical results were provided to study the effects of different system parameters on various performance measures. Shekhar, Jain, Iqbal, and Raina (2017) proposed the computational approach to calculate the transient and steady state probabilities of the machine repair problem with an unreliable server operating under F-policy. Greedy selection and Newton-quasi methods were used to determine the optimal solutions at minimum cost. Jain and Sanga (2019a) studied F-policy M/M/1/K retrial queueing system with state-dependent rates. A cost function was conducted to determine the optimal service rate at minimum cost. The optimal control F-policy for M/M/R/K queue with an additional server and balking was investigated by Jain and Sanga (2019b). A cost model was framed to determine the optimal threshold parameter and optimal service rate. The finite state dependent queueing model with general retrial times operating under F-policy was explored by Jain and Sanga (2019c). Sensitivity analysis of the cost function with numerical illustrations were provided. Jain and Sanga (2020a) provided a state-of-the-art and survey literature review of the state-dependent queueing models operating under F-policy. Recently, Jain and Sanga (2020b) investigated the machine repair system with general retrial times operating under F-policy in fuzzy environment. The signed distance method was used to defuzzify the cost function. The authors applied the genetic algorithm to find the optimal control parameter and corresponding minimum cost. To the best of our knowledge, system reliability and sensitivity analysis of a RMRP with working breakdowns operating under the F-policy has never been discussed in the literature.

This paper differs from previous work in that:

(i) Queueing problems are broadly divided into two categories, one aims at controlling service category (N-policy) whereas the other aims at controlling arrivals category (F-policy). The F-policy reliability problem has distinct characteristics which are different from the N-policy reliability problem.

(ii) It aims at controlling arrivals to study reliability characteristics of a retrial machine repair problem;

(iii) We simultaneously consider the F-policy and working breakdowns which has never been investigated in the existing work.

The rest of the paper is organized as follows. Section 2 gives the descriptions of the problem. In Section 3, we use the Laplace transform technique to develop the main results such as system reliability and mean time to system failure. Sensitivity and relative sensitivity analysis is also addressed in this section. In Section 4, extensive numerical results were carried out to reveal the influence of system parameters on related analysis such as system reliability, mean time to system failure, sensitivity, and relative sensitivity. Section 5 concludes the paper.

2. Descriptions of the problem

We propose a retrial machine repair problem (RMRP) with N = M + S identical and independent machines and a single server in the repair facility. As many as M machines can be operated simultaneously in parallel, the rest of the S machines are available as warm-standby spares. Operating machines fails according to independent Poisson distribution with parameter λ. Similarly, each of the available warm standbys fails independently of the state of all the others with Poisson rate η (0 < η < λ). Whenever one of operating machines fails, it is immediately replaced by a warm standby, if ready. We assume that when a warm standby machine moves into an operating state, its failure characteristics will be that of an operating machine. Each failed operating or standby machine is instantly sent to a single server. The server is unreliable and may experience a working breakdown during busy periods. When the server is busy it serves at a fast service rate μ₂, but when it is subject to working breakdowns, in place of terminating service completely, it continues at a lower service rate μ₃ (< μ₂). The service time distribution is assumed to be exponentially distributed. It is assumed that the server may encounter working breakdown at any time with breakdown rate a. Whenever the server breaks down, it is immediately repaired at a repair rate β. Breakdown and repair times of the server are assumed to be exponentially distributed. When the server recovers from a working breakdown, the service rate is enhanced from μ₂ to μ₁. Whenever a machine fails, it is immediately sent to the repair facility where repair work is provided in the order of breakdowns; that is, first-come, first-served (FCFS) discipline. Arriving failed machines discern that the server is busy and must enter the orbit after a certain period of time and retry to receive service. Each failed machine in the orbit repeats its service request with retrial time. Retrial times follow the exponential distribution at rate γ.

Furthermore, to control arriving failed machines, an F-policy is implemented. The definition of an F-policy is described below: When the number of failed machines in the system reaches full capacity K, no further arriving failed machines are allowed to enter the system until the number of failed machines in orbit decreases to the threshold value F (0 ≤ F < K − 1). At that moment, the server must adopt exponentially start up time with parameter θ to start allowing failed machines to join the system. Thus, the system operates continuously until the number of failed machines in the system reaches to full capacity K again at which time the above process is repeated over and over. According to the F-policy, to ensure that the server can operate normally, the repair facility prevents new failed machines from entering with probability c. On the contrary, when system is full, the entering of new failed machine causes
the server to be unable to complete the repair due to insufficient system resources. The system then moves to the unsafe failure (uf) state with probability 1−c, and the occurrence of the unsafe failure always results in a system fail.

2.1. Practical illustration of the model

A practical situation related to an online-learning cloud system is presented for illustrative purposes. To satisfy the requests of extensive concurrent online-learning connections, an online-learning cloud system is established via a VM cluster on four quad-core servers. Since there are total 16 CPU cores (4 × 4 cores), 16 VMs (each VM is assigned with a CPU core) are deployed to establish a private cloud (Fig. 1) which 12 VMs served as working online-learning servers, and 4 VMs as warm-standby servers. Meanwhile, one physical server with Linux system and pre-installed Cloudstack cloud administration tool is deployed as the VM repair server. Besides, a 1 TB CEPH distributed storage system is established to improve the I/O efficiency of online-learning system. Each VM consists of a kernel program and a 25 GB root filesystem with pre-installed Debian Linux system and the Moodle online-learning tool. An extra 200 GB root filesystem is mounted to the VM repair server to provide storage pool (treated as the orbit) for failed VMs. Therefore, there are at most 8 (treated as K) failed VMs can be sent to repair server to wait for repair procedure. To promise the quality of service, a failed working VM is replaced with a warm standby VM, and failed VMs are sent to the VM repair server, which could check and repair the malfunctioned root filesystem. When a failed VM arrives, it accepts repair procedure immediately according to its system log file if the VM repair server is available. Otherwise, the failed VM is transmitted to the storage pool of the repair server to wait for filesystem check and reconfiguration. When the number of failed VMs reaches 8, the Cloudstack has a 80% (treated as c) chance to prohibit any failed VMs from entering to repair server due to storage space limitation K. The Cloudstack may fail to prevent new failed VMs from entering the repair server due to storage space limitation K. The Cloudstack may fail to prevent new failed VMs from entering the repair server due to insufficient system resources. The system then moves to the unsafe failure (uf) state with probability 0.1 and cause an online-learning system error due to excessive use of storage space. A faster service rate is provided when the VM repair server is normal. The VM repair server may be subject to failure and provide a slower service rate due to a software or hardware problem, such as abnormal program execution or disk hot swap.

In the beginning, all the 12 (treated as M) working and 4 (treated as S) warm standby VMs are bootstrapped to speed up the deployment procedure. Let the failure time of the operating VMs be exponentially distributed with parameter λ = 0.12/hr, and the failure time of the warm standby VMs be exponentially distributed with parameters η = 0.04/hr. The failed VM can be repaired at an instant if the repair server is available. The service times are exponentially distributed with parameters μ1 = 2.0/hr and μ2 = 1.0/hr if the server is operating normally or is busy during a working breakdown, respectively. On the contrary, the failed VM is sent to the storage pool if the repair server is busy, and it repeats its demand for service with an exponential amount of retrial time with rate γ = 0.6/hr. No further arriving failed VMs are allowed to come in the system until the number of failed VMs in storage pool decreases to a value F = 3. The repair server must adopt an exponential startup time with parameter θ = 1.2/hr to restart, allowing failed VMs to come in the system. Let the failure time of the repair server be exponentially distributed with parameters α = 0.3/hr, and the service times follow an exponential distribution with parameter β = 0.6/hr.

3. Reliability and sensitivity analysis

In this RMRP, we describe the system states by the pair (i,j) where i denotes the system states and j represents the number of failed machines in the retrial orbit, j = 0,1,2,...,K − 1. Various values of i define system states as follows:

- i = 0: the server is operating normally but is idle, and arriving/retrial failed machines are allowed to join the system for service;
- i = 1: the server is operating normally but is busy, one failed machine is being served and arriving failed machines will be forced to enter the orbit;
- i = 2: the server is operating normally but is busy and the system is blocked, and arriving failed machines are not allowed to enter the system;
- i = 3: the server is busy during a working breakdown, and failed machines are allowed to join the system;
- i = 4: the server is busy during a working breakdown, and failed machines are not allowed to join the system;
- i = 5: the server is idle during a working breakdown, and arriving/retrial failed machines are allowed to join the system.

With the above definition of system states, the state space can be presented as

Ω = \{(i,j) : i = 0,1,...,5; j = 0,1,...,K − 2 \} ∪ \{(2,K − 1),(4,K − 1)\}.

At time t, the probabilities of different states in the system are defined as follows:

\[ P_{ij}(t) \equiv \text{probability of having } j \text{ failed machines in the retrial orbit when the server is in system state } i, \text{ where } (i,j) \in \Omega; \]

\[ P_{ii}(t) \equiv \text{probability of system fails in system state } i = 1, 2; \]

![Fig. 1. The online-learning cloud system.](image-url)
The state transition-rate diagram for the RMRP with working breakdowns operating under the F-policy can be shown in Fig. 2. The differential-difference equations for each state at time \( t \) can be derived by Fig. 2. For instance, the differential difference equation for the state \((0, 0)\) is

\[
\frac{d}{dt}P_{00}(t) = -\lambda_0 P_{00}(t) + \mu_1 P_{10}(t) + \mu_1 P_{20}(t).
\]

Implementing the Laplace transform and the technique of integration by parts in calculus as follows

\[
\int_0^\infty e^{-st}P_{ij}(t)dt = \left[ e^{-st}P_{ij}(t) \right]_0^\infty - \int_0^\infty P_{ij}(t)de^{-st}
\]

\[
s \int_0^\infty P_{ij}(t)e^{-st}dt = sP_{ij}(s) - P_{ij}(0).
\]

We have

\[
sP_{00}(s) - P_{00}(0) = -\lambda_0 P_{00}(s) + \mu_1 P_{10}(s) + \mu_1 P_{20}(s).
\]

Similarly, by taking the Laplace transform to the differential equation of each system state, the Laplace transform Eqs. (1)-(21) are established.

\[
(\lambda_0 + s)p_{00}^*(s) - \mu_1 p_{10}^*(s) - \mu_1 p_{20}^*(s) = p_{00}(0),
\]

\[
(\lambda_1 + s)p_{10}^*(s) - \mu_1 p_{10}^*(s) = p_{00}(0), \quad 1 \leq j < K - 2,
\]

\[
(\lambda_1 + s)p_{10}^*(s) - \mu_1 p_{10}^*(s) = p_{00}(0), \quad 1 \leq j < F - 1,
\]

\[
(\lambda_1 + s)p_{10}^*(s) - \mu_1 p_{10}^*(s) = p_{00}(0), \quad F \leq j < K - 3.
\]
(12)

\[(\lambda_{i+1} + \beta + \mu_2 + s)\tilde{p}_{i+1}(s) - \alpha \tilde{p}_{p_i+1}(s) - \lambda_k \tilde{p}_{K-1}(s) = 0, \quad F \in \mathbb{S}K - 3, \]

\[(\lambda_{k-1} + \beta + \mu_2 + s)\tilde{p}_{k-1}(s) - \alpha \tilde{p}_{k-1}(s) - \lambda_k \tilde{p}_{K-1}(s) - \lambda_{k-1} \tilde{p}_{K-2}(s) = 0, \quad F \in \mathbb{S}K = 3, \]

\[(\beta + \theta + s)\tilde{p}_{i+1}(s) - \alpha \tilde{p}_{p_i+1}(s) - \mu_2 \tilde{p}_{i+1}(s) = 0, \quad F \in \mathbb{S}F - 1, \]

\[(\beta + \mu_2 + s)\tilde{p}_{i+1}(s) - \alpha \tilde{p}_{p_i+1}(s) - \mu_2 \tilde{p}_{i+1}(s) = 0, \quad F \in \mathbb{S}K - 2, \]

\[(\beta + \mu_2 + s)\tilde{p}_{i+1}(s) - \alpha \tilde{p}_{p_i+1}(s) - \mu_2 \tilde{p}_{i+1}(s) = 0, \quad F \in \mathbb{S}K - 1, \]

\[(\alpha_0 + s)\tilde{p}_{i+1}(s) - \alpha \tilde{p}_{p_i+1}(s) = 0, \quad \gamma_0 \in \mathbb{S}F - 0, \]

\[(\lambda_0 + s)\tilde{p}_{i+1}(s) - \alpha \tilde{p}_{p_i+1}(s) = 0, \quad \gamma_0 \in \mathbb{S}K - 0, \]

\[s \tilde{p}_{i+1}(s) - (1 - \alpha) \tilde{p}_{i+1}(s) = 0, \quad \gamma_0 \in \mathbb{S}F - 0, \]

\[s \tilde{p}_{i+1}(s) - (1 - \alpha) \tilde{p}_{i+1}(s) = 0, \quad \gamma_0 \in \mathbb{S}K - 0, \]

Let \(P' \tilde{s}\) represent the corresponding vector of Laplace transform. By partitioning the vector \(P' \tilde{s}\) as

\[P' \tilde{s} = \begin{bmatrix} \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s) \end{bmatrix}^T, \]

where the column vectors \(P_{i+1}^{1,6}(s), \ldots, P_{i+1}^{1,5} \) are given by

\[P_{i+1}^{1,6}(s) = \begin{bmatrix} \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s) \end{bmatrix}^T, \]

\[P_{i+1}^{1,5}(s) = \begin{bmatrix} \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s), & \tilde{p}_{i+1}(s) \end{bmatrix}^T, \]

Eqs. (1)-(19) can be represented in matrix form as follows:

\[\mathbf{R}(s)P' \tilde{s} = (Q + d_{i+1})P' \tilde{s} = P(0), \quad \mathbf{I}_{K-2} \]

\[Q = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & 0 & 0 & 0 & 0 & 0 \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & 0 & 0 & 0 & 0 & 0 \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & 0 & 0 & 0 & 0 & 0 \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} & 0 & 0 & 0 & 0 \\ \alpha_{51} & \alpha_{52} & \alpha_{53} & \alpha_{54} & \alpha_{55} & 0 & 0 & 0 \\ \alpha_{61} & \alpha_{62} & \alpha_{63} & \alpha_{64} & \alpha_{65} & \alpha_{66} & 0 & 0 \\ \alpha_{71} & \alpha_{72} & \alpha_{73} & \alpha_{74} & \alpha_{75} & \alpha_{76} & \alpha_{77} & 0 \\ \alpha_{81} & \alpha_{82} & \alpha_{83} & \alpha_{84} & \alpha_{85} & \alpha_{86} & \alpha_{87} & \alpha_{88} \end{bmatrix} \quad (\mathbf{Q}) \]

Notation \(O\) denotes the zero matrix with appropriate dimensions. Let \(A_{i,j}\) denotes the element of matrix \(A\) in the \(i^{th}\) row and \(j^{th}\) column. The sub-matrix \(Q_{11}\) and \(Q_{12}\) are diagonal matrices of order \(K - 1\) defined by

\[Q_{11} = \begin{cases} \lambda_{i-1}, & i = 1; \\
\lambda_{i-1} + \gamma_{i-1}, & 2 \leq i \leq K - 1, \end{cases} \]

and \(Q_{12} = -\mu_1, \quad 1 \leq i \leq K - 1\). The sub-matrix \(Q_{13}\) is a \((K - 1) \times K\) matrix with only one non-zero element \(Q_{13}(1,1) = -\mu_1\). The sub-matrices \(Q_{23}\) and \(Q_{22}\) are \((K - 1) \times (K - 1)\) matrix defined by

\[Q_{23} = \begin{cases} -\lambda_{i-1}, & 1 \leq i \leq K - 1, \quad i = 1; \\
-\gamma_{i-1}, & 1 \leq i \leq K - 2, \quad j = i + 1; \\
0, & \text{other } i, j. \end{cases} \]

and \(Q_{22} = -\mu_1, \quad 1 \leq i \leq K - 1\). The sub-matrix \(Q_{23}\) is a \((K - 1) \times (K - 1)\) matrix defined by

\[Q_{23} = \begin{cases} -\lambda_{i-1}, & 2 \leq i \leq K - 1, \quad i = 1; \\
-\gamma_{i-1}, & 2 \leq i \leq K - 2, \quad j = i + 1; \\
0, & \text{other } i, j. \end{cases} \]

The sub-matrix \(Q_{33}\) is \((K - 1) \times K\) matrix defined by

\[Q_{33} = \begin{cases} \alpha + \mu_1 + \theta, & 1 \leq i \leq F, \quad j = i; \\
\alpha + \mu_1, & 1 \leq i \leq K - 1, \quad j = i + 1; \\
0, & \text{other } i, j. \end{cases} \]

and \(Q_{35} = -\beta I_{K-1}, \quad Q_{44} = -d_{i+1}\). The sub-matrix \(Q_{34}\) is \(K \times (K - 1)\) matrix defined by

\[Q_{34} = \begin{cases} \lambda_0 + \beta + \mu_1, & 1 \leq i \leq K - 1, \quad j = i; \\
-\lambda_{i-1}, & 1 \leq i \leq K - 1, \quad j = i + 1; \\
0, & \text{other } i, j. \end{cases} \]

The sub-matrix \(Q_{43}\) and \(Q_{46}\) are \((K - 1) \times K\) and \((K - 1) \times (K - 1)\) matrices defined by

\[Q_{43} = \begin{cases} -\lambda_{i-1}, & 1 \leq i \leq K - 1; \\
0, & \text{other } i, j. \end{cases} \]

The sub-matrix \(Q_{44}\) and \(Q_{46}\) are \((K - 1) \times K\) and \((K - 1) \times (K - 1)\) matrices defined by

\[Q_{44} = \begin{cases} -\beta + \theta, & i = 1, j = 1; \\
0, & \text{other } i, j. \end{cases} \]

and \(Q_{46} = -\mu_1 I_{K-1}, \quad Q_{26}\) is a \((K - 1) \times (K - 1)\) diagonal matrix of order \(K - 1\) defined by

\[Q_{26} = \begin{cases} \lambda_0, & i = 1; \\
\lambda_{i-1} + \gamma_{i-1}, & 2 \leq i \leq K - 1. \end{cases} \]
6

\[ \text{Q}_{72} \text{ and } \text{Q}_{84} \text{ are } 2 \times (K - 1) \text{ matrices defined by} \]

\[
\begin{align*}
\text{Q}_{72}(i,j) &= \left\{ \begin{array}{cl}
-1 (1 - c) \mathbf{k}_{i-j}, & i = 1, j = K - 1; \\
0, & \text{otherwise}.
\end{array} \right.
\end{align*}
\]

and

\[
\begin{align*}
\text{Q}_{84}(i,j) &= \left\{ \begin{array}{cl}
-1 (1 - c) \mathbf{k}_{i-j}, & i = 2, j = K - 1; \\
0, & \text{otherwise}.
\end{array} \right.
\end{align*}
\]

To solve the Equation (22), let \( \mathbf{N}(s) \) be the matrix by replacing the \( i \)th column of \( \mathbf{R}(s) \) with \( \mathbf{P}(0) \) and then the probability \( P_{ij}(s) \) can be solved by Cramer’s rule as

\[
P_{ij}(s) = \frac{\det[\mathbf{N}(s)]}{\det[\mathbf{R}(s)]}.
\]

Notations \( \det[\cdot] \) denotes the determinant of the associated matrix. Maple software can be used to obtain \( P_{ij}(s) \). Since the Laplace transform is a one-to-one mapping, once the transformed probability \( P_{ij}(s) \) is obtained, the corresponding steady-state probability \( P_{ij}(t) \) can be calculated by the inverse transformation. Particularly, in the next section, it is shown that the mean time to system failure can be calculated by \( \lim_{s \to \infty} P_{ij}(s) \) and \( \lim_{s \to \infty} P_{ij}(s) \).

### 3.2. Reliability function and mean time to system failure

Let \( Y \) be the random variable denoting the first time of system failure with probability density function \( f(y) \) and cumulative distribution function \( F(y) \). Since \( P_{00}(t) \) and \( P_{00}(t) \) are the probability of the system has failed at or before time \( t \), the reliability function is obtained in the following:

\[
R_Y(t) = P(Y \geq t) = 1 - F(t) = 1 - P_{00}(t) - P_{00}(t), \quad t \geq 0.
\]

Then, the Laplace transform of \( R_Y(t) \) is defined as

\[
R_Y(s) = \int_0^\infty e^{-st}R_Y(t)dt = \int_0^\infty e^{-st}[1 - P_{00}(t) - P_{00}(t)]dt = \frac{1}{s} - P_{00}(s) - P_{00}(s).
\]

It is always finite since the integrated function is a monotonic decreasing and bounded function of \( t \). Thus, it implies that the mean time to system failure (MTTF) exists and can be calculated by

\[
\begin{align*}
\int_0^\infty R_Y(t)dt &= \int_0^\infty [1 - F(t)]dt = \int_0^\infty P(Y \geq t)dt \\
&= \int_0^\infty \int_0^\infty f(y)dydt = \int_0^\infty \int_0^\infty f(y)dydy \\
&= \int_0^\infty yf(y)dy = E(Y) = \text{MTTF}.
\end{align*}
\]

Consequently, the MTTF can be obtained as

\[
\text{MTTF} = \int_0^\infty R_Y(t)dt = \lim_{s \to \infty} R_Y(s) = \lim_{s \to \infty} \frac{1}{s} - P_{00}(s) - P_{00}(s).
\]

### 3.3. Sensitivity and relative sensitivity analysis

For each system parameter \( \omega \), differentiating (25) with respect to \( \omega \), we obtain the gradient of system reliability for the sensitivity analysis, shown below

\[
\frac{\partial R_Y(t)}{\partial \omega} = \left[ \frac{\partial P_{00}(s)}{\partial \omega} + \frac{\partial P_{00}(s)}{\partial \omega} \right].
\]

Similarly, the gradient of MTTF can be derived by differentiating (27) with respect to \( \omega \), it yields

\[
\frac{\partial \text{MTTF}}{\partial \omega} = \lim_{s \to \infty} \frac{\partial R_Y(t)}{\partial \omega} = \lim_{s \to \infty} \frac{\partial R_Y(t)}{\partial \omega} R_Y(t).
\]

Moreover, the relative gradient on \( R_Y(t) \) and MTTF can be implemented by

\[
\sigma = \frac{\partial R_Y(t)}{\partial \omega} R_Y(t) = \frac{\partial \text{MTTF}}{\partial \omega} \frac{\omega}{\text{MTTF}}.
\]

and

\[
\xi = \frac{\partial \text{MTTF}}{\partial \omega} = \frac{\omega}{\text{MTTF}}.
\]

Since the derived results \( R_Y(t) \) and MTTF are complex non-linear function of the system parameters, the explicit closed-form of the associated derivative with respective to parameter \( \omega \) is difficult to be obtained. Hence, we calculate the gradient numerically. Taking \( \omega = \alpha \) as an example, \( \partial P_{00}(\alpha) / \partial \delta \) at point \( \alpha = \alpha_0 \) in Equation (28) is computed by

\[
\frac{\partial P_{00}(\alpha)}{\partial \delta} \bigg|_{\alpha=\alpha_0} = \frac{P_{00}(\alpha)}{\partial \delta} \bigg|_{\alpha=\alpha_0} - \frac{P_{00}(\alpha)}{\partial \delta} \bigg|_{\alpha=\alpha_0},
\]

with a small difference magnitude \( \delta = 10^{-8} \).
4. Numerical results for related analysis

This section presents some numerical results to perform several related analyses. In Sections 4.1 and 4.2, we discuss how system reliability \( R_Y(t) \) and MTTF are affected by changing the values of various system parameters. Moreover, in Section 4.3, numerical experiments for sensitivity and relative sensitivity analysis are provided.

4.1. System reliability analysis

First, as the illustration example of online-learning cloud system provided in Section 2.1, the base (initial) parameters are selected as \( M = 12, S = 4, K = 8, F = 3, \lambda = 0.12, \eta = 0.04, \mu_1 = 2.0, \mu_2 = 1.0, \alpha = 0.3, \beta = 0.6, \gamma = 0.6, \theta = 1.2, c = 0.8 \). Based on the above basic setting, except for parameters \( M \) and \( S \) (which are always fixed), we change one parameter at a time to observe its impact on system reliability and MTTF.
specific parameter at one time and keep the values of the other parameters. That is, we investigate the effects of each parameter on the system reliability to recognize which parameters are critical and needed to be controlled well. The numerical results of $R_Y(t)$ for system reliability assessment are depicted in Figs. 3–13. From these figures, we observe that $\lambda$, $c$, $K$ significantly affect system reliability (see Figs. 3, 11, 13). Parameters $\gamma$ and $\mu_1$ affect the system reliability moderately (see Figs. 5 and 6). Parameters $\mu_2$, $\alpha$, $\beta$, $F$ affects the system reliability slightly (see Figs. 7, 8, 9, 12) while $\eta$ and $\theta$ affect the system reliability rarely (see Figs. 4 and 10). It shows the managers several possibilities for reliability enhance: (i) enlarge the system capacity; (ii) reduce the mean failure rate or improve the VM stability; (iii) add some necessary mechanisms for successfully blocking the failed VMs when the system is full.

4.2. MTTF analysis

In this subsection, some numerical experiments are performed to investigate how each system parameter affects the mean time to system failure (MTTF). Because of the significant effects of parameter $c$ on the
Computers & Industrial Engineering 150 (2020) 106885

Fig. 12. System reliability affected by $F$.

Fig. 13. System reliability affected by $K$.

Table 1
MTTF affected by $c$ and $\lambda$.

| $c$ | 0.08 | 0.09 | 0.1 | 0.11 | 0.12 |
|-----|------|------|-----|------|------|
| 0.5 | 74.960 | 57.930 | 47.261 | 40.176 | 35.214 |
| 0.6 | 89.396 | 68.770 | 55.776 | 47.146 | 41.116 |
| 0.7 | 112.960 | 86.662 | 69.898 | 58.731 | 50.936 |
| 0.8 | 158.304 | 121.812 | 97.894 | 81.792 | 70.523 |
| 0.9 | 281.595 | 222.462 | 179.967 | 150.151 | 128.899 |

Table 2
MTTF affected by $c$ and $\eta$.

| $c$ | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 |
|-----|------|------|------|------|------|
| 0.5 | 49.277 | 48.235 | 47.261 | 46.349 | 45.492 |
| 0.6 | 58.178 | 56.937 | 55.776 | 54.688 | 53.666 |
| 0.7 | 72.924 | 71.361 | 69.898 | 68.525 | 67.235 |
| 0.8 | 102.097 | 99.930 | 97.894 | 95.979 | 94.176 |
| 0.9 | 187.165 | 183.471 | 179.967 | 176.644 | 173.489 |

Table 3
MTTF affected by $c$ and $\gamma$.

| $c$ | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|-----|-----|-----|-----|-----|-----|
| 0.5 | 42.386 | 41.921 | 41.261 | 40.826 | 40.349 |
| 0.6 | 49.262 | 48.697 | 48.146 | 47.688 | 47.235 |
| 0.7 | 60.575 | 59.906 | 59.348 | 58.894 | 58.437 |
| 0.8 | 82.668 | 81.993 | 81.348 | 80.894 | 80.437 |
| 0.9 | 145.002 | 143.402 | 141.894 | 140.437 | 139.002 |

Table 4
MTTF affected by $c$ and $\mu_1$.

| $c$ | 1.8 | 2.0 | 2.2 | 2.4 | 2.6 |
|-----|-----|-----|-----|-----|-----|
| 0.5 | 41.432 | 44.167 | 47.261 | 50.719 | 54.541 |
| 0.6 | 48.495 | 51.912 | 55.776 | 60.090 | 64.858 |
| 0.7 | 60.205 | 64.756 | 69.898 | 75.636 | 81.975 |
| 0.8 | 83.395 | 90.207 | 97.894 | 106.465 | 115.929 |
| 0.9 | 151.208 | 164.740 | 179.967 | 196.918 | 215.612 |

Table 5
MTTF affected by $c$ and $\mu_2$.

| $c$ | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|-----|-----|-----|-----|-----|-----|
| 0.5 | 45.907 | 46.553 | 47.261 | 48.027 | 48.847 |
| 0.6 | 54.063 | 54.883 | 55.776 | 56.738 | 57.765 |
| 0.7 | 67.582 | 68.694 | 69.898 | 71.189 | 72.564 |
| 0.8 | 94.341 | 96.055 | 97.894 | 99.854 | 101.931 |
| 0.9 | 172.485 | 176.127 | 179.967 | 184.009 | 188.255 |

Table 6
MTTF affected by $c$ and $\alpha$.

| $c$ | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|-----|-----|-----|-----|-----|-----|
| 0.5 | 53.892 | 50.034 | 47.261 | 45.169 | 43.531 |
| 0.6 | 63.830 | 59.137 | 55.776 | 53.247 | 51.195 |
| 0.7 | 75.205 | 64.756 | 60.348 | 56.894 | 53.347 |
| 0.8 | 112.207 | 104.048 | 97.894 | 93.283 | 89.701 |
| 0.9 | 206.378 | 191.012 | 179.967 | 171.657 | 165.187 |

Table 7
MTTF affected by $c$ and $\beta$.

| $c$ | 0.5 | 0.6 | 0.7 | 0.8 |
|-----|-----|-----|-----|-----|
| 0.5 | 45.548 | 46.490 | 47.261 | 47.903 |
| 0.6 | 53.384 | 54.698 | 55.776 | 56.674 |
| 0.7 | 66.341 | 68.294 | 69.898 | 71.237 |
| 0.8 | 91.886 | 95.185 | 97.894 | 102.070 |
| 0.9 | 165.742 | 173.556 | 179.968 | 185.312 |

\[ \text{T.-C. Yen et al.} \]
In sensitivity analysis, the order of the parameters that affect system reliability $R(t)$ is determined by their gradient. Table 10 provides a relative sensitivity analysis on $MTTF$.

**Table 10**

| $\lambda$ | $\xi_1$ | $\eta$ | $\xi_2$ | $\gamma$ | $\xi_3$ |
|-----------|---------|--------|---------|---------|--------|
| 0.08      | -2.21401| 0.02   | -0.04365| 0.3     | 0.23133|
| 0.09      | -2.11627| 0.03   | -0.06286| 0.4     | 0.24180|
| 0.10      | -1.94109| 0.04   | -0.08039| 0.5     | 0.24869|
| 0.11      | -1.75881| 0.05   | -0.09644| 0.6     | 0.25223|
| 0.12      | -1.59328| 0.06   | -0.11103| 0.7     | 0.25468|

Extensive numerical results are carried out on sensitivity and relative sensitivity analysis. First, sensitivity analysis on system reliability is provided and depicted in Fig. 14. One can see from Fig. 14 that the order of sensitivity analysis on system reliability affected by each parameter can be ranked below: $\lambda > \xi > \eta > \alpha > \mu_1 \approx \gamma > \beta > \mu_2 > \theta$. Furthermore, the signs of gradient on system reliability with respect to parameters $\lambda, \eta, \alpha, \theta$ are negative. It indicates that system reliability decreases as these parameters increase. In addition, the results show that parameters $\lambda, \eta, \alpha, \theta$ have significant effects and therefore they should be carefully estimated as accessing the system reliability. Therefore, a monitoring system with continuous data collection may be crucial to control them well.

Next, as for sensitivity analysis on $MTTF$ affected by each system parameter, Table 9 tabulates the gradient of $MTTF$ with respect to various parameters. We observe from Table 9, the order of sensitivity
analysis on MTTF affected by each parameter can be ranked below: \( \gamma > \mu_1 > \lambda > \alpha > \beta > \gamma > \theta > \mu_2 > \eta \). Besides, the signs of sensitivity on MTTF affected by parameters \( \gamma, \mu_1, \mu_2, \beta \), and \( \gamma \) are positive which indicates that the increment of the parameter improves sensitivity on MTTF.

Finally, the related sensitivity analyses on system reliability and MTTF are also performed. Extensive numerical results for relative sensitivity on system reliability and MTTF are shown in Fig. 15 and Table 10, respectively. It appears from Fig. 15 that the order of relative sensitivity on system reliability affected by each parameter can be ranked below: \( \gamma > \mu_1 > \mu_2 > \lambda > \beta > \alpha > \eta > \theta \). From Table 10, one can see that the order of relative sensitivity on MTTF affected by each parameter can be ranked below: \( \gamma > \mu_1 > \eta > \alpha > \mu_2 > \lambda > \beta > \gamma > \theta \).

From Figs. 10, 14 and 15, and Tables 8 and 10, we demonstrate that the impact of the mean startup rate \( \theta \) may be considered to be ignored for this system.

5. Conclusions

In this paper, we introduced a retrial machine repair problem with working breakdowns operating under the probabilistic F-policy. We applied the Laplace transform technique and the matrix-analytic method to solve the differential-difference equations of the investigated system. Then, the explicit expressions of the system reliability and the mean time to failure were derived. A practical illustration of the online-learning cloud system that the proposed model can be applied to is provided. Based on the application case, a reliability-based sensitivity analysis for the effects of system parameters was performed. The associated numerical results were tabulated and displayed graphically to evaluate the impacts of different parameters. The results indicated that the effect of the mean startup rate is ignorable. Increasing the prohibit success probability, the mean service rate, or the system capacity can effectively enhance the system reliability. The results also show that the mean arrival rate and the prohibit probability need to be evaluated carefully. Once they are changed but the manager does not notice it, the system reliability may become low and cause additional operating cost. In the future, we provide some possible research topics shown below: 1. some random times (e.g. repair times or retrial times) obey general distribution, 2. system availability of the model, 3. server may be subject to unpredictable breakdowns, and 4. the optimal management strategy.

Acknowledgements

The authors would like to thank the editor and anonymous reviewers for their valuable comments, which have led to a significant improvement in the manuscript. This work was partially supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 106-2221-E-126-003-MY2 and MOST 107-2218-E-009-058-MY2.

References

Artalejo, J. R. (1999b). A classified bibliography of research on retrial queues: Progress in 1990-1999. Top, 7, 187-211.

Artalejo, J. R., & Falin, G. (2002). Standard and retrial queueing systems: A comparative analysis (pp. 101–129). XV: Revista Matematica Complutense.

Artalejo, J. R., & Gomez-Corral, A. (2008). Retrial Queueing Systems: A Computational Approach. Springer-Verlag.

Chang, P.-M., Lee, Y.-T., Chang, C.-J., & Yeh, C. (2018). Analysis of a standby redundant system with controlled arrival of failed machines. International Journal of Industrial and Systems Engineering, 28, 117–134.

Chen, W.-L., & Wang, K.-H. (2018). Reliability analysis of a retrial machine repair problem with warm standby and a single server with N-policy. Reliability Engineering and System Safety, 180, 476–486.

Chen, W.-L. (2018). System reliability analysis of retrial machine repair systems with warm standbys and a single server of working breakdown and recovery policy. Systems Engineering, 21, 59–65.

Efrosinin, D. V., & Winkler, A. (2011). Queueing system with a constant retrial rate, non-reliable server and threshold-based recovery. European Journal of Operational Research, 210, 594–605.

Gharbi, N., & Jousselain, M. (2006). GSPN analysis of retrial systems with servers breakdowns and repairs. Applied Mathematics and Computation, 174, 1151–1168.

Jain, M., & Sanga, S. S. (2017). Control F-policy for fault tolerance machining system with general retrial attempts. National Academy Science Letters, 40, 359–364.

Jain, M., & Sanga, S. S. (2019a). F-policy for M/M/1/K retrial queueing model with state-dependent rates. In Performance Prediction and Analytics of Fuzzy, Reliability and Queueing Models (pp. 127–138). Singapore: Springer.

Jain, M., & Sanga, S. S. (2019b). Optimal control F-policy for M/M/R/K queue with an additional server and balking. International Journal of Applied and Computational Mathematics, 5, 163.

Jain, M., & Sanga, S. S. (2019c). Admission control for finite capacity queueing model with mixed standby components. Applied Mathematical Modelling, 38, 5946–5962.

Jain, M., & Sanga, S. S. (2020a). State dependent queueing models under admission control F-policy: A survey. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-019-01638-y

Jain, M., & Sanga, S. S. (2020b). Fuzzy cost optimization and admission control for machine interference problem with general retrial. Journal of Testing and Evaluation, 48(6). https://doi.org/10.1520/JTE20180882

Jain, M., Sanga, S. S., & Meena, R. K. (2012). Control F-policy for Markovian retrial queue with server breakdowns. In 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy System, New Delhi, India (pp. 1–5).

Jain, M., Shekhar, C., & Shukla, S. (2016). Queueing analysis of machine repair problem with controlled rates and working vacation under F-policy. In Proceedings of the National Academy of Sciences, India Section A: Physical Sciences (Vol. 86, pp. 21–31).

Ke, J.-C., Yang, D.-Y., Sheu, S.-H., & Kuo, C.-C. (2013). Availability of a repairable retrial system with warm standby components. International Journal of Computer Mathematics, 90, 2279–2297.

Kuo, C.-C., Sheu, S.-H., Ke, J.-C., & Zhang, G. Z. (2014). Reliability-based measures for a retrial system with mixed standby components. Applied Mathematical Modelling, 38, 4640–4651.

Phung-Duc, T. (2019). Retrial queueing models: A survey on theory and applications, arXiv preprint arXiv:1906.09560.

Shukla, S. (2016). Queueing analysis of machine repair problem with general retrial times and state dependent rates. Journal of Industrial and Management Optimization, 12. https://doi.org/10.9394/jimo.2019073

Jain, M., Sanga, S. S., & Meena, R. K. (2012). Control F-policy for Markovian retrial queue with server breakdowns. In 1st IEEE International Conference on Power Electronics, Intelligent Control and Energy System, New Delhi, India (pp. 1–5).