Towards Optimal Multi-Level Checkpointing
Anne Benoit, Aurélien Cavelan, Valentin Le Fèvre, Yves Robert, Hongyang Sun

To cite this version:

HAL Id: hal-02082416
https://inria.hal.science/hal-02082416
Submitted on 28 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Towards Optimal Multi-Level Checkpointing

Anne Benoit, Aurélien Cavelan, Valentin Le Fèvre, Yves Robert, Hongyang Sun

Abstract—We provide a framework to analyze multi-level checkpointing protocols, by formally defining a $k$-level checkpointing pattern. We provide a first-order approximation to the optimal checkpointing period, and show that the corresponding overhead is in the order of $\sum_{\ell=1}^{k} \sqrt{2 \lambda_{\ell} C_{\ell}}$, where $\lambda_{\ell}$ is the error rate at level $\ell$, and $C_{\ell}$ the checkpointing cost at level $\ell$. This nicely extends the classical Young/Daly formula on single-level checkpointing. Furthermore, we are able to fully characterize the shape of the optimal pattern (number and positions of checkpoints), and we provide a dynamic programming algorithm to determine the optimal subset of levels to be used. Finally, we perform simulations to check the accuracy of the theoretical study and to confirm the optimality of the subset of levels returned by the dynamic programming algorithm. The results nicely corroborate the theoretical study, and demonstrate the usefulness of multi-level checkpointing with the optimal subset of levels.

Index Terms—resilience, fail-stop errors, multi-level checkpointing, optimal pattern.

1 INTRODUCTION

Checkpointing is the de-facto standard resilience method for HPC platforms at extreme-scale. However, the traditional single-level checkpointing method suffers from significant overhead, and multi-level checkpointing protocols now represent the state-of-the-art technique. These protocols allow different levels of checkpoints to be set, each with a different checkpointing overhead and recovery ability. Typically, each level corresponds to a specific fault type, and is associated to a storage device that is resilient to that type. For instance, a two-level system would deal with (i) transient memory errors (level 1) by storing key data in main memory; and (ii) node failures (level 2) by storing key data in stable storage (remote redundant disks).

We consider a very general scenario, where the platform is subject to $k$ levels of faults, numbered from 1 to $k$. Level $\ell$ is associated with an error rate $\lambda_{\ell}$, a checkpointing cost $C_{\ell}$, and a recovery cost $R_{\ell}$. A fault at level $\ell$ destroys all the checkpoints of lower levels (from 1 to $\ell-1$ included) and implies a roll-back to a checkpoint of level $\ell$ or higher. Similarly, a recovery of level $\ell$ will restore data from all lower levels. Typically, fault rates are decreasing and checkpoint/recovery costs are increasing when we go to higher levels: $\lambda_{1} \geq \lambda_{2} \geq \cdots \geq \lambda_{k}$, $C_{1} \leq C_{2} \leq \cdots \leq C_{k}$, and $R_{1} \leq R_{2} \leq \cdots \leq R_{k}$.

The idea of multi-level checkpointing is that checkpoints are taken for each level of faults, but at different periods. Intuitively, the less frequent the faults, the longer the checkpointing period: this is because the risk of a failure striking is lower when going to higher levels; hence the expected re-execution time is lower too; one can safely checkpoint less frequently, thereby reducing failure-free overhead (checkpointing is useless in the absence of fault). There are several natural approaches to implement multi-level checkpointing. The first option is to use independent checkpointing periods for each level. This option raises several difficulties, the most prominent one being overlapping checkpoints. Typically, we need to checkpoint different levels in sequence (e.g., writing into memory before writing onto disk), so we would need to delay some checkpoints, which might not be possible in some environments, and which would introduce irregular periods. The second option is to synchronize all checkpoint levels by nesting them inside a periodic pattern that repeats over time, as illustrated in Figure 1(a). In this figure, the pattern has five computational segments, each followed by a level-1 checkpoint. A segment is a chunk of work between two checkpoints, and a pattern consists in segments and checkpoints. The second and fifth level-1 checkpoints are followed by a level-2 checkpoint. Finally, the pattern ends with a level-3 checkpoint. When using patterns, a checkpoint at level $\ell$ is always preceded by checkpoints at all lower levels 1 to $\ell-1$, which makes good sense in practice (e.g., with two levels, main memory and disk, one writes the data into memory before transferring it to disk).

Using periodic patterns simplifies the orchestration of checkpoints at all levels. In addition, repeatedly applying the same pattern is optimal for on-line scheduling problems, or for jobs running a very long (even infinite) time on the platform. Indeed, in this scenario, we seek the best pattern, i.e., the one whose overhead is minimal. The overhead of a pattern is the price per work unit to pay for resilience in the pattern; hence minimizing overhead is equivalent to optimizing platform throughput. For a pattern $P(W)$ with $W$ units of work (the cumulated length of all its segments), the overhead $H(P(W))$ is defined as the ratio of the pattern’s expected execution time $E(P(W))$ over its total work $W$ minus 1:

$$H(P(W)) = \frac{E(P(W))}{W} - 1. \quad (1)$$

If there were neither checkpoint nor fault, the overhead would be zero. Determining the optimal pattern (with minimal overhead), and then repeatedly using it until job completion, is the optimal approach with Exponential failure distributions and long-lasting jobs. Indeed, once a pattern is successfully executed, the optimal strategy is to re-execute the same pattern. This is because of the memoryless property of exponential distributions: the history of failures has no impact on the solution, so if a pattern is optimal at some point in time, it stays optimal.
The first sub-pattern has 2 segments while the second one has 3. The memoryless property does not imply that all sub-patterns are identical, because the state after completing the first sub-pattern is not the same as the initial state when beginning the execution of the pattern. In the general case with \( k \) levels, the shape of the pattern will be even more complicated, with different-shaped sub-patterns (each ended by a level 1 checkpoint). In turn, each sub-pattern may have different-shaped sub-sub-patterns (each ended by a level 2 checkpoint), and so on. The major contribution of this work is to provide an analytical characterization of the optimal pattern with an arbitrary number \( k \) of checkpointing levels, with closed-form formulas for the pattern length \( W_{\text{opt}} \), the number of checkpoints at each level, and the optimal overhead \( H_{\text{opt}} \). In particular, we obtain the following beautiful result:

\[
H_{\text{opt}} = \sum_{i=1}^{k} \sqrt{2\lambda_i C_i} + O(\Lambda),
\]

where \( \Lambda = \sum_{i=1}^{k} \lambda_i \). However, we point out that this analytical characterization relies on a first-order approximation, so it is valid only when resilience parameters \( C_i \) and \( R_i \) are small in front of the platform Mean Time Between Failures (MTBF) \( \mu = 1/\Lambda \). Also, the optimal pattern has rational number of segments, and we use rounding to derive a practical solution. Still, Equation (2) provides a lower bound on the optimal overhead, and this bound is very closely in all our experimental scenarios.

Finally, in many practical cases, there is no obligation to use all available checkpointing levels. For instance, with \( k = 3 \) levels, one may choose among four possibilities: level 3 only, levels 1 and 3, levels 2 and 3, and all levels 1, 2 and 3. Of course, we still have to account for all failure types, which translates into the following:

- level 3: use \( \lambda_3 \leftarrow \lambda_1 + \lambda_2 + \lambda_3 \);
- levels 1 and 3: use \( \lambda_1 \leftarrow \lambda_3 + \lambda_2 + \lambda_3 \);
- levels 2 and 3: use \( \lambda_2 \leftarrow \lambda_1 + \lambda_2 + \lambda_3 \);
- all levels: use \( \lambda_1, \lambda_2 \) and \( \lambda_3 \).

Our analytical characterization of the optimal pattern leads to a simple dynamic programming algorithm for selecting the optimal subset of levels.

The rest of this paper is organized as follows. Section 2 surveys the related work. Section 3 is the heart of the paper and shows how to compute the optimal pattern as well as the optimal subset of levels. Section 4 is devoted to simulations assessing the accuracy of the first-order approximation. Finally, Section 5 provides concluding remarks and hints for future work.

2 Related work

Given the checkpointing cost and platform MTBF, classical formulas due to Young [18] and Daly [6] are well known to determine the optimal checkpointing period in the single-level checkpointing scheme. However, this method suffers from the intrinsic limitation that the cost of checkpointing/recovery grows with failure probability, and becomes unsustainable at large scale [9], [4] (even with diskless or incremental checkpointing [15]).

To reduce the I/O overhead, various two-level checkpointing protocols have been studied. Vaidya [17] proposed a two-level recovery scheme that tolerates a single node failure using a local checkpoint stored on a parter node. If more than one failure occurs during any local checkpointing interval, the scheme resorts to the global checkpoint. Silva and Silva [16] advocated a similar scheme by using memory to store local checkpoints, which is protected by XOR encoding. Di et al. [8] analyzed a two-level checkpointing pattern, and proved equal-length segments in the optimal solution. They also provided mathematical equations that can be solved numerically to compute the optimal pattern length and number of segments. Benoit et al. [3] relied on disk checkpoints to cope with fail-stop failures and memory checkpoints coupled with error detectors to handle silent data corruptions. They derived first-order approximation formulas for the optimal pattern length and the number of memory checkpoints between two disk checkpoints.

Some authors have also generalized two-level checkpointing to account for an arbitrary number of levels. Moody et al. [14] implemented this approach in a three-level Scalable Checkpoint/Restart (SCR) library. They relied on a rather complex Markov model to recursively compute the efficiency of the scheme. Bautista-Gomez et al. [2] designed a four-level checkpointing library, called Fault Tolerance Interface (FTI), in which partner-copy and Reed-Solomon coding are employed as two intermediate levels between local and global disks. Based on FTI, Di et al. [7] proposed an iterative method to compute the optimal checkpointing interval for each level with prior knowledge of the application’s total execution time. Hakkarinen and Chen [11] considered multi-level diskless checkpointing for...

While all of these works relied on numerical methods to compute the checkpointing intervals at different levels, this paper is the first one to provide explicit formulas on the optimal parameters in a multi-level checkpointing protocol (up to first-order approximation as in Young/Daly’s classical result).

3 Computing the optimal pattern

This section computes the optimal multi-level checkpointing pattern. We first state our assumptions in Section 3.1, and then analyze the simple case with \( k = 2 \) levels in Section 3.2, before proceeding to the general case in Section 3.3. Finally, the algorithm to compute the optimal subset of levels is described in Section 3.4.

3.1 Assumptions

In this paper, we assume that failures from different levels are independent\(^2\). For each level \( \ell \), the arrival of failures follows a Poisson process with error rate \( \lambda_\ell \). In order to deal with the interplay of failures from different levels, we make use of the following well-known properties of independent Poisson processes \([10, \text{Chapter 2.3}]\).

Property 1. During the execution of a segment with length \( w \), let \( X_\ell \) denote the time when the first error (of any level) strikes. Thus, \( X_\ell \) is a random variable following an Exponential distribution with parameter \( \lambda_\ell \), for all \( \ell = 1, 2, \ldots, k \).

1. Let \( X \) denote the time when the first error (of any level) strikes. We have \( X = \min\{X_1, X_2, \ldots, X_k\} \), which follows an Exponential distribution with parameter \( \Lambda = \sum_{\ell=1}^k \lambda_\ell \). The probability of having an error (from any level) in the segment is therefore \( P(X \leq w) = 1 - e^{-\Lambda w} \).

2. Given that an error (from any level) strikes during the execution of the segment, the probability that the error belongs to a particular level is proportional to the error rate of that level, i.e., \( P(X = X_{\ell} | X \leq w) = \frac{\lambda_{\ell}}{\Lambda} \), for all \( \ell = 1, 2, \ldots, k \).

Moreover, we assume that error rates of different levels are of the same order, i.e., \( \lambda_\ell = \Theta(\Lambda) \) for all \( \ell = 1, 2, \ldots, k \), and that errors only strike during the computations, while checkpointing and recovery are error-free. Indeed, the durations of checkpoints and recoveries are generally small compared to the pattern length, so the probability of a failure striking during these operations is low. It has been shown in [3] that removing this assumption does not impact the first-order approximation of the pattern overhead.

3.2 Optimal two-level pattern

We start by analyzing the two-level pattern shown in Figure 1(b). The goal is to determine a first-order approximation to the optimal pattern length \( W \), the number \( n \) of level-1 checkpoints in the pattern, as well as the length \( w_i = \alpha_i W \) of the \( i \)-th segment, for all \( 1 \leq i \leq n \), where \( \sum_{i=1}^n \alpha_i = 1 \).

3.2.1 With a single segment

We first consider a special case of the two-level pattern, in which only a single segment is present, i.e., \( n = 1 \). The result establishes the order of the optimal pattern length \( W^{\text{opt}} \), which will be used later for analyzing the general case. Recall that \( \Lambda = \lambda_1 + \lambda_2 \) and, for convenience, let us also define \( C = C_1 + C_2 \). The following proposition shows the expected time of such a pattern with fixed length \( W \).

Proposition 1. The expected execution time of a two-level pattern with a single segment and fixed length \( W \) is

\[
E = W + C + \frac{1}{2} \Lambda W^2 + O(\max\{\Lambda^2 W^3, \Lambda W\}).
\]

Proof. We can express the expected execution time of the pattern recursively as follows:

\[
E = P \left( W^{\text{lost}}(W, \Lambda) + \frac{\lambda_1}{\Lambda} (R_1 + E) + \frac{\lambda_2}{\Lambda} (R_2 + R_1 + E) \right) + (1 - P) (W + C),
\]

where \( P = 1 - e^{-\Lambda W} \) denotes the probability of having a failure (either level-1 or level-2) during the execution of the pattern based on Property 1.1, and \( W^{\text{lost}}(w_i, \Lambda) \) denotes the expected time lost when such a failure occurs. In this case, and based on Property 1.2, if the failure belongs to level 1, which happens with probability \( \frac{\lambda_1}{\Lambda} \), we can recover from the latest level-1 checkpoint \( (R_1) \). Otherwise, the failure belongs to level 2 with probability \( \frac{\lambda_2}{\Lambda} \), and we need to first recover from the latest level-2 checkpoint \( (R_2) \) before restoring the level-1 checkpoint \( (R_1) \). In both cases, the entire pattern needs to be re-executed again. Finally, if no error (of any level) strikes, which happens with probability \( 1 - P \), the pattern is completed after \( W \) time of execution followed by the time \( C \) to perform the two checkpoints, which are assumed to be error-free.

From [13, Equation (1.13)], the expected time lost when executing a segment of length \( W \) with error rate \( \Lambda \) is

\[
W^{\text{lost}}(W, \Lambda) = \frac{1}{\Lambda} - \frac{W}{e^{\Lambda W} - 1}.
\]

Substituting Equation (4) into Equation (3) and solving for \( E \), we get:

\[
E = \left( e^{\Lambda W} - 1 \right) \left( \frac{1}{\Lambda} + R_1 + \frac{\lambda_2}{\Lambda} R_2 \right) + C_1 + C_2,
\]

which is an exact formula on the expected execution time of the pattern. Now, using Taylor series to expand \( e^{\Lambda W} = 1 + \Lambda W + \frac{\Lambda^2 W^2}{2} + O(\Lambda^3 W^3) \) while assuming \( W = \Theta(\Lambda^{-x}) \), where \( 0 < x < 1 \), we can re-write Equation (5) as

\[
E = W + \frac{1}{2} \Lambda W^2 + C_1 + C_2 + O(\Lambda^2 W^3)
\]

\[
+ \left( \Lambda W + \frac{\Lambda^2 W^2}{2} + O(\Lambda^3 W^3) \right) \left( R_1 + \frac{\lambda_2}{\Lambda} R_2 \right).
\]

Since recovery costs \( (R_1, R_2) \) are assumed to be constants, and error rates \( (\lambda_1, \lambda_2, \Lambda) \) are in the same order, the expected execution time can be expressed as follows:

\[
E = W + C_1 + C_2 + \frac{1}{2} \Lambda W^2 + O(\Lambda^2 W^3) + O(\Lambda W),
\]

which completes the proof of the proposition.

From Proposition 1, the expected execution overhead of the

\footnote{In practice, failures from different checkpointing levels can exhibit potential correlation \([12], [7]\). Consideration of correlated failures is beyond the scope of this paper.}
pattern can be derived as
\[ H = \frac{C}{W} + \frac{1}{2}AW + O(\max\{\Lambda^2 W^2, \Lambda\}). \]

Assume that the platform MTBF \( \mu = 1/\Lambda \) is large in front of the resilience parameters, and consider the first two terms of \( H \): the overhead is minimized when the pattern has length \( W = \Theta(\Lambda^{-1/2}) \), and in that case both terms are in the order of \( \Theta(\Lambda^{1/2}) \), so we have \( H = \Theta(\Lambda^{1/2}) + O(\Lambda) \). Indeed, the last term \( O(\Lambda^2 W^2) = O(\Lambda) \) becomes negligible compared to \( \Theta(\Lambda^{1/2}) \). Hence, the optimal pattern length \( W_{\text{opt}} \) can be obtained by balancing the first two terms in \( H \), which gives
\[ W_{\text{opt}} = \sqrt{\frac{2C}{\Lambda}} = \Theta(\Lambda^{-1/2}), \]  
and the optimal execution overhead becomes
\[ H_{\text{opt}} = \sqrt{2\lambda C} + O(\Lambda). \]  

Remarks. Unlike in single-level checkpointing, the checkpoint to roll back to in a two-level pattern depends on which type of error strikes first. Under first-order approximation and assuming that the resilience parameters are small compared to the platform MTBF and pattern length, the formulas shown in Equations (6) and (7) reduce exactly to Young/Daly’s classical result by aggregating the error rates and checkpointing costs of both levels.

### 3.2.2 With multiple segments

We now consider the general two-level pattern with multiple segments, and derive the optimal pattern parameters. As in the single-segment case, we start with a proposition showing the expected time to execute a two-level pattern with fixed parameters.

**Proposition 2.** The expected execution time of a given two-level pattern is
\[ E = W + nC_1 + C_2 + \frac{1}{2} \left( \lambda_1 \sum_{i=1}^{n} \alpha_i^2 + \lambda_2 \right) W^2 + O(\Lambda^{1/2}). \]

**Proof.** We first prove the following result (by induction) on the expected time \( E_i \) to execute the \( i \)-th segment of the pattern (up to the level-1 checkpoint at the end of the segment):
\[ E_i = w_i + C_1 + \frac{\lambda_1}{2} w_i^2 + \lambda_2 \left( \frac{w_i^2}{2} + \sum_{j=1}^{i-1} w_j w_i \right) + O(\Lambda^{1/2}). \]  

According to the result with a single segment, we know that the optimal pattern length and hence the segment length are in the order of \( O(\Lambda^{-1/2}) \), which implies that \( E_i = w_i + O(1) \).

For the ease of analysis, we assume that there is a hypothetical segment at the beginning of the pattern with length \( w_0 = 0 \) (hence no need to checkpoint). For this segment, we have \( E_0 = w_0 = 0 \), satisfying Equation (8). Suppose the claim holds up to \( E_{i-1} \). Then, \( E_i \) can be recursively expressed as follows:
\[ E_i = P_i \left( E_{\text{lost}}(w_i, \Lambda) + \frac{\lambda_1}{\Lambda} (R_1 + E_i) \right) + \frac{\lambda_2}{\Lambda} \left( R_2 + R_1 + \sum_{j=1}^{i-1} E_j + E_i \right) + (1 - P_i)(w_i + C_1), \]  
where \( P_i = 1 - e^{-\Lambda w_i} \) denotes the probability of having a failure (either level-1 or level-2) during the execution of the segment, and \( E_{\text{lost}}(w_i, \Lambda) \) denotes the expected time lost when such a failure occurs.

Equation (9) is very similar to Equation (3), except when a level-2 failure occurs we need to re-execute all the segments (up to segment \( i \)) that have been executed so far. Following the derivation of Proposition 1 and applying \( E_j = w_j + O(1) \) for \( j = 1, 2, \ldots, i-1 \), we can derive the first-order approximation of \( E_i \) as follows:
\[ E_i = w_i + C_1 + \frac{1}{2} \left( \lambda_1 w_i^2 + \lambda_2 w_i^2 + 2\lambda_2 w_i \sum_{j=1}^{i-1} w_j + O(1) \right) + O(\Lambda^{1/2}) \]
\[ = w_i + C_1 + \frac{1}{2} \left( \lambda_1 w_i^2 + \lambda_2 w_i^2 + 2\lambda_2 \sum_{j=1}^{i-1} w_j \right) + O(\Lambda^{1/2}). \]  

Since the level-2 checkpoint at the end of the pattern is also assumed to be error-free, we can compute the expected execution time of the pattern as
\[ E = \sum_{i=1}^{n} E_i + C_2 \]
\[ = W + nC_1 + C_2 + \frac{1}{2} \left( \lambda_1 \sum_{i=1}^{n} \alpha_i^2 + \lambda_2 \right) W^2 + O(\Lambda^{1/2}), \]

since \( \sum_{i=1}^{n} w_i^2 + 2 \sum_{i=1}^{n} \sum_{j=1}^{i-1} w_j w_i = (\sum_{i=1}^{n} w_i)^2 = W^2 \). \( \square \)

**Theorem 1.** A first-order approximation to the optimal two-level pattern is characterized by
\[ n_{\text{opt}} = \sqrt{\frac{\lambda_1}{C_1}}, \]  
\[ \alpha_i^{\text{opt}} = \frac{1}{n_{\text{opt}}} \forall i = 1, 2, \ldots, n_{\text{opt}}, \]  
\[ W_{\text{opt}} = \frac{n_{\text{opt}} C_1 + C_2}{\frac{1}{2} \left( \frac{\lambda_1}{n_{\text{opt}}} + \lambda_2 \right)}, \]  

where \( n_{\text{opt}} \) is the number of segments, \( \alpha_i^{\text{opt}} W_{\text{opt}} \) is the length of the \( i \)-th segment, and \( W_{\text{opt}} \) is the pattern length.

The optimal pattern overhead is
\[ H_{\text{opt}} = \sqrt{2\lambda_1 C_1} + \sqrt{2\lambda_2 C_2} + O(\Lambda). \]  

**Proof.** For a given pattern with a fixed number \( n \) of segments, \( \sum_{i=1}^{n} \alpha_i^2 \) is minimized subject to \( \sum_{i=1}^{n} \alpha_i = 1 \) when \( \alpha_i = \frac{1}{n} \) for all \( i = 1, 2, \ldots, n \). Hence, we can derive the expected execution overhead from Proposition 2 as follows:
\[ H = \frac{nC_1 + C_2}{W} \left( 1 + 2 \left( \frac{\lambda_1}{n} + \lambda_2 \right) \right) W + O(\Lambda). \]  

For a given \( n \), the optimal work length can then be computed from Equation (15), and it is given by \( W_{\text{opt}} = \sqrt{\frac{n C_1 + C_2}{2 \left( \frac{\lambda_1}{n} + \lambda_2 \right)}} \). In that case, the execution overhead becomes
\[ H = \sqrt{2 \left( \frac{\lambda_1}{n} + \lambda_2 \right)} (nC_1 + C_2) + O(\Lambda), \]
which is minimized as shown in Equation (14) when \( n \) satisfies Equation (11). Indeed, 
\[
2 \left( \frac{\lambda_1}{n^2} + \lambda_2 \right) (n^{opt}C_1 + C_2) = 2\lambda_1 C_1 + 2\lambda_2 C_2 + 4\sqrt{\lambda_1 \lambda_2 C_1 C_2} = (\sqrt{2}\lambda_1 C_1 + \sqrt{2}\lambda_1 C_1)^2.
\]
In practice, since the number of segments can only be a positive integer, the optimal solution is either \( \max(1, [n^{opt}]) \) or \( [n^{opt}] \), whichever leads to a smaller value of the convex function \( H \) as shown in Equation (16).

Remarks. Consider the example given in [8] with \( C_1 = R_1 = 20, C_2 = R_2 = 50, \lambda_1 = 2.78 \times 10^{-4} \) and \( \lambda_2 = 4.63 \times 10^{-5} \). The optimal solution\(^3\) provided by [8] gives \( n^{opt} = 3.83, \)
\( W^{opt} = 1302.49 \) and \( H^{opt} = 0.1879 \), while Theorem 1 suggests \( n^{opt} = 3.87, W^{opt} = 1378.27 \) and \( H^{opt} = 0.1735 \), which is quite close to the exact optimum. The difference in overhead is due to the negligence of lower-order terms in the first-order approximation. We point out that the solution provided by [8] relies on numerical methods to solve rather complex mathematical equations, whose convergence is not always guaranteed, and it is only applicable to two levels. Our result, on the other hand, is able to provide fast and good approximation to the optimal solution when the error rates are sufficiently small, and it can be readily extended to an arbitrary number of levels, as shown in the next section.

3.3 Optimal \( k \)-level pattern

In this section, we derive the first-order approximation to the optimal \( k \)-level pattern by determining its length \( W \), the number \( N_\ell \) of level-\( \ell \) checkpoints for all \( 1 \leq \ell \leq k \), as well as the positions of all checkpoints in the pattern.

3.3.1 Observations

Before analyzing the optimal pattern, we make several observations. First, we can obtain the orders of the optimal length and pattern overhead as shown below (recall that \( \Lambda = \sum_{\ell=1}^{k} \lambda_{\ell} \)).

Observation 1. Consider the simplest \( k \)-level pattern with a single segment of length \( W \). We can conduct the same analysis as in Section 3.2.1 to show that the optimal pattern length satisfies \( W^{opt} = \Theta(\Lambda^{-1/2}) \), and the corresponding overhead satisfies \( H^{opt} = \Theta(\Lambda^{1/2}) \).

From the analysis of the two-level pattern, we can also observe that the overall execution overhead of any pattern comes from two distinct sources defined below.

Observation 2. There are two types of execution overheads for a pattern:

(1) Error-free overhead, denoted as \( o_{ef} \), is the total cost of all the checkpoints placed in the pattern. For a given set of checkpoints, the error-free overhead is completely determined regardless of their positions in the pattern.

(2) Re-executed fraction overhead, denoted as \( o_{re} \), is the expected fraction of work that needs to be re-executed due to errors. The re-executed fraction overhead depends on both the set of checkpoints and their positions.

For example, in the two-level pattern with \( n \) level-1 checkpoints and given values of \( \alpha_i \) for all \( i = 1, 2, \ldots, n \), the two types of overheads are given by
\[
o_{ef} = nC_1 + C_2 \quad \text{and} \quad o_{re} = \frac{1}{2} \left( f_1 \sum_{i=1}^{n} \alpha_i^2 + f_2 \right), \]
where \( f_1 = \frac{\Lambda}{\lambda} \) for \( \ell = 1, 2 \). Assuming that checkpoints at all levels have constant costs and that the error rates at all levels are in the same order, then both \( o_{ef} \) and \( o_{re} \) can be considered as constants, i.e., \( o_{ef} = O(1) \) and \( o_{re} = O(1) \).

A trade-off exists between these two types of execution overheads, since placing more checkpoints generally reduces the re-executed work fraction when an error strikes, but it can adversely increase the overhead when the execution is error-free. Therefore, in order to achieve the best overall overhead, a resilience algorithm must seek an optimal balance between \( o_{ef} \) and \( o_{re} \).

For a given pattern with fixed overheads \( o_{ef} \) and \( o_{re} \), we can make the following observation based on Propositions 1 and 2, which partially characterizes the optimal pattern.

Observation 3. For a given pattern (with fixed \( o_{ef} \) and \( o_{re} \)), the expected execution time is given by
\[
E = W + o_{ef} + \frac{AW}{\text{expected } \# \text{ errors}} \cdot o_{re} W + O(\Lambda^{1/2}), \tag{17}
\]
and the optimal pattern length and the resulting expected execution overhead of the pattern are
\[
W^{opt} = \sqrt{o_{ef}} \cdot o_{re}, \tag{18}
\]
\[
H^{opt} = 2\sqrt{\Lambda \cdot o_{ef} \cdot o_{re}} + O(\Lambda). \tag{19}
\]

Equation (19) shows that the trade-off between \( o_{ef} \) and \( o_{re} \) is manifested as the product of the two terms. Hence, in order to determine the optimal pattern, it suffices to find the pattern parameters (e.g., \( n \) and \( \alpha_i \)) that minimize \( o_{ef} \cdot o_{re} \).

3.3.2 Analysis

We now extend the analysis to derive the optimal multi-level checkpointing patterns. Generally, for a \( k \)-level pattern, each computational segment \( s^{(c)}_{i_{k-1}, \ldots, i_{1}} \) can be uniquely identified by its level \( \ell \) as well as its position \( i_{k-1}, \ldots, i_{1} \) within the multi-level hierarchy. For instance, in a four-level pattern, the segment \( s^{(2)}_{1,3} \) denotes the third level-2 segment inside the first level-3 segment of the pattern (see Figure 2). Note that a segment can contain multiple sub-segments at the lower levels (except...
for bottom-level segments) and is a sub-segment of a larger segment at a higher level (except for top-level segments). The entire pattern can be denoted as \( s^{(k)} \), which is the only segment at level \( k \).

For any segment \( s^{(\ell)}_{i_\ell,\ldots,i_\ell} \) at level \( \ell \), where \( 1 \leq \ell \leq k \), let \( w^{(\ell)}_{i_\ell,\ldots,i_\ell} \) denote its length. Hence, we have \( w^{(\ell+1)}_{i_\ell+1,\ldots,i_\ell+1} = \sum_{\ell} w^{(\ell)}_{i_\ell,\ldots,i_\ell} \) and \( w^{(k)} = W \). Also, let \( n^{(\ell)}_{i_\ell,\ldots,i_\ell} \) denote the number of sub-segments contained by \( s^{(\ell)}_{i_\ell,\ldots,i_\ell} \) at the lower level \( \ell - 1 \). We have \( n^{(1)}_{i_1,\ldots,i_1} = 1 \) for all \( i_1,\ldots,i_1 \). For convenience, we further define

\[
\alpha^{(\ell)}_{i_\ell,\ldots,i_\ell} = \frac{w^{(\ell)}_{i_\ell,\ldots,i_\ell} \times R}{W} 
\]

as the fraction of the length of segment \( s^{(\ell)}_{i_\ell,\ldots,i_\ell} \) inside the pattern, and define \( N_\ell \) to be the total number of level-\( \ell \) segments in the entire pattern. Therefore, we have \( N_\ell = 1 \), \( N_{\ell-1} = n^{(k)} \), and in general

\[
N_\ell = \sum_{i_\ell \in \ldots \in i_{\ell+1}} n^{(\ell+1)}_{i_{\ell+1},\ldots,i_{\ell+1}}. 
\]

The following proposition shows the expected time to execute a given k-level pattern.

**Proposition 3.** The expected execution time of a given k-level pattern is

\[
E = W + \sum_{\ell=1}^{k-1} N_\ell C_\ell + C_k 
+ \frac{W^2}{2} \left( \sum_{\ell=1}^{k} \sum_{i_\ell \in \ldots \in i_{\ell+1}} \frac{\left( \sum_{i_\ell \in \ldots \in i_{\ell+1}} \alpha^{(\ell)}_{i_\ell,\ldots,i_{\ell+1}} \right)^2}{\lambda_\ell} \right) + O(\Lambda^{1/2}).
\]

**Proof.** We show that the expected time to execute any segment \( s^{(h)}_{i_h,\ldots,i_h} \) at level \( h \), where \( 1 \leq h \leq k \), satisfies the following (without counting the time to execute all the checkpoints inside the segment):

\[
E^{(h)}_{i_h,\ldots,i_h} = w^{(h)}_{i_h,\ldots,i_h} + \frac{W^2}{2} \left( \sum_{i_\ell \in \ldots \in i_{\ell+1}} \alpha^{(\ell)}_{i_\ell,\ldots,i_{\ell+1}} \right)^2 + \Lambda_{[h+1,k]} \left( \frac{w^{(h)}_{i_h,\ldots,i_h}}{2} + \sum_{i_{h+1} \in \ldots \in i_{h+2}} E^{(h+1)}_{i_{h+1},\ldots,i_{h+2}} \right) + \sum_{i_{h+1} \in \ldots \in i_{h+2}} \sum_{j_{h+1} \in \ldots \in j_{h+2}} \frac{E^{(h)}_{i_{h+1},\ldots,i_{h+2}}}{2} + O(\Lambda^{1/2}),
\]

which satisfies Equation (20).

Suppose Equation (20) holds up to any segment \( s^{(h)}_{i_h,\ldots,i_h} \) at level \( h \). Following the proof of Proposition 2 (in particular, the derivation of Equation (10)), we can show by induction that

\[
E^{(h+1)}_{i_{h+1},\ldots,i_{h+1}} = w^{(h+1)}_{i_{h+1},\ldots,i_{h+1}} + O(1). 
\]

Hence, for segment \( s^{(h+1)}_{i_{h+1},\ldots,i_{h+1}} \) at level \( h + 1 \), we have:

\[
E^{(h+1)}_{i_{h+1},\ldots,i_{h+1}} = \sum_{i_{h+1} \in \ldots \in i_{h+1}} E^{(h)}_{i_{h},\ldots,i_{h}} 
+ \sum_{i_{h+1} \in \ldots \in i_{h+1}} w^{(h)}_{i_{h+1},\ldots,i_{h+1}} \sum_{i_{h+1} \in \ldots \in i_{h+1}} \left( \frac{E^{(h)}_{i_{h},\ldots,i_{h}}}{2} + \sum_{i_{h+1} \in \ldots \in i_{h+1}} \sum_{i_{h+1} \in \ldots \in i_{h+1}} \sum_{j_{h+1} \in \ldots \in j_{h+1}} \frac{E^{(h)}_{i_{h+1},\ldots,i_{h+1}}}{2} \right) + O(\Lambda^{1/2})
\]

where \( \Lambda_{[x,y]} = \sum_{\ell=x}^{y} \alpha^{(\ell)}_{i_\ell,\ldots,i_\ell} \). The proposition can then be proven by setting \( E = E^{(h+1)} \) and \( \sum_{\ell=1}^{k-1} N_\ell C_\ell + C_k \), since checkpoints are assumed to be error-free.
with the re-executed work overhead is given by

\[ o = \sum_{\ell=1}^{k} \sum_{j=1}^{i_{\ell}-1} E_j + O(\Lambda^{1/2}) \]

and it is achieved when its level-\( \ell \) checkpoints are equally spaced, for all \( \ell \leq h - 1 \). The proposition can then be proven by setting \( o_{\text{opt}} = o_{\text{opt}}(s(\ell)) \), since \( N_1(s(\ell)) = N_\ell, N_k = 1, \) and \( \alpha(\ell) = 1 \).

Now, we prove Equation (25) by induction on the level \( h \). For the base case, i.e., when \( h = 1 \), we have \( o_{\text{opt}}(s(1)) = \frac{f_1}{2} \cdot (\alpha(1)) \). Then, for segment \( s(\ell+1) \) at level \( h+1 \), we have:

\[ o_{\text{opt}}(s(\ell+1)) = \frac{f_{\ell+1}}{2} \cdot (\alpha(\ell+1)) \]

subject to \( \sum_{i_h} (\alpha(\ell)) \).

Since \( y \) is clearly a convex function of \( \alpha(\ell) \), we can readily get, using Lagrange multiplier [5], the minimum value of \( y \) as follows:

\[ y_{\min} = \frac{1}{\sum_{i_h} 1/x(\ell)} \cdot (\alpha(\ell)) \]

which is obtained at

\[ \alpha(h) = \frac{1}{\sum_{i_h} 1/x(\ell)} \cdot (\alpha(\ell)) \]

Let us define \( z = \sum_{i_h} 1/x(\ell) \). We now need to solve the following maximization problem:

\[ \text{maximize} \quad z = \sum_{i_h} \frac{1}{\sum_{\ell=1}^{N_1(s(\ell))}} \]

subject to \( \sum_{i_h} N_\ell(s(\ell)) = N_\ell(s(\ell)) \), \( \forall \ell = 1, \ldots, h \). Again, \( z \) is a convex function of \( N_\ell(s(\ell)) \), and it can be
shown to be maximized when

\[ N_\ell \left( s_{i_{k-1}, \ldots, i_h}^{(h)} \right) = \frac{N_\ell \left( s_{i_{k-1}, \ldots, i_{h+1}}^{(h+1)} \right)}{N_{i_{k-1}, \ldots, i_{h+1}}}, \quad \forall \ell = 1, \ldots, h, \]

which gives \( \tilde{\alpha}_{i_{k-1}, \ldots, i_h}^{(h)} = \frac{1}{n_{i_{k-1}, \ldots, i_{h+1}}} \cdot \alpha_{i_{k-1}, \ldots, i_{h+1}}^{(h+1)} \) according to Equation (28). This implies that all level-\( \ell \) checkpoints are also equally spaced inside segment \( s_{i_{k-1}, \ldots, i_{h+1}} \), for all \( \ell \leq h \). The maximum value of \( z \) in this case is

\[ z_{\text{max}} = \frac{1}{\sum_{\ell=1}^{h} \frac{n_{i_{k-1}, \ldots, i_{h+1}}}{N_\ell \left( s_{i_{k-1}, \ldots, i_{h+1}}^{(h+1)} \right)}}, \]

and the optimal value of \( y_{\text{min}} \) according to Equation (27) is then given by

\[ y_{\text{opt}} = \frac{1}{z_{\text{max}}} \left( \alpha_{i_{k-1}, \ldots, i_h}^{(h+1)} \right)^2 = \left( \sum_{\ell=1}^{h} \frac{f_\ell}{N_\ell \left( s_{i_{k-1}, \ldots, i_{h+1}}^{(h+1)} \right)} \right) \left( \alpha_{i_{k-1}, \ldots, i_h}^{(h+1)} \right)^2. \]

Substituting \( y_{\text{opt}} \) into Equation (26), we get the optimal value of \( \alpha_{\text{re}} \left( s_{i_{k-1}, \ldots, i_{h+1}}^{(h+1)} \right) \) as follows:

\[ \alpha_{\text{re}}^{\text{opt}} \left( s_{i_{k-1}, \ldots, i_{h+1}}^{(h+1)} \right) = \frac{f_{h+1}}{2} \cdot \left( \alpha_{i_{k-1}, \ldots, i_h}^{(h+1)} \right)^2 + \frac{1}{2} y_{\text{opt}} \]

\[ = \frac{1}{2} \left( \sum_{\ell=1}^{h} \frac{f_\ell}{N_\ell \left( s_{i_{k-1}, \ldots, i_{h+1}}^{(h+1)} \right)} \right) \left( \alpha_{i_{k-1}, \ldots, i_h}^{(h+1)} \right)^2. \]

This shows that Equation (25) also holds for segment \( s_{i_{k-1}, \ldots, i_{h+1}}^{(h+1)} \) at level \( h+1 \) and, hence, completes the proof of the proposition.

We are now ready to characterize the optimal \( k \)-level pattern. The result is stated in the following theorem.

**Theorem 2.** A first-order approximation to the optimal \( k \)-level pattern and its overhead are characterized by

\[ W_{\text{opt}} = \sqrt{2 \left( \sum_{\ell=1}^{k-1} N_\ell^{\text{opt}} C_\ell + C_k \right) / \sum_{\ell=1}^{k-1} \frac{N_\ell^{\text{opt}}}{N_\ell^{\text{opt}}}}, \quad \forall \ell = 1, 2, \ldots, k - 1, \]

\[ N_\ell^{\text{opt}} = \frac{\lambda_\ell C_\ell}{\lambda_k}, \quad \forall \ell = 1, 2, \ldots, k - 1, \]

\[ H_{\text{opt}} = \sum_{\ell=1}^{k} \sqrt{2 \lambda_\ell C_\ell} + O(\Lambda). \]

**Proof.** From Observation 3, Equation (22) and Proposition 4, we know that the optimal pattern can be obtained by minimizing the following function:

\[ F = \alpha_{\text{ef}} \cdot \alpha_{\text{re}}^{\text{opt}} = \frac{1}{2} \left( \sum_{\ell=1}^{k-1} N_\ell C_\ell + C_k \right) \left( \sum_{\ell=1}^{k-1} \frac{f_\ell}{N_\ell^{\text{opt}}} + f_k \right). \]

We first compute the optimal number of checkpoints at each level using a two-phase iterative method. Towards this end, let us define

\[ \alpha_{\text{ef}}(h) = \sum_{\ell=h}^{k-1} N_\ell C_\ell + C_k, \]

\[ \alpha_{\text{re}}^{\text{opt}}(h) = \frac{1}{2} \left( \sum_{\ell=h}^{k-1} \frac{f_\ell}{N_\ell^{\text{opt}}} + f_k \right). \]

In the first phase, we set initially

\[ F(1) = \alpha_{\text{ef}}(1) \cdot \alpha_{\text{re}}^{\text{opt}}(1). \]

The optimal value of \( N_1 \) that minimizes \( F(1) \) can then be obtained by setting

\[ \frac{\partial F(1)}{\partial N_1} = C_1 \alpha_{\text{re}}^{\text{opt}}(1) - \alpha_{\text{ef}}(1) \cdot \frac{f_1}{2 N_1^2} = 0, \]

which gives \( N_1^{\text{opt}} = \frac{f_1}{C_1}. \)

Substituting it into \( F(1) \) and simplifying, we can get the value of \( F \) after the first iteration as

\[ F(2) = \frac{1}{2} \left( \sqrt{f_1 C_1^{\text{opt}}} + \sqrt{\alpha_{\text{re}}^{\text{opt}}(2) \cdot \alpha_{\text{re}}^{\text{opt}}(2)} \right)^2. \]

Repeating the above process, we can get the optimal value of \( F \) after \( k - 1 \) iterations as

\[ F_{\text{opt}} = F(k) = \frac{1}{2} \left( \sum_{\ell=1}^{k} \sqrt{f_\ell C_\ell} \right)^2, \]

and the optimal value of \( N_\ell \) as

\[ N_\ell^{\text{opt}} = \frac{f_\ell}{C_\ell} \cdot \frac{\alpha_{\text{ef}}(\ell + 1)}{2 \alpha_{\text{re}}^{\text{opt}}(\ell + 1)} \cdot \lambda_\ell, \quad \forall \ell = 1, 2, \ldots, k - 1. \]

In the second phase, we first compute from Equation (33)

\[ N_{k-1}^{\text{opt}} = \frac{f_{k-1}}{C_{k-1}} \cdot \frac{\alpha_{\text{ef}}(k)}{2 \alpha_{\text{re}}^{\text{opt}}(k)}, \]

\[ = \frac{f_{k-1}}{C_{k-1}} \cdot \frac{C_k}{f_k} = \frac{\lambda_{k-1}}{C_{k-1}} \cdot \frac{C_k}{\lambda_k}. \]

Substituting it into \( N_{k-2}^{\text{opt}} \), we obtain:

\[ N_{k-2}^{\text{opt}} = \frac{f_{k-2}}{C_{k-2}} \cdot \frac{\alpha_{\text{ef}}(k-1)}{2 \alpha_{\text{re}}^{\text{opt}}(k-1)} \]

\[ = \frac{\lambda_{k-2}}{C_{k-2}} \cdot \frac{N_{k-1}^{\text{opt}} C_{k-1} + C_k}{\lambda_{k-1} C_{k-1} + \lambda_k} \]

\[ = \frac{\lambda_{k-2}}{C_{k-2}} \cdot \frac{\lambda_{k-1} C_{k-1} + C_k}{\lambda_{k-1} \lambda_k C_{k-1} C_{k-2} + \lambda_k}. \]
For example, with \( k \) levels, we determine the optimal subset, we identify the following two subsets are possible: \( \{1, 2\} \) and \( \{2\} \). In the fixed independent cost model, these costs will stay unchanged regardless of the subset chosen. In the incremental cost model, since \( C_2 \) is the additional cost paid after \( C_1 \) is done, when using subset \( \{2\} \), i.e., only placing level-2 checkpoints in the pattern, we need to adjust its cost as \( C'_2 = 10 + 20 = 30 \). In both cases, once the subset is decided, the checkpoint costs at the selected levels can be computed and therefore considered as fixed costs. The theoretical analysis presented in Section 3.3 can then be used to compute the optimal pattern.

But how to determine the optimal subset of levels? Consider again the example with \( k = 2 \) levels. In the incremental cost model, Equation (31) suggests that the optimal solution (ignoring lower-order terms) uses both levels if and only if

\[
\sqrt{2\lambda_1 C_1} + \sqrt{2\lambda_2 C_2} \leq \sqrt{2(\lambda_1 + \lambda_2)(C_1 + C_2)}
\]

which is always true when assuming \( \lambda_1 \geq \lambda_2 \) and \( C_1 \leq C_2 \). We can easily apply the same argument to show that the optimal subset must contain all levels available as long as all checkpoint costs are positive.

In the fixed independent cost model, however, it is not clear whether all available levels should be used. Consider the same example with \( k = 2 \) levels, and define \( \alpha = \frac{\lambda_1}{\lambda_2} \) and \( \beta = \frac{C_2}{C_1} \). The optimal solution uses both levels if and only if

\[
\sqrt{2\lambda_1 C_1} + \sqrt{2\lambda_2 C_2} \leq \sqrt{2(\lambda_1 + \lambda_2)(C_1 + C_2)}
\]

which is not true when \( \alpha = 0.5 \) and \( \beta = 2 \). In this case, using only level-2 checkpoints leads to a smaller overhead.

### 3.4 Optimal subset of levels

The preceding section characterizes the optimal pattern by using \( k \) levels of checkpoints. In many practical cases, there is no obligation to use all available levels. This section addresses the problem of selecting the optimal subset of levels in order to minimize the overall execution overhead.

#### 3.4.1 Checkpoint cost models

So far, we have assumed that all the checkpoint costs are fixed under a multi-level checkpointing scheme. In practice, the checkpoint costs may vary depending upon the implementation, and upon the subset of selected levels. In order to determine the optimal subset, we identify the following two checkpoint cost models:

- **Fixed independent costs.** The checkpoint cost \( C_\ell \) at level \( \ell \) is the cost paid to save data at level \( \ell \), independently of the subset of levels used. In this model, the checkpoint costs stay the same for all possible subsets.

- **Incremental costs.** The checkpointing cost \( C_\ell \) at level \( \ell \) is the additional cost paid to save data when going from level \( \ell - 1 \) to \( \ell \). In this model, the checkpoint cost at a particular level depends on the subset of levels selected.

For example, with \( k = 2 \) levels and \( C_1 = 10, C_2 = 20 \), two subsets are possible: \( \{1, 2\} \) and \( \{2\} \). In the fixed independent cost model, these costs will stay unchanged regardless of the subset chosen. In the incremental cost model, since \( C_2 \) is the additional cost paid after \( C_1 \) is done, when using subset \( \{2\} \), i.e., only placing level-2 checkpoints in the pattern, we need to adjust its cost as \( C'_2 = 10 + 20 = 30 \). In both cases, once the subset is decided, the checkpoint costs at the selected levels can be computed and therefore considered as fixed costs. The theoretical analysis presented in Section 3.3 can then be used to compute the optimal pattern.

![Image](image-url)
4 Simulations

In this section, we conduct a set of simulations whose goal is threefold: (i) to verify the accuracy of the first-order approximation; (ii) to confirm the optimality of the subset of levels found by the dynamic programming algorithm; and (iii) to evaluate the performance of our approach and to compare it with other multi-level checkpointing algorithms. After introducing the simulation setup in Section 4.1, we proceed in two steps. First, in Section 4.2, we instantiate the model with realistic parameters from the literature and run simulations for all possible subsets of levels and roundings. Then, in Section 4.3, we instantiate the model with different test cases from the recent work of Di et al. [7], [8] on multilevel checkpointing and compare the overheads obtained with three approaches: (a) Young/Daly’s classical formula; (b) our first-order approximation formula; and (c) Di et al.’s iterative/optimal algorithm. The simulator code is publicly available at http://perso.ens-lyon.fr/aurelien.cavelan/multilevel.zip, so that interested readers can experiment with it and instantiate the model with parameters of their own choice.

4.1 Simulation setup

Checkpoint and recovery costs both depend on the volume of data to be saved, and are mostly determined by the hardware resource used at each level. As such, we assume that recovery cost for a given level is equivalent to the corresponding checkpointing cost, i.e., \( R_\ell = C_\ell \) for \( 1 \leq \ell \leq k \) (unless specified otherwise). This a common assumption [14], [7], even though in practice the recovery cost can be expected to be somewhat smaller than the checkpoint cost [7], [8]. All costs are fixed and independent (as discussed in Section 3.4.1).

The simulator is fed with \( k \) levels of errors and their MTBFs \( \mu_\ell = 1/\lambda_\ell \) as well as the resilience parameters \( C_\ell \) and \( R_\ell \). For each of the \( 2^{k-1} \) possible subsets of levels (the last level is always included), we take the optimal pattern given in Theorem 2 and Corollary 1, and then try all possible roundings (floor and ceiling) based on the optimal (rational) number of checkpoints \( n_\ell^{\text{opt}} \) given in Equation (34). For each rounding, we compare the following three overheads:

- **Simulated overhead**, obtained by running the simulation 10000 times and averaging the results;
- **Corresponding theoretical overhead**, obtained from Equations (19), (22) and (24) using the integer solution that corresponds to the rounding;
- **Theoretical lower bound**, obtained from Equation (31) with the optimal rational solution.

In the following, we associate Young/Daly’s classical formula, defined as \( W_{\text{opt}} = \sqrt{\frac{2C}{\mu_k}} \), with the highest checkpointing level available, i.e., \( C = C_k \). Note that in this case, Young/Daly’s formula and Equation (29) can be used interchangeably, and the corresponding theoretical overhead is obtained with \( H_{\text{opt}} = \sqrt{2kC} \).

4.2 Assessing accuracy of first-order approximation

In this section, we run simulations with two sets of parameters, described in Table 1. For each set of parameters, we consider all possible subsets of levels. Then, for each subset, we compute the optimal pattern length and number of checkpoints to be used at each level. We show the accuracy of our approach in both scenarios, and we confirm the optimality of the subset of levels returned by the dynamic programming algorithm.

<table>
<thead>
<tr>
<th>Set</th>
<th>From</th>
<th>Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>Moody et al. [14]</td>
<td>MTBF</td>
<td>5.006</td>
<td>5.566</td>
<td>2.506</td>
<td>-</td>
</tr>
<tr>
<td>(B)</td>
<td>Balaprakash et al. [1]</td>
<td>MTBF</td>
<td>3.006</td>
<td>7.206</td>
<td>1.436</td>
<td>7.206</td>
</tr>
</tbody>
</table>

4.2.1 Using set of parameters (A)

The first set of parameters (shown in set (A) of Table 1) corresponds to the Coastal platform, a medium-sized HPC system of 1104 nodes at the Lawrence Livermore National Laboratory (LLNL). The Coastal platform has been used to evaluate the Scalable Checkpoint/Restart (SCR) library by Moody et al. [14], who provided accurate measurements for the checkpoint costs using real applications (given in the first row of Table 1). There are \( k = 3 \) levels of checkpoints. First-level checkpoints are written to the local RAMs of the nodes, and this is the fastest method (0.5 s). Second-level checkpoints are also written to local RAMs, but small sets of nodes collectively compute and store parity redundancy data, which takes a little while longer (4.5 s). Lastly, Lustre is used to store third-level checkpoints onto the parallel file system, which takes significantly longer time (1051 s). Failures were analyzed in [14], and the error rates are given in the second row of Table 1. Note that the error rate at level 2 is higher than those of levels 1 and 3.

**Results:** Table 2 and Figure 3 present the simulation results. Table 2 shows, from left to right, the subset of levels used, the number of checkpoints computed by our first-order approximation formula for each possible rounding \( (N_1, N_2, N_3) \), the corresponding optimal pattern length \( (W_{\text{opt}}) \), the simulated overhead (Sim. Ov.), the corresponding theoretical overhead (Th. Ov.), the absolute and relative differences of these two overheads \( (\text{Diff.} = \text{Sim. Ov.} - \text{Th. Ov.}) \), and finally the theoretical lower bound for this subset (Th. L.B.).

With \( k = 3 \), there are four possible subsets of levels, and both the best simulated overhead and the corresponding theoretical overhead are achieved for the subset \{2, 3\}, with \( N_2 = 35 \) and \( N_3 = 1 \) (highlighted in bold in the table). First, the difference between the simulated and theoretical overheads is very small, with a difference < 0.7% in absolute values, and a relative difference ranging from 2.9% (for subset \{1, 2, 3\}) to 8.14% (for subset \{3\}), which shows the accuracy of the first-order approximation for this set of parameters. The simulated overhead is always higher than the theoretical one, which is expected, because the first-order approximation ignores some lower-order terms. Next, we observe that, for each subset, all roundings of the number of checkpoints yield similar overheads on this platform, and the difference between the best and worst roundings is almost negligible.

Furthermore, using the best subset \{2, 3\} improves the overhead by over 50% compared to using level-3 checkpoints alone (as in Young/Daly’s result). This is indeed the subset returned by the dynamic programming algorithm, and the result matches closely the minimum theoretical lower bound. Finally, comparing our result to the one obtained by the optimal two-level algorithm by Di et al. [8] on this best subset, we see that the simulated overheads are similar under the optimal subset, as the patterns found using both approaches share the same number of checkpoints and the pattern lengths are also almost identical.
4.2.2 Using set of parameters (B)

The second set of parameters correspond to the execution of the LAMMPS application on the large BG/Q platform Mira at the Argonne National Laboratory (ANL) [1]. The parameters are presented in set (B) of Table 1. In this setting, the Fault Tolerance Interface (FTI) [2] was used, which has four checkpoint levels \( (k = 4) \): Local checkpoint; Local checkpoint + Partner-copy; Local checkpoint + Reed-Solomon coding; and PFS-based checkpoint. The MTBFs correspond to the failure rates typically observed for petascale HPC applications [2], [14], [7].

Table 3: Simulation results using set of parameters (B).

<table>
<thead>
<tr>
<th>Levels</th>
<th>( N_1 ), ( N_2 ), ( N_3 ), ( N_4 )</th>
<th>( W^{\text{opt}} )</th>
<th>Sim. Over.</th>
<th>Th. Over.</th>
<th>Abs. Diff.</th>
<th>Rel. Diff.</th>
<th>Th. L.B.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[4]</td>
<td>- - - -</td>
<td>2.43e-02</td>
<td>1.43e-1</td>
<td>1.22e-1</td>
<td>1.9%</td>
<td>13.3%</td>
<td>7.22e-1</td>
</tr>
<tr>
<td>[1,4]</td>
<td>- - - 1</td>
<td>3.13e-01</td>
<td>1.18e-1</td>
<td>1.05e-1</td>
<td>1.3%</td>
<td>11.0%</td>
<td>1.05e-1</td>
</tr>
<tr>
<td>[2,4]</td>
<td>- - 5 -</td>
<td>6.00e-03</td>
<td>1.11e-1</td>
<td>1.00e-1</td>
<td>1.1%</td>
<td>9.9%</td>
<td>1.00e-1</td>
</tr>
<tr>
<td>[3,4]</td>
<td>- - - 11</td>
<td>1.56e-01</td>
<td>9.96e-2</td>
<td>9.02e-2</td>
<td>0.94%</td>
<td>9.44%</td>
<td>9.01e-2</td>
</tr>
<tr>
<td>[1,2,4]</td>
<td>- - - 10</td>
<td>1.44e-01</td>
<td>9.01e-2</td>
<td>9.01e-2</td>
<td>0.90%</td>
<td>9.08%</td>
<td>1.02e-1</td>
</tr>
<tr>
<td>[1,3,4]</td>
<td>- - 18 -</td>
<td>1.04e-01</td>
<td>9.95e-2</td>
<td>9.95e-2</td>
<td>0.89%</td>
<td>9.92%</td>
<td>8.96e-2</td>
</tr>
<tr>
<td>[2,3,4]</td>
<td>- - - 12</td>
<td>1.36e-01</td>
<td>1.04e-1</td>
<td>9.73e-2</td>
<td>0.67%</td>
<td>6.4%</td>
<td>0.68e-2</td>
</tr>
<tr>
<td>[1,2,3,4]</td>
<td>- - - 9</td>
<td>1.11e-01</td>
<td>1.05e-1</td>
<td>9.73e-2</td>
<td>0.75%</td>
<td>7.1%</td>
<td>0.92e-2</td>
</tr>
</tbody>
</table>

Figure 4: Simulated and (corresponding) theoretical overheads for all possible subsets of levels with the best and worst roundings for each subset using set of parameters (B).

Results: Table 3 and Figure 4 present the simulation results for this set of parameters. There are 8 possible subsets of levels. As before, we observe that the theoretical overhead is always slightly smaller than the simulated one, with an absolute difference of less than 2%, and a relative difference between 6-14%, demonstrating the accuracy of the model. Again, the results are very close to the theoretical lower bound. For this platform, the simulated overheads vary from 9.68% (with optimal subset of levels \( \{1, 3, 4\} \)) found by the dynamic programming algorithm to 14.3% (with level-4 checkpoints alone). For a given subset of levels, the rounding does not play a significant role, as \( W^{\text{opt}} \) is also adjusted accordingly (increased or decreased) as a result of rounding. For instance, we observe that, for subset \( \{1, 2, 3, 4\} \), the number of checkpoints at levels 1 and 2 are halved for the third rounding compared to the first rounding in Table 3, but \( W^{\text{opt}} \) is also reduced by 31%, so that for the same amount of work, the number of checkpoints does not change by much. We can also see that the pattern length \( W^{\text{opt}} \) for the smallest overhead is around 10400s, but only 2450s for the largest overhead. In fact, the largest pattern lengths are obtained for the highest *cumulated checkpoint cost*, which turns out to be 830s for \( \{2, 3, 4\} \) with \( N_2 = 16, N_3 = 4, N_4 = 1 \), and for \( \{1, 2, 3, 4\} \) with \( N_1 = 24, N_2 = 8, N_3 = 4 \) and \( N_4 = 1 \). This is because using more checkpoints both increases the error-free overhead and reduces the time lost due to re-executions upon errors. As a consequence, and to mitigate the aforementioned overhead, the length of the pattern increases for \( \{2, 3, 4\} \) and \( W^{\text{opt}} = 16600s \) for \( \{1, 2, 3, 4\} \). And the converse is also true: when using fewer checkpoints, the error-free overhead decreases and the time lost upon errors increases. In order to compensate, the pattern length decreases for \( \{1, 2, 3, 4\} \) with \( N_1 = 6, N_2 = 3, N_3 = 3 \) and \( N_4 = 1 \).

We note that, in this case, our first-order solution slightly outperforms the iterative method by Di et al. [7] on multi-level checkpointing (with a simulated overhead of 9.68e-2 compared to 9.75e-2). The reason is that their algorithm computes a solution under the independent checkpointing model, i.e., checkpoints at different levels are taken according to different independent periods. However, it is not clear how such a model can be implemented in practice due to the difficulties as explained in Section 1 and the different options to rollback to a checkpoint in case of a fault. Therefore, we transformed their result to a pattern-based solution by rounding the differ-
We first run simulations for Cases #A and #B, whose parameters are presented in Table 4. These parameters are based on the FTI multilevel checkpointing model and have been used by Di et al. [7] to evaluate the performance of their approach. Note that the recovery cost is about half that of the checkpointing cost in Case #B.

In their work, Di et al. considered independent checkpointing periods, as opposed to the nested method based on periodic patterns (as discussed in Section 1). Although they provided an optimal solution, an iterative approach was used to compute it numerically in contrast to the simple formula we propose in this paper. Recall that using independent checkpointing periods allows checkpoints at different levels to be taken simultaneously, which can hardly be done in practice. Adapting their solution to our model results in rational numbers of checkpoints, and we again use rounding to resolve this issue. We find that, using the best roundings for both approaches, their solution turns out to be very similar to ours (with the same number of checkpoints, and close periods with < 1% difference).

**Results:** Figure 5 presents the overheads for both cases. First, we observe that Di et al.’s optimal iterative algorithm has almost identical performance to our solution, with a simulated overhead around 45% for Case #A and 140% for Case #B under both approaches. However, using Young/Daly’s formula to checkpoint only at the highest level yields significantly worse overheads (around 90% for Case #A and 170% for Case #B). Overall, our solution is as good as Di et al.’s optimal numerical one (but has much less complexity), and it is up to 45% better than Young/Daly’s formula in Case #A and 30% better in Case #B.

Note that the corresponding theoretical overhead of our solution is close to the simulated one for Case #A, but starts to diverge for Case #B. This is because first-order approximation is only accurate when the resilience parameters and pattern length are small compared to the MTBF, which is no longer true for Case #B. Specifically, we have:

- In Case #A, the optimal subset of levels is \{2, 4\}. The optimal pattern has length \(W^{opt} = 1052s\) and consists of \(N_2 = 8\) level-2 checkpoints followed by \(N_1 = 1\) level-4 checkpoint, meaning that we have a level-2 checkpoint every 131.5s of computation. So a level-2 checkpoint is saved every 141.5s and a level-4 checkpoint is saved every 1222s. On the other hand, the combined MTBF for errors at levels 1 and 2 (handled by level-2 checkpoints) is 864s and the combined MTBF for errors at levels 3 and 4 (handled by level-4 checkpoints) is 6171s. Hence, we have \(\frac{141.5}{864} = 0.164\) and \(\frac{1222}{6171} = 0.198\), which are reasonably small, making our solution accurate.

- In Case #B, the optimal subset of levels is \{1, 4\}, and the optimal pattern has \(W^{opt} = 223s\), \(N_1 = 5\) and \(N_2 = 1\). Thus, we have a level-1 checkpoint every 44.6s of computation. So a level-1 checkpoint is saved every 45.6s and a level-4 checkpoint is saved every 298s. The MTBF for errors at level 1 is 864s and the combined MTBF for errors at levels 2, 3 and 4 (handled by level-4 checkpoints) is 360s. Thus, we have \(\frac{44.6}{864} = 0.052\), which is fine, but \(\frac{298}{360} = 0.828\), which is too high and essentially makes the first-order solution inaccurate.

Despite the difference between the theoretical and simulated overheads under Case #B, the proximity of our solution to Di et al.’s optimal numerical solution nevertheless shows the usefulness of first-order approximation for determining the optimal multi-level checkpointing patterns.

### Table 4: Set of parameters (C) used as input for simulations.

<table>
<thead>
<tr>
<th>Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case #A</td>
<td>C (s)</td>
<td>10</td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>Case #A</td>
<td>R (s)</td>
<td>10</td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>Case #B</td>
<td>MTBF (s)</td>
<td>1440</td>
<td>864</td>
<td>2160</td>
</tr>
</tbody>
</table>

### Table 5: Set of parameters (D) used as input for simulations.

<table>
<thead>
<tr>
<th>Level</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1</td>
<td>C (s)</td>
<td>20</td>
</tr>
<tr>
<td>Case 1</td>
<td>MTBF (s)</td>
<td>3600</td>
</tr>
<tr>
<td>Case 2</td>
<td>C (s)</td>
<td>3</td>
</tr>
<tr>
<td>Case 2</td>
<td>MTBF (s)</td>
<td>660</td>
</tr>
<tr>
<td>Case 3</td>
<td>C (s)</td>
<td>10</td>
</tr>
<tr>
<td>Case 3</td>
<td>MTBF (s)</td>
<td>660</td>
</tr>
<tr>
<td>Case 4</td>
<td>C (s)</td>
<td>864</td>
</tr>
<tr>
<td>Case 4</td>
<td>MTBF (s)</td>
<td>864</td>
</tr>
</tbody>
</table>
We run simulations for eight cases, whose parameters are presented in Table 5. These parameters have been used by Di et al. [8] to evaluate their two-level checkpointing model, and as such, each case consists of only two checkpointing levels. In their work, the authors proposed an optimal solution by solving complex mathematical equations using numerical method. Again, for each case, we compare the simulated overheads obtained with the three different approaches.

In this set of parameters, the MTBF has a large variation, ranging from more than 1 hour (Case 1) to less than 4 minutes (Case 8). Similarly, the checkpointing costs vary from 10s (Cases 4 and 5) to 300s (Case 8). Note that Cases 7 and 8 have both very short MTBFs and very high checkpointing costs, resulting in a lot of errors and recoveries. In particular, the checkpointing cost at level 2 in Case 8 (300s) is larger than the MTBF at level 1 (216s).

Results: Figure 6 presents the simulation results for the eight cases. First, we observe that the optimal algorithm by Di et al. only yields a slightly better simulated overhead compared to our simple first-order approximation solution (by less than 2% in Cases 1 to 6). However, our solution always improves significantly over Young/Daly’s formula, from 2% (Case 1) up to 100% (Case 6). Due to their short MTBFs, Cases 7 and 8 stand out and incur much higher overheads compared to the first six cases (thus their results are presented in a separate plot). Still, considering Case 8, we are able to improve over Young/Daly’s solution by as much as 2500% (in absolute value of the overhead), and we are off the optimal simulated overhead by only 300%. In addition, Figure 6 shows the theoretical overheads obtained both with our formula and the solution provided by Di et al. in [8]. As expected, our first-order approximation remains accurate when the MTBF is large, as in Cases 1, 2 and 4. However, it becomes less accurate with shorter MTBFs and higher error rates, especially in Cases 7 and 8 (which do not represent healthy HPC platforms).

4.4 Summary of results

From the simulation results, we conclude that first-order approximation remains a valuable performance model for evaluating checkpointing solutions in HPC systems (as long as the error rates stay reasonably low). We have demonstrated, through an extensive set of simulations with a wide range of parameters, the usefulness of multi-level checkpointing (over using only one level of checkpoints) with significantly reduced overheads. The results also corroborate the analytical study by showing the benefit of selecting an optimal subset of levels among all the levels available. Overall, our approach achieves the optimal or near-optimal performance in almost all cases, except when the MTBF is too small, in which case even the optimal solution yields an unacceptably high overhead (e.g., Case 8 of Table 5).

5 Conclusion

This paper has studied multi-level checkpointing protocols, where different levels of checkpoints can be set; lower levels deal with frequent errors that can be recovered at low cost (for instance with a memory copy), while higher levels allow us to recover from all errors, such as node failures (for instance with a copy in stable storage). We consider a general scenario with $k$ levels of faults, and we provide explicit formulas to characterize the optimal checkpointing pattern, up to first-order approximation. The overhead turns out to be of the order of $\sum_{l=1}^{k} \sqrt{2\lambda_l C_l}$, which elegantly extends Young/Daly’s classical formula.

The first-order approximation to the optimal $k$-level checkpointing pattern uses rational numbers of checkpoints, and we prove that all segments should have equal lengths. We corroborate the theoretical study by an extensive set of simulations, demonstrating that greedily rounding the rational values leads to an overhead very close to the lower bound. Furthermore, we provide a dynamic programming algorithm to determine those levels that should be selected, and the simulations confirm the optimality of the subset of levels returned by the dynamic programming algorithm.

The problem of finding a first-order optimal pattern with an integer number of segments to minimize the overhead remains open. It may well be the case that such an integer pattern is not periodic at each level and uses different-length segments. However, the good news is that the rounding of the rational solution provided in this paper seems quite efficient in practice.

Acknowledgment

We would like to thank the anonymous reviewers for their valuable comments, which have greatly improved the quality of this paper. This research was funded in part by the European project SCoRPIO, by the LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program “Investissements d'Avenir” (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR), by the PIA ELCI project, and by the ANR RESCUE project. Yves Robert is with Institut Universitaire de France.
Anne Benoit received the PhD degree from Institut National Polytechnique de Grenoble in 2003, and the Habilitation à Diriger des Recherches (HDR) from École Normale Supérieure de Lyon (ENS Lyon) in 2009. She is currently an associate professor in the Computer Science Laboratory LIP at ENS Lyon, France. She is the author of 7 books, 78 papers published in international conferences, and 147 papers published in international journals, and 30 PhD theses. Her research interests include algorithm design and scheduling techniques for parallel and distributed platforms, and also the performance evaluation of parallel systems and applications, with a focus on energy awareness and resilience. She is Associate Editor of IEEE TPDS, JPDC, and SUSCOM. She is the program chair of several workshops and conferences, in particular she is program chair for HiPC’2016, program co-chair for ICPP’2017, and technical papers chair for SC’2017. She is a senior member of the IEEE, and she has been elected a Junior Member of Institut Universitaire de France in 2009.

Aurélien Cavelan completed his master at the University of Orleans in 2014, and then moved to École Normale Supérieure de Lyon (ENS Lyon), where he is currently a PhD candidate under the supervision of Anne Benoit and Yves Robert. As part of completing his PhD, he is also a visiting student at Argonne National Laboratory, where he is working with Franck Cappello on optimizing the execution of large application workloads subject to faults. His main topics of interest include fault tolerance, energy efficiency and scheduling techniques for large scale platforms.

Valentin Le Fèvre received the M.S. degree in fundamental computer science from École Normale Supérieure de Lyon (ENS Lyon, France) in 2016, where he also received his B.S. degree in 2014. He is mainly interested in high-performance computing, in particular resilience and scheduling problems.

Yves Robert received the PhD degree from Institut National Polytechnique de Grenoble. He is currently a full professor in the Computer Science Laboratory LIP at ENS Lyon. He is the author of 7 books, 147 papers published in international journals, and 219 papers published in international conferences. He is the editor of 11 book proceedings and 13 journal special issues. He is the advisor of 30 PhD theses. His main research interests are scheduling techniques and resilient algorithms for large-scale platforms. Yves Robert served on many editorial boards, including IEEE TPDS and JPDC. He was the program chair of HiPC’2006 in Bangalore, IPDPS’2008 in Miami, ISPDC’2009 in Lisbon, ICPP’2013 in Lyon and HiPC’2013 in Bangalore. He is a Fellow of the IEEE. He has been elected a Senior Member of Institut Universitaire de France in 2007 and renewed in 2012. He has been awarded the 2014 IEEE TCSC Award for Excellence in Scalable Computing. He also holds an IEEE TCPP Outstanding Service Award. He holds a Visiting Scientist position at the University of Tennessee Knoxville since 2011. Contact him at Yves.Robert@inria.fr.

Hongyang Sun received his Ph.D. in Computer Science from Nanyang Technological University, Singapore in 2011. Prior to that, he completed his B.Eng. in Computer Engineering from the same university, and his M.Sc. in Computer Science from the National University of Singapore under the Singapore-MIT Alliance program. He has held research position at the Toulouse Institute Computer Science Research, France. Currently, he is a Postdoc Researcher at ENS de Lyon and INRIA, France. His main research interests include high-performance computing, scheduling and resource management, energy efficiency, design and analysis of algorithms, and fault tolerance. He has published over 30 peer-reviewed papers in international journals and conferences, and has served in the program committee of SBAC-PAD’15, IPDPS’16.