RELIABILITY ANALYSIS OF CHECKPOINTING MODEL WITH MULTIPLE VERIFICATION MECHANISM

YUTAE LEE

ABSTRACT. We consider a checkpointing model for silent errors, where a checkpoint is taken every fixed number of verifications. Assuming generally distributed i.i.d. inter-occurrence times of errors, we derive the reliability of the model as a function of the number of verifications between two checkpoints and the duration of work interval between two verifications.

1. Introduction

Reactive failure management techniques to mitigate the impact of errors are required to ensure a correct and uninterrupted execution of an application in high performance computing [17]. The classical fail-stop errors are defined as fatal interruptions, such as hardware failures or crashes, that call the faulty node for a reboot or replacement [7]. In addition to fail-stop errors, silent errors, also called silent data corruptions or latent errors, constitute another threat that can no longer be ignored and should also be accounted for [1,18,19,23–25]. Contrarily to a fail-stop error whose detection is immediate, a silent error is not detected immediately because the error is identified only when the corrupted data is activated or leads to an unusual application behavior [1,4].

Checkpoint with rollback and recovery is the standard recovery technique for coping with fail-stop errors [12,14]. An application employing checkpoints periodically saves its state, so that when an error occurs while some task is executing, the application is rolled back to its last checkpointed task and resumes execution from that task onward [1,4]. Given the value of Mean Time Between Failures (MTBF), an approximation of the optimal checkpoint interval has been computed as a function of the key parameters: downtime, checkpointing time, and recovery time. The first estimate was made by Young [22], and later by Daly [13]. Both used a first-order approximation for exponential failure distributions. Formulas for Weibull failure distribution, one of the most
widely used lifetime distributions, were provided in [10, 15, 20]. However, the optimal checkpoint interval is known only for exponential failure distributions [9]. Dynamic programming heuristics for arbitrary distributions were proposed in [9, 11, 21].

While checkpoint with rollback and recovery is the standard approach to address fail-stop errors, there is no widely used technique to cope with silent errors [4, 5]. A major challenge with silent errors is the latency of error detection. If an error stroke before the last checkpoint, which saved an already corrupted state, and is detected after that checkpoint, then the checkpoint is corrupted and may not be used to recover from the error [4]. To alleviate this issue, Lu et al. [16] introduced a multiple checkpointing model with error detection latency. The error is detected after a random time and one has to rollback up to the last valid checkpoint that precedes the occurrence of the error. A problem concerning this concept is that it is not clear how one can determine when the error has indeed occurred and hence how one can identify the last valid checkpoint. To address the problem with silent errors, researchers have recently proposed many verification mechanisms. If a verification is successful, then the output of the task is correct, and one can safely either proceed to the next task directly or save the result beforehand by taking a checkpoint. Otherwise, if a verification fails, we have to rollback to the last saved checkpoint and re-execute the work since that point on [4]. Aupy et al. [1] proposed a model with periodic patterns coupling $k$ verifications and 1 checkpoint, where silent errors are detected only through some verification mechanism. They computed the values $k$ and $\tau$ to minimize the waste, i.e., maximize the reliability, where $\tau$ is the work interval between two verifications. Benoit et al. [8] extend the analysis of [1] by including $k$ verifications and $l$ checkpoints. In [1, 8], they neglected the possibility of having more than one error within a periodic pattern and assumed that an error, if any, is located uniformly within the first $k$ verifications in a periodic pattern. Due to the assumption, the optimal values derived in [1] only depend on the mean value of the inter-occurrence times of errors, not the shape of their distribution.

Bautista-Gomez et al. [2] first investigated the use of different types of partial detectors while taking both recall and precision into consideration. The objective was to find the optimal pattern that minimizes the expected execution time of an application. They first showed that detectors with imperfect precision offer limited usefulness. Then, focusing on detectors with perfect precision but imperfect recall, they conducted a comprehensive complexity analysis of the optimization problem and proposed a greedy algorithm to solve the problem. In [2], the inter-occurrence times of silent errors were assumed to follow an exponential distribution.

Although numerous studies have dealt with either fail-stop or silent errors, very few studies have dealt with both simultaneously [3–7]. Benoit et al. [5] introduced a general-purpose model to deal with both fail-stop and silent errors, combining checkpoints with a verification mechanism. They considered
three execution scenarios: (i) a single speed is used during the whole execution; (ii) a second, possibly higher speed is used for any potential re-execution; (iii) different pairs of speeds can be used throughout the execution. For the three execution scenarios and for both makespan and energy objectives, they provided a dynamic programming algorithm that determines the best locations of checkpoints and verifications. In [6], they presented a unified framework and optimal algorithmic solutions to the double problem of fail-stop and silent errors. Silent errors are handled via verification mechanisms and in-memory checkpoints. Fail-stop errors are processed via disk checkpoints. All verification and checkpoint types are combined into computational patterns. They designed a detailed model based upon the computational patterns and determined the optimal pattern. In [7], they combined multi-level checkpointing with guaranteed and partial verifications to deal with both fail-stop and silent errors in linear workflows. Benoit et al. [3] and Benoit et al. [4] combined checkpointing and replication for the reliable execution on platforms subject to both fail-stop and silent errors. All of these results assume that the inter-occurrence times of errors are exponentially distributed.

In this paper, we focus on silent errors whose inter-occurrence times are generally distributed i.i.d. random variables. We consider a checkpointing model with multiple verification mechanism, where a checkpoint is taken every $k$ verifications. Moreover, we remove the conventional assumption that at most one error may occur between two checkpoints. Given the cost of checkpointing, downtime, recovery, and verification, we drive the reliability as a function of the number of verifications between two checkpoints and the duration of the work interval between two verifications.

2. Model

We consider a checkpointing model where silent errors are detected only through some verification mechanism [1]. Consider a scenario where $k$ consecutive successful verifications and 1 checkpoint are included in a periodic pattern $T$ that repeats over time (Figure 1).

![Figure 1. Model with $k = 3$](image)

Each verification is performed after executing a work of duration $\tau$. Let $V$ be the time it takes to perform a verification. If $k$ consecutive verifications reveal no errors, then a checkpoint directly follows the last verification, so as to save correct results. If a verification detects an error, then we roll back to the last checkpoint at the end of the previous pattern and re-execute the work.
Let $C$ be the checkpointing time it takes to create a checkpoint. A series of unexpected errors can be occurred and the inter-occurrence time $X$ of errors is generally distributed with probability distribution function $F_X(x)$. When an error is detected, the downtime and the recovery time, denoted by $D$ and $R$, respectively, is required. The downtime $D$ represents the unavoidable time to rejuvenate a process after an error, which is required for stopping the failed process and restoring a new one that will load the checkpoint image [1]. The recovery time $R$ represents the time to reload the information stored at the checkpoint back into primary memory [1]. The values $\tau$, $C$, $D$, and $R$ are assumed to be constant. If there are no errors during recovery time, then at the end of the recovery time the system operates normally and restarts from the last checkpoint. It is assumed that errors can not occur during checkpoint and downtime.

We assume that the inter-occurrence times of errors are generally distributed i.i.d. random variables with well-known distribution function, not limited to exponential distribution. Thus, if the inter-occurrence times of errors in a platform with several processors are generally distributed i.i.d. random variables and all the processors are rejuvenated anew after each error, we can apply the derivation in this paper to the platform with several processors. However, even if the error occurrence distribution of one processor is known, it is difficult to compute, or even approximate, the error distribution of the platform with $n$ processors, because it is the superposition of $n$ random variables. Furthermore, the assumption that all the processors are rejuvenated anew after each error is unreasonable for a large parallel platform [9]. Thus, on the practical side, the derivation in the next section applies only with a single processor.

3. Stochastic analysis

Let $S(t)$ be the number of checkpoints created after the last error occurrence at time $t$. Note that each periodic pattern begins immediately after a checkpoint is created. Let the time sequence $\{t_n, n = 1, 2, \ldots\}$ be the epochs at which checkpoints are created. Then, the time sequence $\{t_n, n = 1, 2, \ldots\}$ is the Markov points embedded in the process $\{S(t), t \geq 0\}$. The embedded Markov chain $\{S_n, n = 1, 2, \ldots\}$ is defined by

$$(1) \quad S_n \equiv S(t_n).$$

We also define a semi-Markov process $\bar{S}(t)$ by

$$(2) \quad \bar{S}(t) \equiv S_{\sup\{n, t_n \leq t\}}, \quad t \geq 0.$$ 

Notation “sup” (superior limit) means the limit from above. The value $\bar{S}(t)$ remains constant from one Markov point to the next.

Let $T_i$, $i = 1, 2, \ldots$, be the residence time that the semi-Markov process is in state $i$, i.e., $T_i$ be the length of a periodic pattern beginning with $S_n = i$, which does not depend on $n$. To obtain the reliability, we need to calculate the
expected value of $T_i$:

$$E(T_i) = E(t_{n+1} - t_n | S_n = i), \quad i = 1, 2, \ldots$$

Once the expected value $E(T_i)$ is obtained, the mean value of the length $T$ of an arbitrary periodic pattern is given by

$$E(T) = \lim_{n \to \infty} E(t_{n+1} - t_n) = \sum_{i=1}^{\infty} \pi_i E(T_i),$$

where $\{\pi_i, i = 1, 2, \ldots\}$ is the limiting distribution of $\{S_n, n = 1, 2, \ldots\}$, defined by

$$\pi_i \equiv \lim_{n \to \infty} P(S_n = i), \quad i = 1, 2, \ldots$$

First, we will obtain the limiting distribution $\{\pi_i, i = 1, 2, \ldots\}$, which is given as the unique solution to the following equations:

$$\pi_j = \sum_{i=1}^{\infty} \pi_i p_{ij}, \quad j = 1, 2, \ldots,$$

$$\sum_{i=1}^{\infty} \pi_i = 1,$$

where

$$p_{ij} \equiv P(S_{n+1} = j | S_n = i)$$

is the state-transition probability, which is independent of $n$. The state transition from $S_n = i$ to $S_{n+1} = 1$ occurs when there is at least one error during the $n$-th periodic pattern with $S_n = i$ and the transition from $S_n = i$ to $S_{n+1} = i+1$ occurs when there are no errors during the $n$-th periodic pattern with $S_n = i$.

Thus, we have

$$p_{ij} = \begin{cases} 1 - q_{i,k}^+, & i = 1, 2, \ldots, \quad j = 1, \\ q_{i,k}^+, & i = 1, 2, \ldots, \quad j = i + 1, \\ 0, & i = 1, 2, \ldots, \quad j \neq 1, \quad j \neq i + 1, \end{cases}$$

where

$$q_{i,j}^+ \equiv P(X - [R + ik(\tau + V)] > j(\tau + V) | X > R + ik(\tau + V))$$

for $i = 1, 2, \ldots$ and $j = 1, 2, \ldots$. Using the state-transition probabilities, we have

$$\pi_i = \pi_{i-1} q_{i-1,k}, \quad i = 2, 3, \ldots$$

and thus

$$\pi_i = \pi_1 \prod_{j=1}^{i-1} q_{j,k}^+ = \pi_1 q_{1,(i-1)k}^+, \quad i = 1, 2, \ldots.$$
The unknown probability \( \pi_1 \) is determined from the normalization condition
\[
\sum_{i=1}^{\infty} \pi_i = 1:
\]
(13)
\[
\pi_1 = \frac{1}{1 + \sum_{j=1}^{\infty} q_{1,jk}^+}.
\]

Note that
\[
1 + \sum_{j=1}^{\infty} q_{1,jk}^+ = E \left( \left\lfloor \frac{X - \lfloor R + k(\tau + V) \rfloor}{k(\tau + V)} \right\rfloor \middle| X > R + k(\tau + V) \right),
\]
where \( \lfloor a \rfloor \) is the least integer greater than or equal to \( a \).

Now, we obtain the expected value \( E(T_i) \) of the length \( T_i \), which is given by
\[
E(T_i) = k(\tau + V) + C + \sum_{j=1}^{k} q_{i,j} [j(\tau + V) + D + R]
\]
\[
+ \left( 1 - q_{i,k}^+ \right) \frac{1 - r_k^+}{r_k^+} \sum_{j=1}^{k} \frac{r_j}{1 - r_k^+} [j(\tau + V) + D + R],
\]
where
\[
r_1 \equiv P(X \leq R + (\tau + V)),
\]
(16)
\[
r_j \equiv P(R + (j - 1)(\tau + V) < X \leq R + j(\tau + V)), \quad j = 2, 3, \ldots, k,
\]
(17)
\[
r_k^+ \equiv P(X > R + k(\tau + V)),
\]
(18)
\[
q_{i,j} \equiv P((j - 1)(\tau + V) < X - \lfloor R + ik(\tau + V) \rfloor \leq j(\tau + V) \middle| X > R + ik(\tau + V))
\]
(19)
for \( i, j = 1, 2, \ldots \). Note that in (15) the part \( k(\tau + V) + C \) corresponds to the length of a periodic pattern with no errors; the part
\[
\sum_{j=1}^{k} q_{i,j} [j(\tau + V) + D + R]
\]
corresponds to the length added due to the first error during a periodic pattern; the part \( 1 - q_{i,k}^+ \) of the last term corresponds to the probability that the first error occurs; the part \( (1 - r_k^+) / r_k^+ \) of the last term corresponds to the mean number of errors, excluding the first one, during a periodic pattern in the condition that the first error occurs; and the last part
\[
\sum_{j=1}^{k} \frac{r_j}{1 - r_k^+} [j(\tau + V) + D + R]
\]
of the last term corresponds to the length added due to each error (not the first one) during the periodic pattern. We rewrite (15) to

$$E(T_i) = k(\tau + V) + C + \frac{1 - q_i^+}{v_i^+} (D + R)$$

$$+ (\tau + V) \sum_{j=1}^{k} \left[ q_{i,j} + \left(1 - q_i^+ \right) \frac{r_j}{v_i^+} \right].$$

Substituting (12), (13), and (20) into (4), we obtain the mean value $E(T)$. Then, the reliability $Rel(\tau, k)$ is given by

$$Rel(\tau, k) = \frac{k\tau}{E(T)}.$$  

3.1. Closed form expression for exponential error occurrence case

The purpose of this subsection is just to give a closed form expression of the reliability (21) in the special case of an exponential error occurrence. The inter-occurrence time $X$ of errors is assumed to be exponentially distributed with rate $\lambda$ in this subsection. Because of the memoryless property of exponential distribution, the probability distribution of $T$ is the same as that of $T_i$ for any $i$. Hence,

$$E(T) = (\tau + V) \frac{1 - e^{-\lambda k(\tau + V)}}{e^{-\lambda (R + k(\tau + V))}} \left[ 1 + \frac{e^{-\lambda (R + \tau + V)}}{1 - e^{-\lambda (\tau + V)}} \right] + C$$

$$+ \frac{1 - e^{-\lambda k(\tau + V)}}{e^{-\lambda (R + k(\tau + V))}} (D + R)$$

and therefore the reliability $Rel(\tau, k)$ is given by

$$Rel(\tau, k) = \frac{k\tau}{(\tau + V) \frac{1 - e^{-\lambda k(\tau + V)}}{e^{-\lambda (R + k(\tau + V))}} \left[ 1 + \frac{e^{-\lambda (R + \tau + V)}}{1 - e^{-\lambda (\tau + V)}} \right] + C}$$

$$= \frac{k\tau \left(1 - e^{-\lambda (\tau + V)}\right) e^{-\lambda (R + k(\tau + V))}}{(\tau + V) \left(1 - e^{-\lambda k(\tau + V)}\right) \left\{1 - e^{-\lambda (\tau + V)} + e^{-\lambda (R + \tau + V)}\right\} + C \left(1 - e^{-\lambda (\tau + V)}\right) e^{-\lambda (R + k(\tau + V))} + (D + R) \left(1 - e^{-\lambda (\tau + V)}\right) \left(1 - e^{-\lambda k(\tau + V)}\right)}.$$

Actually, as for the exponential error occurrence case, researchers have considered more realistic settings of evaluation scenarios than this subsection. Dealing with several errors, different chunk lengths, and more complicated patterns has been the scope of the papers [2,5,6], as mentioned in Section 1.
4. Numerical examples

Table 1. Parameter values

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>20 seconds</td>
<td>2 seconds</td>
</tr>
<tr>
<td>C</td>
<td>600 seconds</td>
<td>60 seconds</td>
</tr>
<tr>
<td>D</td>
<td>0 seconds</td>
<td>0 seconds</td>
</tr>
<tr>
<td>R</td>
<td>600 seconds</td>
<td>60 seconds</td>
</tr>
<tr>
<td>E(X)</td>
<td>0.0001 years</td>
<td>0.0001 years</td>
</tr>
</tbody>
</table>

Let us provide some numerical examples. In our model, a checkpoint is taken every $k$ verifications of the current state of the application. We find the optimal pair $(k, \tau)$ to maximize the reliability, where the value $k$ is the number of verifications between two checkpoints and the value $\tau$ is the duration of the work interval between two verifications. We consider two scenarios. The parameter values for each scenario are summarized in Table 1, in which the second scenario is more optimistic [1]: the values $V$, $C$, and $R$ in Scenario

![Figure 2. Reliability for Scenario 1: (left) Exponential; (right) Weibull](image-url)
Figure 3. Reliability for Scenario 2: (left) Exponential; (right) Weibull

2 are divided by 10 compared to Scenario 1. In both scenarios, the mean inter-occurrence time \( E(X) \) of errors is set to 0.0001 years, equivalent to 52.56 minutes. The downtime is set to 0 seconds because the process rejuvenation time is small compared to the other parameters [1]. In each scenario, two different distributions with same mean are considered for the inter-occurrence times of errors: exponential distribution and Weibull distribution with shape parameter 2.

Figure 2 and Figure 3 illustrate the reliability by varying the number of verifications for different values of \( \tau \) in Scenario 1 and Scenario 2, respectively. Both figures show that the reliability increases, reaches a maximum value, and then decreases as the number of verifications increases. Note that intermediate verifications can detect an error before a checkpoint is created at the end of a periodic pattern, which can reduce the amount of time lost due to the error. On the contrary, introducing too many verifications induces too much overhead, which eventually decreases the reliability. Intuitively, as the value of \( \tau \) is higher, the optimal value of \( k \) to maximize the reliability is usually lower. Since Scenario 2 is more optimistic than Scenario 1, the reliability of Scenario 2 is higher than that of Scenario 1. Moreover, we find that, not only the mean
of the inter-occurrence time $X$ of errors, but the shape of the distribution also affects to the reliability. In Scenario 1, the optimal reliability is achieved at $k = 4$ and $\tau = 6$ minutes in case of exponential distribution for $X$ (the left in Figure 2) and it is achieved at $k = 3$ and $\tau = 6$ minutes in case of Weibull distribution for $X$ (the right in Figure 2), although the mean value $E(X)$ is the same in both cases. In Scenario 2, the optimal reliability is achieved at $k = 5$ and $\tau = 2$ minutes in case of exponential distribution (the left in Figure 3) and it is achieved at $k = 4$ and $\tau = 2$ minutes in case of Weibull distribution (the right in Figure 3).

5. Conclusion

We consider a checkpointing model for silent errors, in which a checkpoint is taken every $k$ verifications. Assuming generally distributed inter-occurrence times of errors, we analytically derive its reliability as a function of $k$ and $\tau$, where the value $k$ is the number of verifications between two checkpoints and the value $\tau$ is the duration of work interval between two verifications. In numerical examples with realistic parameters, we find the optimal values of $k$ and $\tau$ to maximize the reliability. Moreover, it is shown that not only the mean of inter-occurrence times of errors, but the shape of the distribution also affects to the reliability.

References


YUTAE LEE
DEPARTMENT OF INFORMATION AND COMMUNICATIONS ENGINEERING
DONG-EUI UNIVERSITY
BUSAN 47340, KOREA
Email address: ylee@deu.ac.kr