Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 174 (2006) 182–200
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Stochastics and Statistics

On the interaction between maintenance, spare part inventories and repair capacity for a k-out-of-N system with wear-out

Karin S. de Smidt-Destombes [a,*], Matthieu C. van der Heijden [b], Aart van Harten [b]

[a] *TNO Physics and Electronics Laboratory, P.O. Box 96864, Oude Waalsdorperweg 63, 2509 JG The Hague, The Netherlands*
[b] *University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands*

Received 25 March 2004; accepted 17 February 2005
Available online 14 April 2005

Abstract

In this paper we consider a k-out-of-N system with identical, repairable components under a condition-based maintenance policy. Maintenance consists of replacing all failed and/or aged components. Next, the replaced components have to be repaired. The system availability can be controlled by the maintenance policy, the spare part inventory level, the repair capacity and repair job priority setting. We present two approximate methods to analyse the relation between these control variables and the system availability. Comparison with simulation results shows that we can generate accurate approximations using one of these models, depending on the system size.

Keywords: Maintenance; Spare parts; Repair capacity; Priority setting; k-out-of-N systems; Ageing

1. Introduction

Many of today's technological systems, such as aircrafts, military equipments or medical equipments, are becoming more and more complex. At the same time, the requirements concerning availability and reliability are becoming higher. There are several ways to influence the system availability. First, we can prevent system failures using redundancy of critical components and preventive maintenance. Second, we can reduce the length of system downtimes caused by corrective and/or preventive maintenance. The latter can be achieved

* Corresponding author. Tel.: +31 70 374 01 13; fax: +31 70 734 06 42.
E-mail address: desmidt@fel.tno.nl (K.S. de Smidt-Destombes).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2005.02.043

by replacing failed components immediately and repairing off-line. To this end, we need a certain amount of spare part inventories that is balanced with the demand for spare parts and the repair throughput time of failed components. If we have sufficient spare parts, the system downtime is limited to the replacement time of failed components. If not, the downtime is extended with the time needed to repair additional components. In this sense, we have a trade-off between spare part inventory levels and repair shop throughput times. The repair shop throughput time itself can be influenced by the repair capacity and possibly by repair shop priority setting if multiple types of repair jobs share the same repair capacity.

The simultaneous setting of all structural parameters (redundancy, repair shop capacity) and control variables (spare part inventory levels, preventive maintenance policy and repair job priorities) is mathematically a hard problem. A prerequisite for optimisation is that we are able to evaluate the system availability as a function of these factors efficiently. To this end, we present in this paper an approximate method to analyse these relations for the special case of a single parallel redundant (k-out-of-N) system with component wear-out under a condition-based maintenance policy. That is, we initiate maintenance if the number of failed components passes some critical level m. Such a policy can also be seen as an m-failure group replacement policy, see Wang (2002).

We model the wear-out as follows: The time to component failure consists of a first phase where the component moves from the as-good-as-new state to the degraded state and a second phase where the component moves from the degraded state to the failed state. We assume that the sojourn time in each state is exponentially distributed. We consider two types of repair jobs, namely, repair of degraded and failed components, both having an exponential repair time distribution. This wear-out model is inspired by real systems found in practice, like the Active Phased Array Radar (APAR) used by the Royal Netherlands Navy, see Fig. 1 and De Smidt-Destombes et al. (2004). This system consists of many transmit and receive elements. An element is said to have degraded if it is only capable of transmitting signals or receiving signals, but not both. If both functions fail, the element has completely failed. Another interesting example is the ATAS, a towed array sonar consisting of several hydrophones towed to the end of a ship in order to find submarines. Other examples include batteries for pumps, servers and other industrial equipments.

A single repair facility with one or more parallel, identical servers handles all jobs. A certain lead-time may be necessary between maintenance initiation and execution for set-up activities and/or moving the system to the repair facility. Maintenance consists of replacing failed and/or degraded components by new ones (spares). If insufficient spares are available, maintenance is delayed until the required components have been repaired.

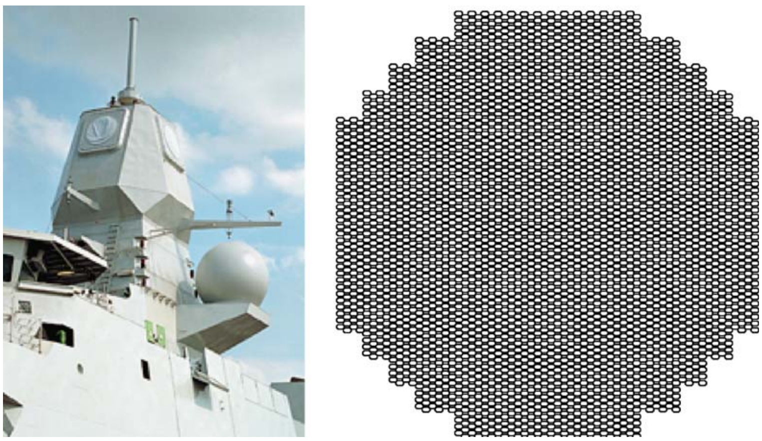

Fig. 1. The Active Phased Array Radar (APAR, left) consists of four "faces", each having a large number of elements (right); a face can be modelled as a k-out-of-N system.

The literature on spare part management is extensive, see Rustenburg (2000) and Kennedy et al. (2002) for a review. Parts of our problem have been discussed as well. Firstly, there is a range of papers on inventory analysis of repairable spare parts under a finite repair capacity (Díaz and Fu, 1997; Sleptchenko et al., 2003; Zijm and Avsar, 2003; Avsar and Zijm, 2003). For a recent overview of these models integrating spare part management and repair capacity we refer to Sleptchenko et al. (2002). Secondly, the literature on the relation between spare part inventories and maintenance mainly focuses on age based maintenance rather than condition-based maintenance. Some papers (Kabir and Al-Olayan, 1996; Park and Park, 1986) consider consumable spares rather than repairables, in which case repair capacity is not relevant. Sarkar and Sarkar (2001) consider a problem related to ours, namely a one-component model with maintenance based on periodic inspections where the function of the component, degraded or failed, is taken over by a spare one. In our paper, however, we use a maintenance policy based on redundancy and also we include the effect of limited repair capacity. Thirdly, the importance of integrating the maintenance policy with spare parts and repair capacity is mentioned by several authors, see e.g. Dinesh Kumar et al. (2000) and Gross et al. (1985). However, only a few papers actually deal with the integration of these three aspects in quantitative models. De Smidt-Destombes et al. (2004) present a similar model with two-state components (up or down only).

Compared to the latter model, the introduction of wear-out complicates the analysis considerably. We have different repair jobs (failed and degraded components) and so we may consider repair priorities to reduce system downtime. Also, the computation of the system uptime and particularly the system downtime is more complex. The introduction of a degraded state also allows for a wider class of maintenance policies if the component states are observable during system uptime, as is true for the APAR example. We may then use a maintenance policy based on the number of failed and degraded components. For example, if the number of failed components is not too high but many components are degraded, it may be wise to initiate maintenance to avoid system failure during the lead-time.

The remainder of this paper is structured as follows. In Section 2, we present our model and notation in detail and we discuss our basic assumptions. Next, we present two approaches to evaluate the system availability approximately, an exact analysis (for major parts of the analysis) in Section 3 and a simple and fast approximate analysis in Section 4. In Section 5, we show some numerical examples and compare them with simulations to examine the accuracy of our approximations and to get some insight in the relations between the parameter settings and the system availability. We discuss several extensions of our basic model in Section 6. We end with conclusions and possibilities for further research in Section 7.

2. Model description

We consider a single k-out-of-N system with deteriorating components and hot standby redundancy, which means that all components are active and have the same failure behaviour. Each component has three possible observable states, namely as-good-as-new (state 0), degraded (state 1) and failed (state 2). When in use, the sojourn time of a component in state $i-1$ is exponentially distributed with mean $1/\lambda_i$ ($i = 1, 2$). The system is fully operational if less than $N-k+1$ components have failed. To prevent system down time, maintenance is performed dependent on the condition of the system. In the simplest case, the maintenance policy consists of a single critical number m, being the number of failed components at which maintenance is initiated. We have a deterministic lead-time or set-up time L between maintenance initiation and execution. Therefore, it may be wise to initiate maintenance before the actual system failure ($m < N-k+1$).

Maintenance consists of replacing all failed and possibly also all degraded components by spares (repair-by-replacement). We denote the number of spares by S. We assume that the replacement time is negligible.

Nomenclature

c	repair capacity
k	the least number of components needed for a functional system
L	lead-time: time from maintenance initiation until the start of maintenance activities
m	the number of failed components to initiate maintenance activities
N	the total number of components in the system
S	the total number of spares
λ_i	the transition rate of a system component from state $i-1$ to state i
μ_i	the repair rate of a component from state i to state 0
$T(i,j)$	time from system state $(N-i-j,i,j)$ until maintenance initiation
$\alpha(i,j)$	probability of system transition from state $(N-i-j,i,j)$ to $(N-i-j-1,i+1,j)$
$\beta(i,j)$	probability of system transition from state $(N-i-j,i,j)$ to $(N-i-j,i-1,j+1)$
$\tau(i,j)$	sojourn time of the system in state $(N-i-j,i,j)$
$Q(i,j,t)$	probability of the system reaching state $(N-i-j,i,j)$ at time t given m failed components at time 0
$p_{ij}(t)$	probability of a component transition from state i to state j during time t
$P(i,m)$	probability of the system being in state $(N-i-m,i,m)$ at maintenance initiation
$P_L(i,j)$	probability of the system being in state $(N-i-j,i,j)$ at the start of maintenance
$\pi(i,j)$	probability of the spares being in state $(i,S-i-j,j)$ at the start of maintenance
$R(r,s_1,s_2)$	time to repair r components from spares state $(S-s_1-s_2,s_1,s_2)$ given capacity c
$H(w,x,y,z,t)$	probability that spares state changes from $(S-w-x,w,x)$ to $(S-y-z,y,z)$ in time t, given repair capacity c
$\widehat{T}$	time from maintenance initiation until system failure, given maint. initiation level m
$\widehat{T}(i,m)$	time from system state $(N-i-m,i,m)$ to failure, given maint. initiation level m
A_i	number of system components in state i at the start of maintenance activities
B_i	number of spare components in state i at the start of maintenance activities
C_i	number of spare components in state i at the end of maintenance activities
W_i	number of components in state i to repair during the maintenance period
$R_\mu(X)$	time needed to repair X components given repair rate μ and repair capacity c
$Z_\mu(X)$	number of components repaired in time X given repair rate μ and capacity c
$Av_{m,S,c}$	the system availability, given the maintenance intiation level m, number of spares S and the repair capacity c
T_m	time until maintenance initiation given maintenance initiation level m
U_m	uptime during the lead-time L, given maintenance initiation level m
$D_{m,S,c}$	downtime caused by maintenance activities, given maintenance initiation level m, number of spares S and repair capacity c

If the number of ready-for-use spares is less than the number of components to be replaced at the start of maintenance, the maintenance period is extended by the time needed to repair the remaining components. The repair time of a component in state i (i.e., the time needed for a transition to state 0) is exponentially distributed with mean $1/\mu_i$ $(i = 1,2)$. Because it is plausible that degraded components can be repaired faster than failed components, we assume $\mu_2 \leqslant \mu_1$. The repair shop contains c parallel, identical servers that are able to handle both types of repair jobs. We denote the system state by the triple (n_0,n_1,n_2), with n_i the number of components in state i (hence $n_0+n_1+n_2 = N$). Equivalently, we denote the state of the spares as (s_0,s_1,s_2), with $s_0+s_1+s_2 = S$.

For this system, we address the following questions in order to obtain a certain prescribed availability level:

1. *What is the influence of the system state/condition to initiate maintenance activities* (*choice of* m)?
2. *What is the influence of the number of spare parts* (*choice of* S)?
3. *What is the influence of the repair capacity* (*choice of* c)?
4. *What is the influence of the priority rule used for the two types of repair jobs*?

Of course, there is a clear interaction between the various parameters, as several choices of m, S, and c might lead to the same system availability. Although this gives rise to a multi-dimensional optimisation problem, the emphasis here is on the analysis of the availability as a function of m, S, and c.

We model the evolution of the system state (n_0, n_1, n_2) as a renewal process, see Fig. 2. A system cycle starts when the system is as-good-as-new. The operational period lasts until maintenance is initiated upon the mth failure. During the lead-time L, the system is still operational and will degrade further, where it may even fail. Then maintenance starts, components are replaced and the system is as-good-as-new again.

A second cycle, the spares cycle, describes the evolution of the spares state (s_0, s_1, s_2). It starts when system maintenance has just been finished. Then, the spares in state 1 and state 2 represent repair jobs that have to be addressed during the next operational period of the system plus the lead-time. When the system arrives for maintenance, the failed and degraded components are replaced by good ones. If the number of ready-for-use spares is insufficient ($s_0 < n_1 + n_2$), we have to wait until the remaining components have been repaired. Note that the spares cycle is *not* a renewal process, because subsequent cycles are generally not independent. However, we treat this cycle as a renewal process. The state of the spare parts in the beginning of each cycle may be different and therefore we use a stationary distribution.

A complication is that the system cycle and the spare cycle are interrelated. We can explain this intuitively as follows. Suppose that in a certain system cycle the operational time is relatively long. Then it is likely that many components have been degraded until the time that m components have failed. Hence, the number of components in state 1 (n_1) is relatively large at the start of maintenance. At the same time, the number of restored spares (s_0) is likely to be relatively large when the operational time is relatively long. Therefore the system state and the state of the spares at the beginning of the maintenance period are not independent. As an approximation, however, we assume that both cycles are independent. Whether these approximations have a significant impact, is discussed when comparing our approximate methods with results from discrete event simulation (Section 5).

As a performance measure, we focus on the *limiting* or *steady state* system availability, defined as the quotient of the system uptime, consisting of the time to maintenance initiation T_m and the uptime during the lead-time U_m, and the total cycle, being the time to maintenance initiation T_m plus the lead-time L plus the maintenance time $D_{m,S,c}$ (see Fig. 2):

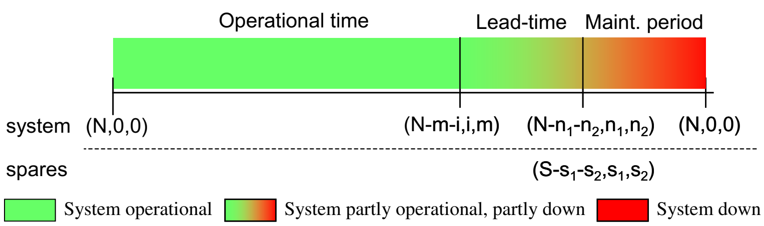

Fig. 2. Schematic presentation of the system's cycle above and the spares' cycle beneath.

$$Av_{m,S,c} = \frac{E[T_m] + E[U_m]}{E[T_m] + L + E[D_{m,S,c}]}. \tag{1}$$

In Section 3, we discuss how to calculate $E[T_m]$, $E[U_m]$ and $E[D_{m,S,c}]$ exactly if there is no correlation between the system state and spares state. Because numerical problems arise for large systems when calculating $E[U_m]$ and $E[D_{m,S,c}]$, we describe approximations for $E[U_m]$ and $E[D_{m,S,c}]$ which are applicable for larger systems in Section 4. In Section 6 we discuss some variants of the model as described in this section: a wider class of maintenance policies based on both degraded and failed spares (Section 6.1), maintenance on failed components only (Section 6.2) and redundant components in the cold standby mode (Section 6.4).

3. Method A

From now on, we omit in our notation, the indices referring to the parameters m, S, c, N and k since these are fixed parameters during the calculations. Assuming that all components in state 1 and state 2 are replaced by new ones during maintenance and there is no correlation between the system state and the spares state we compute $E[T]$, $E[U]$ and $E[D]$ exactly in Sections 3.1 till 3.3 respectively. In Section 3.4 we discuss the computational issues.

3.1. Expected operational time

The operational time until maintenance initiation T is the time until the mth component failure ($1 \leqslant m \leqslant N-k+1$). If $L=0$, it is clear that we should choose $m = N-k+1$. If $L > 0$, m is likely to be chosen smaller. The distribution function $F(t)$ for T is given by

$$F(t) = \Pr\left(\text{number of failed components at } t \geqslant m\right) = \sum_{i=m}^{N} \binom{N}{i} (p_{02}(t))^i (1 - p_{02}(t))^{N-i}, \tag{2}$$

where $p_{02}(t)$, the probability that a component will move from state 0 to state 2 in time t, equals $1 - e^{-\lambda_1 t} - \frac{\lambda_1}{\lambda_1 - \lambda_2}(e^{-\lambda_2 t} - e^{-\lambda_1 t})$. Although we could derive $E[T]$ from $\int_{t=0}^{\infty}(1 - F(t))\,dt$, it is far easier to use a recursive approach. Let us define $T(i,j)$ as the time needed for a transition from state $(N-i-j, i, j)$ to the set of states in which maintenance is initiated ($n_2 = m$). Obviously, we have that $T(i,m) = 0$. If $j < m$, the mean value of $T(i,j)$ equals the expected sojourn time in the current state $(N-i-j, i, j)$ plus the expected time needed from the next state on. The expected sojourn time in state $(N-i-j, i, j)$ equals $\tau(i,j) = ((N-i-j)\lambda_1 + i\lambda_2)^{-1}$. Next, the system state changes to $(N-i-j-1, i+1, j)$ with a probability $\alpha(i,j) = \frac{(N-i-j)\lambda_1}{(N-i-j)\lambda_1 + i\lambda_2}$ and to $(N-i-j, i-1, j+1)$ with a probability $\beta(i,j) = \frac{i\lambda_2}{(N-i-j)\lambda_1 + i\lambda_2}$. Note that if $i = 0$ then $\alpha(i,j) = 1$ and $\beta(i,j) = 0$ and if $i + j = N$ then $\alpha(i,j) = 0$ and $\beta(i,j) = 1$, Hence

$$E[T(i,j)] = \begin{cases} 0 & \text{if } j = m, \\ \tau(i,j) + \alpha(i,j)E[T(i+1,j)] + \beta(i,j)E[T(i-1,j+1)] & \text{else.} \end{cases} \tag{3}$$

Observing that $E[T] = E[T(0,0)]$, we can compute this value starting with $E[T(i,m)] = 0$. It can be shown that we need $\frac{1}{2}m(m+N-3)$ simple computations, which is no problem at all from a computational perspective.

3.2. Expected uptime during L

We denote the uptime during the lead-time L by U, which can be written as L minus the downtime during L, so

$$E[U] = L - \int_{t=0}^{L} \sum_{j=N-k+1}^{N} \sum_{i=0}^{N-j} Q(i,j,t)\,\mathrm{d}t. \tag{4}$$

Here $Q(i,j,t)$ is defined as the probability of reaching state $(N-i-j,i,j)$ at time t, given that there were m failed components at time 0 (the time of maintenance initiation). This implies that each system state $(N-x-m,x,m)$ with $x \in [0,i+j-m]$ is possible at time 0. We define $P(x,m)$ as the probability of the system being in state $(N-x-m,x,m)$ at time 0. We define the number of transitions from state 1 to state 2 as y, with $y \in [\max\{0,x-i\},\min\{j-m,x\}]$. Given the system state at time 0, the system state at time t and the number of transitions from state 1 to state 2, we also know the number of transitions from state 0 to state 1 and the number of transitions from state 0 to state 2. The probability of a component transition from state i to state j is denoted as $p_{ij}(t)$. Hence,

$$\begin{aligned} Q(i,j,t) = {} & \sum_{x=0}^{i+j-m} P(x,m) \sum_{y=\max\{0,x-i\}}^{\min\{j-m,x\}} \binom{x}{y}\binom{N-m-x}{i-x+y}\binom{N-m-i-y}{j-m-y}(p_{00}(t))^{N-i-j}(p_{01}(t))^{i-x+y} \\ & \times (p_{02}(t))^{j-m-y}(p_{11}(t))^{x-y}(p_{12}(t))^{y}. \end{aligned} \tag{5}$$

We can explicitly write the transition probabilities as

$$p_{00}(t) = \mathrm{e}^{-\lambda_1 t},\ \ p_{01}(t) = \frac{\lambda_1}{\lambda_1 - \lambda_2}\left(\mathrm{e}^{-\lambda_2 t} - \mathrm{e}^{-\lambda_1 t}\right) \text{ and } p_{11}(t) = \mathrm{e}^{-\lambda_2 t},$$

where obviously $p_{12}(t) = 1 - p_{11}(t)$ and $p_{02}(t) = 1 - p_{00}(t) - p_{01}(t)$. Now let us derive an expression for the probability $P(x,m)$. Because at this state $(N-x-m,x,m)$ maintenance is initiated, it can only be reached through a transition from state $(N-x-m,x+1,m-1)$. As a result, we obtain a recursive calculation scheme for the probabilities of reaching each possible system state until maintenance initiation. This scheme is given in Eq. (6), and an example is illustrated in Fig. 3.

$$P(i,j) = \begin{cases} 1 & \text{if } i=j=0, \\ \alpha(i-1,j)P(i-1,j) + \beta(i+1,j-1)P(i+1,j-1) & \text{else.} \end{cases} \tag{6}$$

3.3. Expected maintenance duration

For the expected maintenance duration $E[D]$, we condition on the system state and the spares state just before the system arrives for maintenance at the repair shop. Because we assume that the spares cycle and the system cycle are independent (see Section 2), we have that

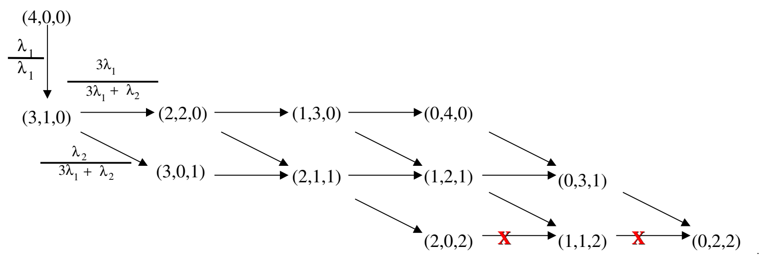

Fig. 3. Example of a 2-out-of-4 system with $m = 2$. Transitions from $(1,1,2)$ to $(0,2,2)$ and from $(2,0,2)$ to $(1,1,2)$ are not taken into account, because these states would have initiated maintenance themselves.

$$E[D] = \sum_{s_0=0}^{S}\sum_{s_2=0}^{S-s_0}\sum_{n_2=m}^{N}\sum_{n_1=0}^{N-n_2} P_L(n_1,n_2)\pi(s_0,s_2)E[R(n_1+n_2-s_0,n_1+S-s_0-s_2,n_2+s_2)]. \tag{7}$$

$P_L(n_1,n_2)$ is the probability of the system having n_1 degraded and n_2 failed components when actual maintenance activities starts and $\pi(s_0,s_2)$ is the steady state probability of the spares inventory consisting of s_0 ready-for-use spares and s_2 failed spares (note that $s_1 = S - s_0 - s_2$). $R(r,\tilde{s}_1,\tilde{s}_2)$ is the time to repair r components with capacity c when there are $\tilde{s}_1$ degraded and $\tilde{s}_2$ failed components available at the start of the repair. Regarding the system state at maintenance initiation, we have that $n_2 = m$ and $P_L(n_1,m) = P(n_1,m)$ if $L = 0$. If $L > 0$ then $P_L(n_1,n_2) = Q(n_1,n_2,L)$. We provide expressions for $R(r,\tilde{s}_1,\tilde{s}_2)$ in Section 3.3.1 and $\pi(s_0,s_2)$ in Section 3.3.2.

3.3.1. Repair time and priority rule

The repair time $R(r,\tilde{s}_1,\tilde{s}_2)$ depends on the order in which the $\tilde{s}_1$ degraded and $\tilde{s}_2$ failed components are repaired, which can be given by a certain *repair priority rule*. The other way round, the repair priority rule influences the spares state at the start of system maintenance. We try to minimise the maintenance duration by repairing as many degraded components as possible. When the number of degraded components is not sufficient to replace all failed and degraded components ($n_1 + s_1 < N - s_0 - n_0$), we have to repair some failed components as well. It is well known that we minimise the makespan (and hence the system repair time) by selecting the longest mean processing times first, see Pinedo and Chao (1999). So we use the following repair priority rule:

If there are sufficient degraded spares (in state 1), then only repair degraded components. If the number of degraded components is insufficient, start repairing the minimum number of failed components needed to repair the system. Next, repair the degraded components.

If we have both degraded and failed spares to be repaired after system repair, we need a second repair priority rule. It is logical to aim for handling as many jobs as possible before $T + L$ (the time between maintenance instances). Therefore we complete the jobs with the shortest mean repair time first.

Now let us apply the repair priority rule to find the mean repair time $E[R(r,\tilde{s}_1,\tilde{s}_2)]$. We define $r = [n_1 + n_2 - s_0]^+$ as the total number of repairs needed to repair the system, where $x^+ = \max\{0,x\}$ for any real number x. Then we need to restore $[r - s_1]^+$ failed spares during the maintenance period. Since we have at most c spares in the repair shop, we start with the $\min\{c,[r - s_1]^+\}$ failed spares in repair at the start of the repair period. Next, we assign $a = \min\{s_1, c - \min\{c,[r - s_1]^+\}\}$ degraded spares to the repair shop. If there is still repair capacity left, we use this capacity for the remainder of the failed components. The number of failed spares in the repair shop is now equal to $b = \min\{s_2, c - a\}$. Let us denote the number of components in state 1 and state 2 at the start of the repair period by a and b respectively. Using $R(r,\tilde{s}_1,\tilde{s}_2) = 0$ if $r = 0$ or $s_1 < 0$ or $s_2 < 0$, we find the following recursive relation for the expected repair time:

$$E[R(r,\tilde{s}_1,\tilde{s}_2)] = \frac{1}{a\mu_1 + b\mu_2} + \frac{a\mu_1}{a\mu_1 + b\mu_2}E[R(r-1,\tilde{s}_1-1,\tilde{s}_2)] + \frac{b\mu_2}{a\mu_1 + b\mu_2}E[R(r-1,\tilde{s}_1,\tilde{s}_2-1)]. \tag{8}$$

3.3.2. The steady state probabilities of the spares states

We use a Markov chain to determine the steady state probabilities $\pi(i)$. Here we use a short hand notation $i = (s_0,s_2)$. We want to solve the steady state conditions $\pi = M^{\mathrm{T}}\pi$ with $\sum\pi(i) = 1$. Each entry (i,j) of the transition matrix M equals the transition probability q_{ij} that $j = (s_0',s_2')$ is the spares state just before the maintenance period starts, while the spares state just before the previous maintenance period was $i = (s_0,s_2)$. We calculate the probability q_{ij} by conditioning on the time to maintenance initiation $T = t$:

$$q_{ij} = \sum_{n_2=m}^{N} \sum_{n_1=0}^{N-n_2} P_L(n_1,n_2) \int_{t=0}^{\infty} f(t) H\big(\min\{s_1+n_1,[S-s_2-n_2]^+\}, \min\{s_2+n_2,S\}, s_1', s_2', t+L\big)\mathrm{d}t, \tag{9}$$

where $f(t)$ is the density function of T and $H(w,x,y,z,t)$ is the probability that the spares state changes from w degraded and x failed spares to y degraded and z failed spares during t with c servers. In the special case $L = 0$, we have that $n_2 = m$ and so the transition probabilities consist of one summation only. The density function $f(t)$ can be found as the derivative of $F(t)$ from Eq. (2):

$$f(t) = \sum_{g=m}^{N} \sum_{h=0}^{N-g} \binom{N}{g} \binom{N-g}{h} (-1)^h (g+h) \lambda_2 p_{01}(t) (p_{02}(t))^{g+h-1}. \tag{10}$$

Let us now derive an expression for $H(w,x,y,z,t)$. We first note that only a non-negative number of spares can be restored, so $H(w,x,y,z,t) = 0$ if $w < y$ and/or $x < z$. Because our repair priority rule states that we should first restore degraded components, it is not possible to restore one or more failed components if the number of degraded components remaining is at least equal to the number of servers; thus $H(w,x,y,z,t) = 0$ if $y \geqslant c$ and $x > z$. If no spares are restored (i.e. $w = y$ and $x = z$) we have a repair rate that is equal to $\min\{c,w\}\mu_1 + \min\{x, c - \min\{c,w\}\}\mu_2$ and therefore $H(w,x,y,z,t)$ decreases exponentially with that rate. If spares are restored, we distinguish two cases: one in which all spares are being repaired immediately ($w + x \leqslant c$) and one in which not all repairs, but only c repairs, start immediately ($w + x > c$).

In the first case we have a combination of two binomial distributions, one with parameters w and $\mathrm{e}^{-\mu_1 t}$, and one with parameters x and $\mathrm{e}^{-\mu_2 t}$. In the second case, where $w + x > c$, we can write $H(w,x,y,z,t)$ in a recursive formulation. In case $w \leqslant c$, it is possible to have a failed spare restored before a degraded spares is restored or the other way around. In case $w > c$, the only possibility is to restore a degraded spare. In the recursive formulation, y and z play the role of fixed parameters, which we suppress for readability. We find that

$$H(w,x,t) = \begin{cases} 0, & w<y \vee x<z \vee (y \geqslant c \wedge x>z), \\ \mathrm{e}^{-\left(\min\{c,w\}\mu_1 + \min\left\{x,[c-w]^+\right\}\mu_2\right)t}, & w=y \wedge x=z, \\ \binom{w}{y}\binom{x}{z} \mathrm{e}^{-(y\mu_1+z\mu_2)t}(1-\mathrm{e}^{-\mu_1 t})^{w-y}(1-\mathrm{e}^{-\mu_2 t})^{x-z}, & w+x \leqslant c, \\ \int_{\tau=0}^{t} W(\tau)((c-w)\mu_2 H(w,x-1,t-\tau) + w\mu_1 H(w-1,x,t-\tau))\mathrm{d}\tau, & w+x>c \wedge w \leqslant c, \\ \int_{\tau=0}^{t} w\mu_1 W(\tau) H(w-1,x,t-\tau)\mathrm{d}\tau, & w+x>c \wedge w>c. \end{cases} \tag{11}$$

Here $W(\tau) = \mathrm{e}^{-w\mu_1\tau}\mathrm{e}^{-(c-w)\mu_2\tau}$. From Eqs. (10) and (11) we are able to determine all the elements q_{ij} of the transition matrix M (see Eq. (9)).

3.4. Computational issues

Our approach has several drawbacks. First, Eqs. (4) and (5) contain binominals of high order. Therefore, we encountered numerical problems and long computation times when evaluating the equations for larger systems (say $N > 80$). The computation time for a system with $N = 80$ components is several hours on a Pentium II, 800 MHz pc. A similar problem occurs for the maintenance duration (Eqs. (7)–(9)). Furthermore, we found that the computation time to evaluate the transition probabilities becomes large even for smaller problems, because we evaluated the integrals numerically. The number of integrals is very large because of the recursive character of Eq. (11). The system size for which we can find the maintenance duration within a reasonable amount of time (less than about an hour) is up to 10 or 20 components.

4. Method B

Because of the drawbacks of method A, we developed simpler and faster approximations for $E[U]$ and $E[D]$. These approximations are based on the first two moments by fitting an appropriate distribution. For continuous distributions on $[0,\infty)$ we use phase type distributions, see Tijms (1994). For discrete distributions on $[0,1,2,\ldots)$ we use either a mixture of two binomial distributions, a mixture of two negative binomial distributions, a mixture of two geometric distributions or a Poisson distribution dependent on the mean and the variance, see Adan et al. (1995).

4.1. Expected uptime during L

Let us denote the time from maintenance initiation to system failure by $\widehat{T}$, the time from m component failures until the $(N-k+1)$th component fails if $L\to\infty$. Then the mean uptime during the lead-time equals $E[\min\{\widehat{T},L\}]=E[\widehat{T}]-E[[\widehat{T}-L]^+]$. We can evaluate such an expression easily for specific classes of probability distributions, particularly for phase type distributions (for example, hyperexponential distributions or mixtures of Erlang distributions). Therefore, a simple approximation is to calculate the first two moments of $\widehat{T}$ exactly and next to approximate the distribution of $\widehat{T}$ by a mixture of Erlang distributions with the same first two moments. Such an approach has appeared to be fruitful in many applications where the performance measure to be approximated does not depend heavily on the tails of the probability distribution, see Tijms (1994).

The first two moments of $\widehat{T}$ can be found by conditioning on the system state at maintenance initiation:

$$E[\widehat{T}]=\sum_{i=0}^{N-m}P(i,m)E[\widehat{T}(i,m)], \tag{12}$$

$$E[\widehat{T}^2]=\sum_{i=0}^{N-m}P(i,m)E[\widehat{T}^2(i,m)]. \tag{13}$$

Here $\widehat{T}(i,m)$ is the time until the $(N-k+1)$th component failure occurs when the system is in state $(N-i-m,i,m)$ at maintenance initiation. $E[\widehat{T}(i,m)]$ is found analogously to Eq. (3) with only a small difference in the restriction, which becomes equal to $j=N-k+1$. For the second moment, the recursion is not straightforward because the transition depends on the sojourn time. After some algebra we find Eq. (14) with $E[\widehat{T}^2(i,j)]=0$ if $j=N-k+1$.

$$E\left[\widehat{T}^2(i,j)\right]=\left\{2\tau(i,j)E\left[\widehat{T}(i,j)\right]+\alpha(i,j)E\left[\widehat{T}^2(i+1,j)\right]+\beta(i,j)E\left[\widehat{T}^2(i-1,j+1)\right]\right\}. \tag{14}$$

4.2. Expected maintenance duration

The basic idea for approximation of the mean system downtime is to use a moment iteration scheme as has been proposed by De Kok (1989) for the analysis of the waiting time in the $G/G/1$ queue. First, we define W_1 and W_2 as stochastic variables for the number of repairs of type 1 and type 2 respectively during the maintenance time. We can approximate the maintenance duration by

$$E[D]\approx\frac{E[W_1]}{c\mu_1}+\frac{E[W_2]}{c\mu_2}. \tag{15}$$

This is an approximation, because we pretend as if first all c servers are busy with failed components at a joint rate $c\mu_2$ and next they are all busy with degraded components at a joint rate $c\mu_1$. The reality is that failed and degraded items can be repaired simultaneously and that the repair rate can be less than $c\mu_1$ at the end of the maintenance period if less than c components are available, leaving one or more servers idle. The variables W_1 and W_2 depend on the system state and the spares state at the start of maintenance. We define A_i as the number of system components and B_i as the number of spare components in state i ($i = 0, 1, 2$) when the system arrives for maintenance. Because of our repair priority rule, failed spares are only repaired if the total number of failed components exceeds the number of spares S, hence

$$W_2 = [A_2 + B_2 - S]^+. \tag{16}$$

The number of type 1 repairs equals the total number of components needed, which equals $[N - A_0 - B_0]^+$, minus the components that are obtained by repairing failed components:

$$W_1 = [N - A_0 - B_0]^+ - W_2. \tag{17}$$

The B_i depend on the number of spares in each state at the end of the previous maintenance period. Defining the variables C_i as the number of spare components in state i after the maintenance is finished:

$$C_0 = [B_0 - (A_1 + A_2)]^+ = [B_0 + A_0 - N]^+,$$

$$C_1 = S - C_0 - C_2,$$

$$C_2 = \min\{A_2 + B_2, S\}.$$

Because we start with repairing type 1 spares when maintenance is finished, C_1 decreases with the number of type 1 spares that can be repaired during $T + L$ with capacity c and repair rate μ_1, which is denoted by $Z_{\mu_1}(T + L)$. If time is left, C_2 decreases with the number of failed spares that can be restored during the remaining time. Therefore we denote $R_{\mu_1}(C_1)$ as the time needed to restore C_1 components with repair rate μ_1 and capacity c. For B_i we find

$$B_0 = S - B_1 - B_2,$$

$$B_1 = \left[C_1 - Z_{\mu_1}(T + L)\right]^+,$$

$$B_2 = \left[C_2 - Z_{\mu_2}\left(\left[T + L - R_{\mu_1}(C_1)\right]^+\right)\right]^+.$$

Unfortunately, we face correlations between A_i and B_i (see Section 2). To simplify calculations we assume that $B_1 = 0$. This means that $T + L$ is long enough to restore all spares that are degraded at the end of the maintenance period. This is a reasonable assumption, since the degraded spares have priority to be repaired. The set of equations is simplified to

$$B_0 = S - B_2, \tag{18}$$

$$B_2 = \left[C_2 - Z_{\mu_2}\left(T + L - R_{\mu_1}(C_1)\right)\right]^+, \tag{19}$$

$$C_0 = [B_0 + A_0 - N]^+, \tag{20}$$

$$C_1 = S - C_0 - C_2, \tag{21}$$

$$C_2 = \min\{A_2 + B_2, S\}. \tag{22}$$

Now we find $E[W_1]$ and $E[W_2]$ using the following moment iteration algorithm:

Step 0: Set the first two moments of B_2 to 0, determine the first two moments of A_0 and A_2.
Step 1: Fit a discrete distribution to $A_2 + B_2$ assuming that A_2 and B_2 are uncorrelated.
Step 2: Determine the first two moments of B_0, using Eq. (18) and fit a discrete distribution for $A_0 + B_0$ assuming that A_0 and B_0 are uncorrelated.
Step 3: Calculate the mean and the variance of C_0 and C_2 using two moment approximations (20) and (22).
Step 4: Calculate the mean and the variance of C_1 from (21) taking into account $\mathrm{cov}(C_1, C_2)$.
Step 5: Calculate the mean and the variance of B_2 by approximating the first two moments of $X = S - (C_2 - Z_{\mu_2}(T + L - R_{\mu_1}(C_1)))$ and using $B_2 = [S - X]^+$.
Step 6: Determine $E[W_2]$ and $E[W_1]$ from (16) and (17) using the mean of $W = [N - (A_0 + B_0)]^+$. If the convergence criterion is not satisfied then go to step 1, otherwise stop.

The first two moments of A_0 and A_2 that we need for step 0 are relatively easy to find:

$$E[A_0] = \sum_{i=0}^{N-m} P(i,m)(N-m-i)p_{00}(L),$$

$$E[A_0^2] = \sum_{i=0}^{N-m} P(i,m)(N-m-i)p_{00}(L)\{1 - p_{00}(L) + (N-m-i)p_{00}(L)\},$$

$$E[A_2] = m + \sum_{i=0}^{N-m} P(i,m)\{(N-m-i)p_{02}(L) + ip_{12}(L)\},$$

$$\begin{aligned} E\left[A_2^2\right] = \sum_{i=0}^{N-m} P(i,m)\big\{&(N-m-i)p_{02}(L)\{1 - p_{02}(L) + (N-m-i)p_{02}(L)\} + ip_{12}(L)\{1 - p_{12}(L) + ip_{12}(L)\} \\ &+ 2i(N-m-i)p_{02}(L)p_{12}(L) + 2m(N-i-m)p_{02}(L) + 2mip_{12}(L) + m^2\big\}. \end{aligned}$$

For step 2 we have $E[B_0] = S - E[B_2]$ and $\mathrm{var}[B_0] = \mathrm{var}[B_2]$ and we are able to fit a distribution for $A_0 + B_0$.

For step 3 we have a distribution for $A_0 + B_0$ from step 1, which allows us to determine $E[C_0]$ and $\mathrm{var}[C_0]$. $E[C_2]$ and $\mathrm{var}[C_2]$ are found using the distribution we found for $A_2 + B_2$ in step 0 or in step 5.

In step 4 we use Eq. (23) with only the covariance as an unknown term.

$$\begin{aligned} &E[C_1] = S - E[C_0] - E[C_2], \\ &\mathrm{var}[C_1] = \mathrm{var}[C_0] - \mathrm{var}[C_2] - 2\mathrm{cov}[C_1, C_2]. \end{aligned} \tag{23}$$

For $\mathrm{cov}[C_1, C_2]$ we condition on A_1:

$$\begin{aligned} \mathrm{cov}[C_1, C_2] &= E[\mathrm{cov}[C_1, C_2 \mid A_1]] + \mathrm{cov}[E[C_1 \mid A_1], E[C_2 \mid A_1]] = \mathrm{cov}[\min\{A_1, S - C_2\}, C_2] \\ &= \Pr(A_1 > S - C_2)\mathrm{cov}[S - C_2, C_2] = -\mathrm{var}[C_2]\Pr(A_1 > S - C_2) \\ &= -\mathrm{var}[C_2]\Pr(N - A_0 + B_2 > S) = -\mathrm{var}[C_2]\Pr(A_0 + B_0 < N). \end{aligned}$$

In step 5 we define $X = S - (C_2 - Z_{\mu_2}(T + L - R_{\mu_1}(C_1)))$ and we approximate the mean and the variance of $R_{\mu_1}(C_1)$ and $Z_{\mu_2}(X)$ by

$$E\left[R_{\mu_1}(C_1)\right] \approx \frac{E[C_1]}{c\mu_1},$$

$$\operatorname{var}\left[R_{\mu_1}(C_1)\right] \approx \frac{E[C_1]}{(c\mu_1)^2} + \frac{\operatorname{var}[C_1]}{(c\mu_1)^2},$$

$$E\left[Z_{\mu_2}(X)\right] \approx c\mu_2 E[X],$$

$$\operatorname{var}\left[Z_{\mu_2}(X)\right] \approx c\mu_2 E[X] + (c\mu_2)^2 \operatorname{var}[X].$$

The mean and the variance of X are now written by

$$E[X] = S - E[C_2] + c\mu_2\left(E[T] + L - \frac{E[C_1]}{c\mu_1}\right),$$

$$\operatorname{var}[X] = c\mu_2\left(E[T] + L - \frac{E[C_1]}{c\mu_1}\right) + (c\mu_2)^2\operatorname{var}[T] + \left(\frac{\mu_2}{\mu_1}\right)^2 (E[C_1] + \operatorname{var}[C_1]) + \operatorname{var}[C_2] + 2\frac{\mu_2}{\mu_1}\operatorname{cov}[C_1, C_2].$$

For B_2 we find $E[B_2] = E[[S - X]^+]$ and $\operatorname{var}[B_2] = \operatorname{var}[[S - X]^+]$. With the mean and the variance for A_2 from step 0 and the mean and the variance for B_2 from step 4, we fit a discrete distribution to $A_2 + B_2$.

In step 6 we determine $E[W_2]$ using the distribution we found for $A_2 + B_2$ in step 5. We use the distribution for $A_0 + B_0$ to find the mean for the total workload $W = [N - A_0 - B_0]^+$. Then $E[W_1]$ is found by $E[W] - E[W_2]$.

5. Numerical examples

Because both methods as discussed in Sections 3 and 4 are approximations, we need to test the accuracy of both methods. To this end, we constructed a discrete event simulation model as benchmark. The simulation results given in this paper are based on 5000 cycles.

We computed over 460 scenarios divided into three different system sizes: 7-out-of-10 system, 58-out-of-64 system and 2700-out-of-3000 system.

In Table 1 we give an overview of the different scenarios we used. We used method A as presented in Section 3 (if feasible within a few hours computation time), method B as presented in Section 4. To compare these results with our simulation model, we need to make sure that the simulation results are accurate. In our simulation model we compute T, U and D for a number of cycles. Given 95% confidence intervals for the maintenance duration, we found a relative accuracy of 6%, 2.5% and 0.2% for 7-out-of-10 systems, 58-out-of-64 systems and 2700-out-of-3000 systems, respectively. The values for the availability are even better.

The computation times for the 7-out-of-10 system using method A vary between 0.25 and 30 seconds, dependent on the number of spares. Using method A for the 58-out-of-64 system the computation times are at least 140 minutes per scenario. With method B the dependency on the system size is very small and the computation times found are around 0.01 seconds per instance. All computation times are measured on a Pentium II, 800 MHz pc.

Table 1
Input used for our numerical examples; within brackets the stepsize is given

	L	λ_1	λ_2	μ_1	μ_2	m	S	c
7-out-of-10	30, 168	0.01	0.05	0.2	0.1	$1\cdots 3(1)$	$1\cdots 6(1)$	$1\cdots 3(1)$
58-out-of-64	168	0.000125	0.00025	0.05	0.03	$1\cdots 7(1)$	$1\cdots 10(1)$	$1\cdots 3(1)$
2700-out-of-3000	168	2.9×10^{-5}	5.8×10^{-5}	0.125	0.0625	$250\cdots 300(10)$	$250\cdots 350(20)$	$5\cdots 20(5)$

Table 2
For different system sizes the mean and maximum differences for the repair time per method based on roughly 200 instances for the small system, 100 instances for the medium sized system and 150 instances for the large system

	Mean differences		Max. differences	
	Method A (%)	Method B (%)	Method A (%)	Method B (%)
7-out-of-10 system	2.7	4.2	44.3	22.4
58-out-of-64 system	–	1.4	–	10.6
2700-out-of-3000 system	–	0.2	–	0.9

The maximum correlation found for the systems between the system cycle (number of components in state 1 at the start of maintenance) and the spares cycle (number of spares in state 0 at the start of maintenance) is 0.08 at the most, which justifies our model approximation to neglect this correlation. The differences in $E[D]$ between the computations according to method A and method B compared to simulation are given in Table 2.

For the 7-out-of-10 system the maximum differences are rather large. However, when we only take into account the scenarios with $L = 168$ the maximum difference for method A reduces to 1.2%. This is due to the fact that a lead-time of 30 is too small for the assumption that all spares of type one will be restored. For method B the maximum difference hardly changes, only the mean difference changes to 1.0%. The instances

Table 3
Some results for different system sizes with a comparable availability results for different combinations of values for maintenance initiation, number of spares and repair capacity

Input	$E[T]$		$E[U]$			$E[D]$			Av		
	A	Sim.	A	B	Sim.	A	B	Sim.	A	B	Sim.
$N = 10$, $k = 7$, $L = 30$ $m = 1$, $S = 4$, $c = 2$ $\lambda_1 = 0.01$, $\lambda_2 = 0.05$ $\mu_1 = 0.2$, $\mu_2 = 0.1$	22.44	22.86	27.46	27.48	27.29	2.03	1.85	1.93	0.92	0.92	0.92
$N = 10$, $k = 7$, $L = 30$ $m = 2$, $S = 5$, $c = 3$ $\lambda_1 = 0.01$, $\lambda_2 = 0.05$ $\mu_1 = 0.2$, $\mu_2 = 0.1$	37.76	38.01	22.46	22.44	22.57	1.23	0.99	1.23	0.87	0.88	0.88
$N = 64$, $k = 58$, $L = 168$ $m = 1$, $S = 3$, $c = 2$ $\lambda_1 = 0.000125$, $\lambda_2 = 0.00025$ $\mu_1 = 0.05$, $\mu_2 = 0.03$	950	916	168	168	168	–	52.86	54.16	–	0.95	0.96
$N = 64$, $k = 58$, $L = 168$ $m = 4$, $S = 2$, $c = 3$ $\lambda_1 = 0.000125$, $\lambda_2 = 0.00025$ $\mu_1 = 0.05$, $\mu_2 = 0.03$	2230	2247	167	167	167	–	108	111	–	0.96	0.95
$N = 3000$, $k = 2700$, $L = 168$ $m = 250$, $S = 250$, $c = 10$ $\lambda_1 = 2.9 \times 10^{-5}$, $\lambda_2 = 5.8 \times 10^{-5}$ $\mu_1 = 0.125$, $\mu_2 = 0.0625$	11 740	11 745	–	168	168	–	506	508	–	0.96	0.97
$N = 3000$, $k = 2700$, $L = 168$ $m = 300$, $S = 270$, $c = 10$ $\lambda_1 = 2.9 \times 10^{-5}$, $\lambda_2 = 5.8 \times 10^{-5}$ $\mu_1 = 0.125$, $\mu_2 = 0.0625$	13 102	13 104	–	26.52	27.39	–	580	582	–	0.95	0.95

with the largest differences, have a rather extreme combination of parameters, e.g. $m = 1$, $S = 6$ and $c = 1$. This results in high utilisation rates, which gives uncertainty about the assumption that all type 1 spares are repaired before the next maintenance period. Larger lead-times reduce this uncertainty and therefore give better results for the repair time.

The maximum difference of 10.6% for a 58-out-of-64 system is also obtained in a rather extreme situation where $m = S = 1$ and $c = 3$. Leaving out such scenarios, the maximum difference would be 5%.

For the 2700-out-of-3000 system scenarios we found similar results as we did for the other systems. Note that the *average* error found is smaller than the *average* simulation accuracy for all three systems. Of course, the approximate values may be outside the corresponding 95% confidence interval for individual cases.

Next, we show that various combinations of control parameters (m, S, c) may lead to a similar system availability. In Table 3, we give six examples (two for each system) with comparable availability. To give an impression of the different possibilities for achieving a certain availability see Fig. 4 for a 7-out-of-10 system.

In Fig. 5 we show the availability of the 2700-out-of-3000 system as a function of S for different values of c. The value of m is chosen such that the availability is maximal (without bothering about the effects on the cycle length or cost).

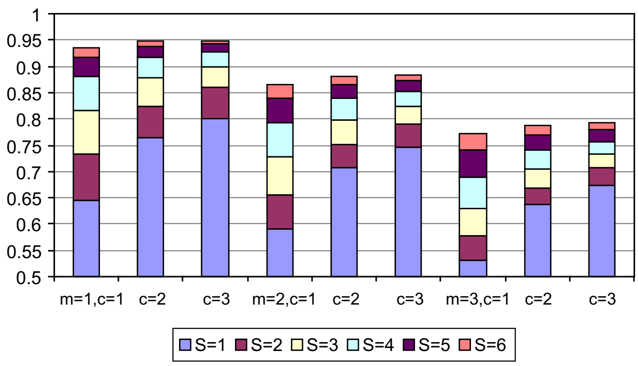

Fig. 4. Columns are depicted for a 7-out-of-10 system with different values for maintenance initiation and capacity. In each column a new shading represents an extra spare. Given a desired availability level, each column shows the parameter combination needed to reach this level.

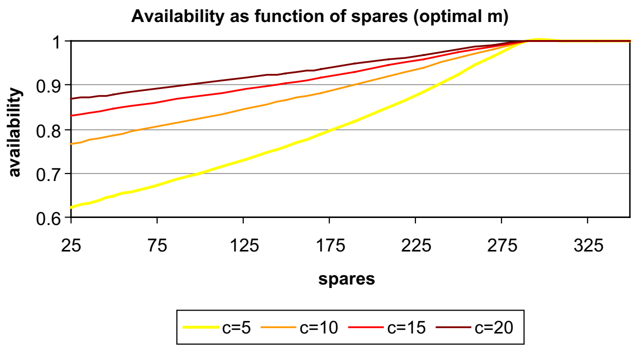

Fig. 5. For a 2700-out-of-3000 system and several values of capacity we show the availability as a function of the spares amount. The maintenance initiation level is chosen such that the availability is the highest.

6. Model extensions

6.1. Maintenance also based on degraded components

Until now, we discussed a maintenance policy dependent only on the number of failed components. If we are able to observe the number of degraded components in the system during the operational time, we could use another rule. Denoting a system state as (i,j), which means there are i degraded components and j failed components in the system, we have one set with all system states $\Omega = (i,j)$ with $i \in [0,N]$ and $j \in [0,N-i]$. We divide Ω into three subsets:

Ω_U: all system states in which the system is operational and maintenance is not yet initiated.
Ω_M: all system states in which the system is operational and maintenance has been initiated.
Ω_D: all system sets in which the system has failed.

Of course, the sets need to be defined such that it is impossible to make a transition to a state in Ω_U once the system state is in one of the other sets, except caused by maintenance.

The expression for the operational time until maintenance initiation only changes slightly. In Eq. (3), the condition $j = m$ changes into $(i,j) \in \Omega_M$. For the expected uptime during the lead-time, Eq. (4) remains similar because the definition of a failed system remains unchanged. We define a subset of Ω_M with only the system states that initiate maintenance, thus the system states $(i,j) \in \Omega_M$ with $(i+1,j-1) \notin \Omega_M$ or $(i-1,j) \notin \Omega_M$, denoted by Ω_I. The uptime is estimated using the equations of Section 4.1. Taking into account the number of degraded components, the expressions for $E[\widehat{T}]$ and $\mathrm{var}[\widehat{T}]$ are modified by replacing m by j and we sum over $(i,j) \in \Omega_I$. In expression (6), for $P(i,j)$ we only change the restriction $j = m$ into $(i,j) \in \Omega_I$. For the expected maintenance duration, we only change $E[A_i]$ and $E[A_i^2]$ for $i = 0, 1, 2$. This change is similar to the other changes: replace m by j and sum over $(i,j) \in \Omega_I$.

Note that a maintenance policy should define the set Ω_M. Optimisation of such a maintenance policy is not straightforward, but at least we are able to evaluate the consequences of a given choice. Explicit optimisation is subject for further research.

6.2. Replacement of failed components only

If it is impossible to distinguish the condition of type 0 and 1 components, we can only replace failed components during maintenance. The system state at the start of a cycle is then unknown and could be any state $(N-i, i, 0)$ with $0 \leqslant i \leqslant N-m$. As a consequence we cannot use $E[T(0,0)]$ because the system is not as-good-as-new at the start of the cycle. We adjust the equation to

$$E[T] = \sum_{i=0}^{N-m} P_{\text{start}}(i) E[T(i,0)]. \tag{24}$$

Here $P_{\text{start}}(i)$ is the probability that the system state at the start of a cycle is equal to $(N-i, i, 0)$. This probability $P_{\text{start}}(i)$ equals the sum of probabilities of the system being in state $(N-i-j, i, j)$ with $j = m, \ldots, N-i$ at the start of the preceding maintenance period.

$$P_{\text{start}}(i) = \sum_{j=m}^{N-i} P_{\text{maint}}(i,j). \tag{25}$$

Here $P_{\text{maint}}(i,j)$ is the probability that the system state equals $(N-i-j, i, j)$ at the start of maintenance. This probability depends on the system state at maintenance initiation and state transitions during the lead-time:

$$P_{\text{maint}}(i,j) = \sum_{h=0}^{i+j-m} P_{\text{init}}(h,m) P_{\text{trans}}((h,m),(i,j),L)\}, \quad j \geqslant m. \tag{26}$$

Here $P_{\text{trans}}((h,m),(i,j),L)$, the probability of a transition from state $(N-h-m,h,m)$ to state $(N-i-j,i,j)$ in time L, which is given by

$$P_{\text{trans}}((h,m),(i,j),L) = \sum_{y=[h-i]^+}^{\min\{j-m,h\}} \binom{h}{y}\binom{N-h-m}{N-i-j}\binom{i+j-h-m}{j-y-m}(p_{00}(L))^{N-i-j}(p_{01}(L))^{i+y-h} \times (p_{02}(L))^{j-y-m}(p_{11}(L))^{h-y}(p_{12}(L))^{y}.$$

The probability $P_{\text{init}}(h,m)$ is defined as the probability of the system state being $(N-h-m,h,m)$ at maintenance initiation, which is a function of the system state at the start of the cycle:

$$P_{\text{init}}(i,m) = \sum_{h=0}^{i+m} P_{\text{start}}(h) P_{h,m}(i,m). \tag{27}$$

$P_{h,m}(i,j)$ is the probability of reaching state $(N-i-j,i,j)$ given initial state $(N-h,h,0)$ and maintenance initiation at m failed components. This probability is found recursively using

$$P_{h,m}(i,j) = \begin{cases} 1 & \text{if } (i,j) = (h,0), \\ \alpha(i-1,j)P_{h,m}(i-1,j) + \beta(i+1,j-1)P_{h,m}(i+1,j-1) & \text{else.} \end{cases}$$

By filling in Eq. (27) into Eq. (26) filled into Eq. (25) we have a set of equations with only $P_{\text{start}}(i)$ which can be solved using $\sum_{i=0}^{N-m} P_{\text{start}}(i) = 1$. With $P_{\text{start}}(i)$, we have $E[T]$.

For the uptime during the lead-time we can use our approximation with $P_{\text{init}}(i,m) = P(i,m)$.

For the maintenance duration our model becomes less complex because we only have type 2 components in our repair shop. This enables us to use the method we used in our model without ageing (see De Smidt-Destombes et al., 2004) with the repair rate equal to μ_2.

When considering large systems we encounter the same problem with $P_{\text{trans}}((h,m),(i,j),L)$ as we did before with $Q(i,j,t)$. An alternative is to use a moment iteration approach. To find the distribution of the system being in state $(N-i,i,0)$ is equal to finding the distribution of A_1 with the first two moments:

$$E[A_1] = \sum_{i=0}^{N-m} P(i,m)\{(N-m-i)p_{01}(L) + ip_{11}(L)\},$$

$$E[A_1^2] = \sum_{i=0}^{N-m} P(i,m)\{(N-m-i)p_{01}(L)\{1-p_{01}(L)+(N-m-i)p_{01}(L)\} + ip_{11}(L)\{1-p_{11}(L) + ip_{11}(L)\} + 2i(N-m-i)p_{01}(L)p_{11}(L)\}.$$

The distribution of $P(i,m)$ is the only expression that changes. We start by choosing an initial distribution for A_1. Then we determine $P(i,m)$ using the recursion of Eq. (3). We then have $E[A_1]$ and $E[A_1^2]$. By iteration we find the system state distribution at the start of the cycle.

6.3. Stochastic lead-time L

In our model we assumed the lead-time to be deterministic. In the case of a stochastic lead-time we have to adjust the calculations for $E[U]$ and $E[D]$. For method A this means changing Eqs. (4) and (5). We could do this by conditioning on the lead-time. Eq. (5) results in terms of the form e^{xL}. The expectation of these terms is found using the Laplace transform of L and taking a Gamma function for instance. Adjusting Eq. (4) can also be done but takes more effort.

In method B only the second expectation of $E[U] = E[\widehat{T}] - E[[\widehat{T} - L]^+]$ changes. Because $\widehat{T}$ and L are independent it is rather easy. For $E[D]$ Eq. (19) for B_2 and the equations for the first and second moments of A_0 and A_2 change. In Eq. (19) we need $Z_{\mu_2}(T + L - R_{\mu_1}(C_1))$ for which the mean and the variance are still the same because T, L and $R_{\mu_1}(C_1)$ are independent of one another. The expressions for the moments of A_i we condition on L and find terms of the form e^{xL}. The expectation of these terms is found by using Laplace transforms and a Gamma distribution for L for instance. See for a more detailed explanation (De Smidt-Destombes et al., submitted for publication).

6.4. Cold standby redundancy

If components are easily switched on, one may choose for cold standby redundancy instead of hot standby redundancy. This results in a system with k active components that degrade while being used, whereas the other components are inactive and therefore are not subject to degradation. This variant is known as cold standby redundancy. This changes the transition probabilities between and the sojourn times in system states. For $E[T]$ we modify $E[T(i,j)]$ from Eq. (3). We change $\tau(i,j) = \frac{1}{(k-i)\lambda_1 + i\lambda_2}$, $\alpha(i,j) = \frac{(k-i)\lambda_1}{(k-i)\lambda_1 + i\lambda_2}$ and $\beta(i,j) = \frac{i\lambda_2}{(k-i)\lambda_1 + i\lambda_2}$. For $E[U]$ we are able to use the approximation given in method B, for which $E[\widehat{T}(i,j)]$, $E[\widehat{T}^2(i,j)]$ and $P(i,j)$ changes equivalently to $E[T]$. For $E[D]$ the only parameters effected in method B are the A_i. If we assume $L = 0$ then we know the first and second moments for A_i by using $P(i,m)$. When $L > 0$, we encounter difficulties with the determination of the first and second moments. This is caused by the fact that we need to take into account the exact timing of the transitions. Otherwise we do not know the number of components in state 0 that are subject to failure.

Hence, we can analyse cold standby redundancy if $L = 0$, but need another approach if $L > 0$.

7. Conclusions and further research

In this paper, we introduced component wear-out in a model for the trade-off between spare part inventories, repair capacity and maintenance policy. This extension implies a lot of complications. The first complication is the correlations between different parameters. The state of the spares at the start of maintenance is not independent of the state of the system at the start of maintenance. Even if we ignore this correlation, we found it impossible to compute the different expressions we need to determine the availability. On the one hand it is impossible because of large binomials in the expression for the uptime during the lead-time. On the other hand it is impossible because of the large state space for the spares needed to compute the steady state probabilities of the spares at the start of the maintenance period. Especially if we want to use the model presented in the paper as a basic model for an optimisation between cost and availability we are in need of an accurate model with small computation times. Our numerical examples show that the second approximation (Section 4) fulfils these requirements and can be used for this purpose.

In our further research we aim to extend to model a situation in which there are several identical systems using the same repair capacity and spare parts in order to keep the system running. This problem can be divided into two different problems. In the first one we assume there are a lot of systems using the same means for their maintenance, which implies the arrival process at the repair shop is approximately Poisson. The second problem is the one in which we have a few systems, which implies that is impossible to assume the arrival process to be Poisson. In this case it may be useful not only to take into account systems that have reached the level m in order to be maintained but also systems that have less than m failed components. This way the workload in the repairshop is more spread over time and the number of available systems is more stable. In practice it is not unreasonable to have such a policy in order to have at least a certain percentage of the total number of systems available at any time.

Another extension would be to take into account multiple types of components. Making the model multi-item could imply a set of m_i, one for each type of components. Also the spares become a set of spares, S_i. The capacity is used for the repair of all the types of components. The repair strategy for this problem is not evident because spares of all kinds are needed during the maintenance period. Last but not the least, further research includes the practical application of the model using field data to show the applicability of the model. In such a case study, attention has to be paid to explicit optimisation of spares and repair capacity.

References

Adan, I., Van Eenige, M., Resing, J., 1995. Fitting discrete distributions on the first two moments. Probability in the Engineering and Informational Sciences 9 (4), 623–632.

Avsar, Z., Zijm, W., 2003. Capacitated Two-echelon Inventory Models for Repairable Item Systems. Kluwer Academic Publishers.

De Kok, A., 1989. A moment-iteration method for approximating the waiting time characteristics of the G/G/1 queue. Probability in the Engineering and Informational Sciences 3, 273–287.

De Smidt-Destombes, K., Van der Heijden, M., Van Harten, A., 2004. On the availability of a k-out-of-N system given limited spares and repair capacity under a condition based maintenance strategy. Reliability Engineering and System Safety 83 (3), 287–300.

De Smidt-Destombes, K., Van der Heijden, M., Van Harten, A., submitted for publication. Spare parts analysis for k-out-of-n systems under block replacement and finite repair capacity. International Journal of Production Economics.

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

Dinesh Kumar, U., Crocker, J., Knezevic, J., El-Haram, M., 2000. Reliability, Maintenance and Logistic Support: A Life Cycle Approach. Kluwer Academic Publishers.

Gross, R., Miller, D., Soland, R., 1985. On common interests among reliability, inventory and queuing. IEEE Transactions on Reliability 34 (3), 204–208.

Kabir, A., Al-Olayan, A., 1996. A stocking policy for spare part provisioning under age based preventive replacement. European Journal of Operational Research 90, 171–181.

Kennedy, W., Patterson, J., Fredendall, L., 2002. An overview of recent literature on spare parts inventories. International Journal of Production Economics 76, 201–215.

Park, Y., Park, S., 1986. Generalized spare ordering policies with random lead time. European Journal of Operational Research 23, 320–330.

Pinedo, M., Chao, X., 1999. Operations Scheduling: With Applications in Manufacturing and Services. McGraw-Hill.

Rustenburg, W., 2000. A System Approach to Budget-Constrained Spare Parts Management. PhD Thesis, BETA Research Institute.

Sarkar, J., Sarkar, S., 2001. Availability of a periodically inspected system supported by a spare unit under perfect repair or upgrade. Statistics & Probability Letters 53 (2), 207–217.

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

Sleptchenko, A., Van der Heijden, M., Van Harten, A., 2003. Trade-off between inventory and repair capacity in spare part networks. Journal of Operations Resarch Society 54 (3), 263–272.

Tijms, H., 1994. Stochastic Models: An Algorithmic Approach. John Wiley & Sons.

Wang, H., 2002. A survey of maintenance policies of deteriorating systems. European Journal of Operational Research 139, 469–489.

Zijm, W., Avsar, Z., 2003. Capacitated two-indenture models for repairable item systems. International Journal of Production Economics 81–82 (C), 573–588.