Two-level checkpointing and partial verifications for linear task graphs

Anne Benoit, Aurélien Cavelan, Yves Robert, and Hongyang Sun

RESEARCH REPORT
N° 8794
October 2015
Project-Team ROMA

ISSN 0249-6399 ISRN INRIA/RR--8794--FR+ENG

Two-level checkpointing and partial verifications for linear task graphs

Anne Benoit*†, Aurélien Cavelan*†, Yves Robert*†‡, and Hongyang Sun*†

Project-Team ROMA

Research Report n° 8794 — October 2015 — 16 pages

Abstract: Fail-stop and silent errors are unavoidable on large-scale platforms. Efficient resilience techniques must accommodate both error sources. A traditional checkpointing and rollback recovery approach can be used, with added verifications to detect silent errors. A fail-stop error leads to the loss of the whole memory content, hence the obligation to checkpoint on a stable storage (e.g., an external disk). On the contrary, it is possible to use in-memory checkpoints for silent errors, which provide a much smaller checkpoint and recovery overhead. Furthermore, recent detectors offer partial verification mechanisms, which are less costly than guaranteed verifications but do not detect all silent errors. In this paper, we show how to combine all these techniques for HPC applications whose dependence graph is a chain of tasks, and provide a sophisticated dynamic programming algorithm returning the optimal solution in polynomial time. Simulations demonstrate that the combined use of multi-level checkpointing and partial verifications further improves performance.

Key-words: resilience, fail-stop errors, silent errors, multi-level checkpoint, verification, dynamic programming.

* École Normale Supérieure de Lyon
† INRIA, France
‡ University of Tennessee Knoxville, USA

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Checkpoint à deux niveaux et vérifications partielles pour des graphes de tâches linéaires

Résumé : Les erreurs fatales et silencieuses ne peuvent plus être ignorées sur des plates-formes à grande échelle. Des techniques de résilience efficaces doivent accommoder les deux types d'erreurs. Une approche traditionnelle de checkpoint et points de reprise peut être utilisée, en rajoutant des vérifications afin de détecter les erreurs silencieuses. Une erreur fatale entraîne la perte de tout le contenu mémoire, d'où l'obligation de faire une sauvegarde sur un support fiable (typiquement un disque). Par contre, il est possible de se satisfaire de checkpoints en mémoire pour les erreurs silencieuses, ce qui donne des surcoûts bien plus faibles. De plus, les détecteurs récents offrent des mécanismes de vérification partielle, qui sont moins coûteux que les vérifications garanties, mais qui ne détectent pas toutes les erreurs silencieuses. Nous montrons comment combiner toutes ces techniques pour des applications HPC dont le graphe de dépendances est une chaîne de tâches, et nous donnons un algorithme de programmation dynamique sophistiqué qui renvoie la solution optimale en temps polynomial. Des simulations démontrent que l'utilisation combinée de checkpoint à deux niveaux et de vérifications partielles améliore la performance.

Mots-clés : résilience, erreurs fatales, erreurs silencieuses, checkpoint multi-niveaux, vérification, programmation dynamique.

1 Introduction

Resilience is one of the major challenges for extreme-scale computing. In particular, several types of errors should be considered. In addition to classical fail-stop errors (such as hardware failures), silent errors, also known as silent data corruptions, constitute another threat that cannot be ignored any longer [11, 14, 13, 10]. In order to deal with both types of errors, a traditional checkpointing and rollback recovery strategy can be used [8], coupled with a verification mechanism to detect silent errors [9].

Because verification mechanisms may be costly, alternative techniques capable to rapidly detect silent errors, with the risk of missing some errors, have been recently developed and studied [2, 7]. We call such verifications *partial verifications*, while perfect verifications (no error missed) are *guaranteed verifications*. Furthermore, rather than checkpointing only on stable storage, a lightweight mechanism of in-memory checkpoints can be provided: one keeps a local copy of the data that has not been corrupted when a silent error stroke, and can therefore be used to recover rapidly. However, such local copies are lost if a fail-stop error occurs, and hence copies on stable storage (i.e., classical disk checkpoints) must also be provided.

Combining all these approaches is challenging even for a simplified, yet realistic, application framework, consisting for instance of a set of application workflows exchanging data at the end of their executions. Such a framework can be modeled as a task graph whose dependences follow a linear chain. This scenario corresponds to an HPC application whose workflow is partitioned into a succession of (typically large) tightly-coupled computational kernels, each of them being identified as a task. At the end of each task, we can perform either a partial or a guaranteed intermediate verification of the task output; or, likely less frequently, we can perform a guaranteed verification followed by a memory checkpoint (we do not take the risk of storing a corrupted checkpoint, hence the need for a guaranteed verification); or again, likely even less frequently, we can perform a guaranteed verification, a memory checkpoint and a disk checkpoint in a row.

The main contribution of this paper is to provide a sophisticated dynamic programming algorithm that returns the optimal solution, i.e., the solution that minimizes the expected execution time. The originality is that we combine both types of verifications and both types of checkpoints. Furthermore, we present extensive simulations that demonstrate the usefulness of mixing these techniques, and in particular we demonstrate the gain obtained thanks to multi-level checkpointing.

To the best of our knowledge, the interplay of verification mechanisms with two types of checkpoints, in-memory and disk-based, has never been investigated for task graphs. Our previous work [5] considers linear chains with a single checkpoint type and guaranteed verifications (for the record, the pioneering paper [12] for linear chains only dealt with a single checkpoint type and no verification). The closest work to this paper is our recent work [6] for divisible applications, where we address the same combined framework (with two error sources, two checkpoint types and two verification types); however, in [6], we target long-lasting executions that are partitioned into periodic patterns that repeat over time, and we compute the best pattern up to first-order approximations. Here we do not have the flexibility of divisible applications, since we insert resilience mechanisms only at the end of the execution of a task. We may well have a limited number of tasks, which prevents the use of any periodic strategy. Instead, we use a completely different approach and design (quite involved) dynamic programming algorithms that provide the optimal solution for any linear task graph. We detail the model in Section 2, before giving the dynamic programming algorithm in Section 3 and providing simulation results in Section 4. Finally, we conclude in Section 5.

2 Model

We consider a chain of tasks $T_1, T_2, \ldots, T_n$, where each task T_i has a weight w_i corresponding to the computational load. For notational convenience, we also define $W_{i,j} = \sum_{k=i+1}^{j} w_k$ to be the time to execute tasks T_{i+1} to T_j for any $i \le j$. Furthermore, we assume that hardware faults (*fail-stop* errors) and silent data corruptions (*silent* errors) coexist, as motivated in Section 1. Since these two types of errors are caused by different sources, we assume that they are independent and that both occurrences follow a *Poisson process* with arrival rates λ_f and λ_s, respectively. The probability of having at least a fail-stop error during the execution of tasks T_{i+1} to T_j is given by $p^f_{i,j} = 1 - e^{-\lambda_f W_{i,j}}$ and that of having at least a silent error during the same execution is $p^s_{i,j} = 1 - e^{-\lambda_s W_{i,j}}$.

To deal with both fail-stop and silent errors, resilience is provided through the use of a two-level checkpointing scheme coupled with an error detection (or verification) mechanism. When a fail-stop error strikes, the computation is interrupted immediately due to a hardware fault, so all the memory content is destroyed: we then recover from the last disk checkpoint or start again at the beginning of the application. On the contrary, when a silent error is detected, either by a partial verification or by a guaranteed one, we roll back to the nearest memory checkpoint, and recover from the memory copy there, which is much cheaper than recovering from the last disk checkpoint.

We enforce that a memory checkpoint is always taken immediately before each disk checkpoint. This can be done with little overhead and it has been enforced in some practical multi-level checkpointing systems [4]. Also, a guaranteed verification is always taken immediately before each memory checkpoint, so that all checkpoints are valid (both memory and disk checkpoints), and hence only one memory checkpoint and one disk checkpoint need to be maintained at any time during the execution of the application. Furthermore, we assume that errors only strike the computations, while verifications, memory copies, and I/O transfers are protected from failures.

Let C_D denote the cost of disk checkpointing, C_M the cost of memory checkpointing, R_D the cost of disk recovery, and R_M the cost of memory recovery. Recall that when a disk recovery is done, we also need to restore the memory state. For simplicity, we assume that the cost R_M is included in the cost R_D. Also, let V^* denote the cost of guaranteed verification and V the cost of a partial verification. The partial verification is also characterized by its *recall*, which is denoted by r and represents the proportion of detected errors over all silent errors that have occurred during the execution. For notational convenience, we define $g = 1-r$ to be the proportion of undetected errors. Note that the guaranteed verification can be considered as one with recall $r^* = 1$. Since a partial verification usually incurs a much smaller cost yet has a reasonable recall [2, 7], it is highly attractive for detecting silent errors, and we make use of them between guaranteed verifications.

Finally, the objective is to decide where to place disk checkpoints, memory checkpoints, guaranteed verifications and partial verifications, in order to minimize the expected execution time of the application.

3 Dynamic programming

The goal is to find which task to verify, which task to checkpoint, and also which type of verification or checkpoint to perform, in order to minimize the expected execution time of the task chain. To solve this problem, we have derived a sophisticated multi-level dynamic programming algorithm. Recall that we assume that a memory checkpoint always comes with a guaranteed verification to ensure that the results are correct, and that a disk checkpoint always comes with a memory checkpoint, as motivated in Section 2. For convenience, we add a virtual task T_0, which is checkpointed on disk (and hence on memory), and whose recovery cost is zero. This accounts for the fact that it is always possible to restart the application from scratch at no extra cost. We first describe in Section 3.1 the general scheme when adding only guaranteed verifications, memory checkpoints and disk checkpoints. We then show how to extend this dynamic programming algorithm to partial verifications in Section 3.2.

3.1 Without partial verifications

Figure 1 illustrates the idea of the general algorithm without using partial verifications. The algorithm contains three dynamic programming levels, which are responsible for placing disk checkpoints, memory checkpoints, and guaranteed verifications, respectively, and an additional step to compute the expected execution time between any two verifications. The following describes each step of the algorithm in detail.

Placing disk checkpoints. The first level focuses on placing disk checkpoints. Let the function $E_{disk}(d_2)$ denote the expected time needed to successfully execute all the tasks from T_1 to T_{d_2}, where task T_{d_2} is verified and checkpointed on both disk and memory. In this function, we try all possible locations for the last checkpoint before T_{d_2}. For each possible location d_1, we call the function recursively on d_1 (to place disk checkpoints before T_{d_1}), and we add the expected time needed to execute the tasks from T_{d_1+1} to T_{d_2}. This is done through the $E_{mem}(d_1, d_2)$ function, which also decides where to place memory checkpoints, and accounts for the cost of memory checkpoints. The cost of the disk checkpoint C_D is finally added after T_{d_2}. Note that a location $d_1 = 0$ means that no further disk checkpoints are added. In this case, we simply let $E_{disk}(0) = 0$, which initializes the dynamic program. We can express $E_{disk}(d_2)$ as follows:

$$E_{disk}(d_2) = \min_{0 \le d_1 < d_2} \{E_{disk}(d_1) + E_{mem}(d_1, d_2) + C_D\}.$$

The total expected time needed to execute all the tasks T_1 to T_n is given by $E_{disk}(n)$.

Placing memory checkpoints. The second level aims at placing additional memory checkpoints between two disk checkpoints. The function is first called from the first level between two disk checkpoints, each of which also comes with a memory checkpoint. We define

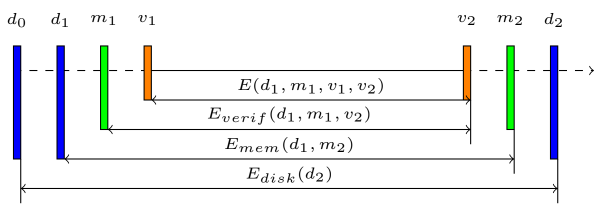

Figure 1: Without partial verifications.

$E_{mem}(d_1, m_2)$ as the expected time needed for successfully executing all the tasks from T_{d_1+1} to T_{m_2}, where there is a disk checkpoint at the end of task T_{d_1}, a memory checkpoint at the end of task T_{m_2}, and no other disk checkpoints. Note that there might be a disk checkpoint after T_{m_2}, for instance when we first call this function, but we do not account for the cost of this disk checkpoint in E_{mem}, only for the cost of the memory checkpoint (the cost of the disk checkpoint is already accounted for in E_{disk}). As before, we try all possible locations for the last memory checkpoint between tasks T_{d_1} and T_{m_2}. For each possible location m_1, we call the function recursively on tasks T_{d_1} to T_{m_1}, and then call the function for the next level, $E_{verif}(d_1, m_1, m_2)$, which computes the expected time needed to execute the tasks from T_{m_1+1} to T_{m_2} (and decides where to place verifications). Finally, we add the cost of the memory checkpoint C_M following T_{m_2}. We can express $E_{mem}(d_1, m_2)$ as follows:

$$\begin{aligned} &E_{mem}(d_1, m_2) = \\ &\quad \min_{d_1 \le m_1 < m_2} \{E_{mem}(d_1, m_1) + E_{verif}(d_1, m_1, m_2) + C_M\}. \end{aligned}$$

If $m_1 = d_1$, there is no extra memory checkpoint between d_1 and m_2, and therefore we initialize the dynamic program with $E_{mem}(d_1, d_1) = 0$.

Placing additional verifications. The third level looks for where to insert additional verifications between two tasks with memory checkpoints. The function is first called from the second level between two memory checkpoints, each of which also comes with a verification. Therefore, we define $E_{verif}(d_1, m_1, v_2)$ as the expected time needed for successfully executing all the tasks from T_{m_1+1} to T_{v_2}, knowing that the last memory checkpoint is after T_{m_1}, the last disk checkpoint is after T_{d_1}, and there are no checkpoints between T_{m_1+1} and T_{v_2}. Note that $E_{verif}(d_1, m_1, v_2)$ accounts only for the time required to execute and verify these tasks. As before, we try all possible locations for the last verification between T_{m_1} and T_{v_2}, and for each possible location v_1, we call the function recursively on tasks T_{m_1} to T_{v_1}. Furthermore, we add the expected time needed to successfully execute the tasks T_{v_1+1} to T_{v_2}, denoted by $E(d_1, m_1, v_1, v_2)$, knowing the position of the last disk checkpoint d_1 and the position of the last memory checkpoint m_1. We express $E_{verif}(d_1, m_1, v_2)$ as follows:

$$\begin{aligned} &E_{verif}(d_1, m_1, v_2) = \\ &\quad \min_{m_1 \le v_1 < v_2} \{E_{verif}(d_1, m_1, v_1) + E(d_1, m_1, v_1, v_2)\}. \end{aligned} \tag{1}$$

Again, the case $v_1 = m_1$ means that no further verifications are added, so we initialize the dynamic program with $E_{verif}(d_1, m_1, m_1) = 0$. The verification cost at the end of T_{v_2} is accounted for in the function $E(d_1, m_1, v_1, v_2)$.

Computing the expected execution time between two verifications. Finally, to compute the expected time needed for successfully executing several tasks between two verifications, we need the position of the last disk checkpoint d_1, the position of the last memory checkpoint m_1, and the positions of the two verifications v_1 and v_2. On the one hand, if a fail-stop error occurs with probability $p^f_{v_1,v_2}$, then the execution stops and we must recover from the last disk checkpoint. In this case, we lose $T^{\text{lost}}_{v_1,v_2}$ time, pay the cost of recovery R_D (set to 0 if $d_1 = 0$), and re-execute the tasks starting from T_{d_1}. The re-execution is done in three steps. First, we call $E_{mem}(d_1, m_1)$ to compute the expected time needed to re-execute the tasks from the last disk checkpoint after T_{d_1} to the last memory checkpoint after T_{m_1}. Then, we call the function $E_{verif}(d_1, m_1, v_1)$ to account for the time needed to re-execute the tasks between the last memory checkpoint after T_{m_1} to the next verification after T_{v_1}. Finally, we re-execute tasks T_{v_1+1} to T_{v_2} with $E(d_1, m_1, v_1, v_2)$.

On the other hand, with probability $1 - p^f_{v_1,v_2}$, there is no fail-stop error. In this case, we pay W_{v_1,v_2} by executing all the tasks from T_{v_1+1} to the next verification after T_{v_2}. Then we add the cost of the guaranteed verification V^*. After the verification, there is a probability $p^s_{v_1,v_2}$ of detecting a silent error. If a silent error is detected, we can recover from the last memory checkpoint with a cost R_M (set to 0 if $m_1 = 0$), and only re-execute the tasks from there by calling the function $E_{verif}(d_1, m_1, v_1)$ followed by $E(d_1, m_1, v_1, v_2)$, as before. Therefore:

$$
\begin{aligned}
E(d_1, m_1, v_1, v_2) = \\
& p^f_{v_1,v_2}\big(T^{\text{lost}}_{v_1,v_2} + R_D + E_{mem}(d_1, m_1) \\
& \quad + E_{verif}(d_1, m_1, v_1) + E(d_1, m_1, v_1, v_2)\big) \\
& + \left(1 - p^f_{v_1,v_2}\right)\Big(W_{v_1,v_2} + V^* + p^s_{v_1,v_2}\big(R_M \\
& \quad + E_{verif}(d_1, m_1, v_1) + E(d_1, m_1, v_1, v_2)\big)\Big).
\end{aligned} \tag{2}
$$

In order to compute the expected execution time, we need to compute $T^{\text{lost}}_{v_1,v_2}$, which is the expected time loss due to a fail-stop error occurring during the execution of tasks T_{v_1+1} to T_{v_2}. We derive:

$$
\begin{aligned}
T^{\text{lost}}_{v_1,v_2} &= \int_0^\infty x\mathbb{P}(X = x | X < W_{v_1,v_2})dx \\
&= \frac{1}{\mathbb{P}(X < W_{v_1,v_2})}\int_0^{W_{v_1,v_2}} x\mathbb{P}(X = x)dx \ ,
\end{aligned}
$$

where $\mathbb{P}(X = x)$ denotes the probability that a fail-stop error strikes at time x. By definition, we have $\mathbb{P}(X = x) = \lambda_f e^{-\lambda_f x}$ and $\mathbb{P}(X < W_{v_1,v_2}) = 1 - e^{-\lambda_f W_{v_1,v_2}}$. Integrating by parts, we get

$$
T^{\text{lost}}_{v_1,v_2} = \frac{1}{\lambda_f} - \frac{W_{v_1,v_2}}{e^{\lambda_f W_{v_1,v_2}} - 1} \ . \tag{3}
$$

Now, substituting $T^{\text{lost}}_{v_1,v_2}$ into Equation (2) and simplifying, we obtain:

$$
\begin{aligned}
E(d_1, m_1, v_1, v_2) &= e^{\lambda_s W_{v_1,v_2}}\left(\frac{e^{\lambda_f W_{v_1,v_2}} - 1}{\lambda_f} + V^*\right) \\
&+ e^{\lambda_s W_{v_1,v_2}}\left(e^{\lambda_f W_{v_1,v_2}} - 1\right)(R_D + E_{mem}(d_1, m_1)) \\
&+ \left(e^{(\lambda_s + \lambda_f) W_{v_1,v_2}} - 1\right) E_{verif}(d_1, m_1, v_1) \\
&+ \left(e^{\lambda_s W_{v_1,v_2}} - 1\right) R_M.
\end{aligned}
$$

Complexity. The complexity is dominated by the computation of the table $E_{verif}(d_1, m_1, v_2)$, which contains $O(n^3)$ entries, and each entry depends on at most n other entries that are already computed. All tables are computed in a bottom-up fashion, from the left to the right of the intervals. Hence, the overall complexity of the algorithm is $O(n^4)$.

3.2 With partial verifications

It may be beneficial to further add partial verifications between two guaranteed verifications. The intuitive idea would be to add yet another level to the dynamic programming algorithm, and to replace $E(d_1, m_1, v_1, v_2)$ in Equation (1) by a call to a function $E^{(intuitive)}_{partial}(d_1, m_1, v_1, p_2, v_2)$, with $p_2 = v_2$, which would compute the expected time needed to execute all the tasks from T_{v_1+1} to T_{p_2} and add further partial verifications (computed from the left to the right).

However, while the dynamic programming approach was rather intuitive without partial verifications, the problem becomes much harder with partial verifications. The main reason is that when computing an interval between two partial verifications, there is a probability g that the error remains undetected after the partial verification. When this happens, we need to account for the time lost executing the following tasks until the error is detected (eventually by the guaranteed verification) or until the execution is interrupted by a fail-stop error. This is only possible if we know the optimal positions of the partial verifications after the interval up to the next guaranteed verification. This requires the dynamic programming algorithm to first compute the values at the right of the current interval, hence progressing the opposite way as what was done so far. Therefore, the function becomes $E_{partial}(d_1, m_1, v_1, p_1, v_2)$ (expected time needed to execute all the tasks from T_{p_1+1} to T_{v_2}), and it tries all positions p_2 for the next partial verification. But then, it also requires to remove some terms that account for re-executed work from the intervals on the left of the current interval (because we do not have this information yet), and to re-inject them later in the computation. Altogether we have quite a complicated algorithm!

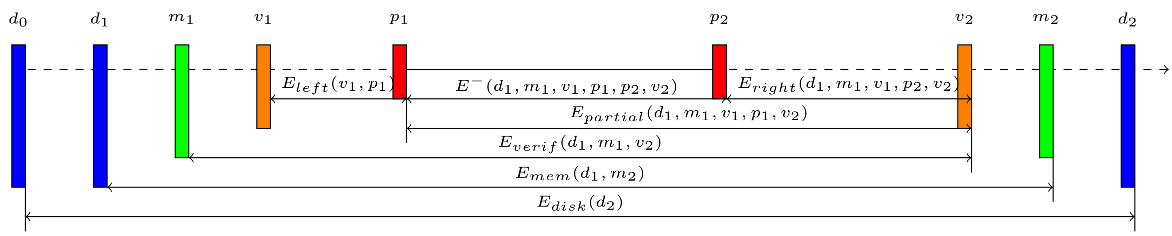

Figure 2: With partial verifications.

Expected lost time in case of silent error. First, we compute $E_{right}(d_1, m_1, v_1, p_1, v_2)$, the expected time lost executing the tasks T_{p_1+1} to T_{v_2}, assuming that there was a silent error in this interval. This computation uses p_2, the optimal position of the verification immediately following p_1, which is computed with the dynamic programming. Indeed, T_{p_2} may detect the error or not. If the error is detected by T_{p_2}, we loose the work $W_{p_1,p_2} + V + R_M$, while we use $E_{right}(d_1, m_1, v_1, p_2, v_2)$ if the error remains undetected. Also, we consider fail-stop errors only between T_{p_1+1} and T_{p_2}, because fail-stop errors between T_{p_2+1} and T_{v_2} will be accounted for in $E_{right}(d_1, m_1, v_1, p_2, v_2)$. Note that even if we know that there is a silent error in the interval, we may need to recover from a fail-stop error if it strikes before the silent error is detected. Altogether:

$$
\begin{aligned}
&E_{right}(d_1, m_1, v_1, p_1, v_2) = \\
&\quad p^f_{p_1,p_2}\left(T^{\text{lost}}_{p_1,p_2} + R_D + E_{mem}(d_1, m_1)\right) \\
&\quad + (1 - p^f_{p_1,p_2})(W_{p_1,p_2} + V + (1-g)R_M \\
&\qquad + gE_{right}(d_1, m_1, v_1, p_2, v_2)),
\end{aligned}
$$

and hence:

$$
\begin{aligned}
&E_{right}(d_1, m_1, v_1, p_1, v_2) = \\
&\quad (1 - e^{-\lambda_f W_{p_1,p_2}})\Big(\frac{1}{\lambda_f} - \frac{W_{p_1,p_2}}{e^{\lambda_f W_{p_1,p_2}} - 1} \\
&\qquad + R_D + E_{mem}(d_1, m_1)\Big) \\
&\quad + e^{-\lambda_f W_{p_1,p_2}}(W_{p_1,p_2} + V + (1-g)R_M \\
&\qquad + gE_{right}(d_1, m_1, v_1, p_2, v_2)).
\end{aligned}
$$

The initialization is $E_{right}(d_1, m_1, v_1, v_2, v_2) = R_M$. Indeed, in this case, there is no task to execute, and if there was a silent error, it is therefore immediately detected by v_2 (a guaranteed verification), and we just pay R_M. Knowing p_2, we are therefore able to compute all values of E_{right}. We will see later how we use this knowledge in E_{right}. Note that the time to re-execute the tasks after a recovery is omitted here, since it will be accounted for when computing $E(d_1, m_1, v_1, p_1, p_2, v_2)$, the expected time needed to successfully execute all the tasks between two partial verifications (from T_{p_1+1} to T_{p_2}).

Expected time to compute tasks T_{p_1+1} to T_{p_2}. Figure 2 shows all the tasks involved in the computation of an interval consisting of several tasks between two partial verifications at p_1 and p_2. Let $E(d_1, m_1, v_1, p_1, p_2, v_2)$ denote the expected time needed to successfully execute all the tasks from T_{p_1+1} to T_{p_2}, knowing that the last disk checkpoint is after T_{d_1}, the last memory checkpoint is after T_{m_1}, the last guaranteed verification is after T_{v_1}, and the next guaranteed verification is after T_{v_2}.

On the one hand, if a fail-stop error occurs with probability $p^f_{p_1,p_2}$, then the task stops and we must recover from the last disk checkpoint. We lose $T^{\text{lost}}_{p_1,p_2}$ time, we pay the cost for the disk recovery R_D, and we need to re-execute the tasks starting from T_{d_1}. This is done in three steps: first we call $E_{mem}(d_1, m_1)$ to compute the expected time needed to re-execute the tasks from the last disk checkpoint after T_{d_1} to the last memory checkpoint after T_{m_1}. Then we call the function $E_{verif}(d_1, m_1, v_1)$ to account for the time needed to re-execute the tasks between the last memory checkpoint after T_{m_1} to the next guaranteed verification after T_{v_1}, and finally we are left with the remaining tasks between T_{v_1+1} and T_{p_1}. Let $E_{left}(v_1, p_1)$ denote the expected time needed to re-execute all the tasks from T_{v_1+1} to T_{p_1}. Finally, we can re-execute tasks T_{v_1+1} to T_{v_2} by calling $E(d_1, m_1, v_1, p_1, p_2, v_2)$.

On the other hand, there is a probability $(1 - p^f_{p_1,p_2})$ of having no fail-stop errors. In that case, we execute all the tasks from T_{p_1+1} to the next verification after T_{p_2} and we pay W_{p_1,p_2}. Then we add the cost for the verification V. After the partial verification, there is a probability p^s of having a silent error. In this case, we pay a recovery from the last memory checkpoint (R_M) and re-executed the tasks from there: we call $E_{verif}(d_1, m_1, v_1)$, followed by $E_{left}(v_1, p_1)$ and $E(d_1, m_1, v_1, p_1, p_2, v_2)$. Furthermore, if the error was not detected (with probability g), we use $E_{right}(d_1, m_1, v_1, p_2, v_2)$ to compute the expected time lost executing the tasks following T_{p_2}, knowing that there is an undetected silent error (as explained earlier). Therefore:

$$
\begin{aligned}
&E(d_1, m_1, v_1, p_1, p_2, v_2) = \\
&\quad p^f_{p_1,p_2}\big(T^{\text{lost}}_{p_1,p_2} + R_D + E_{mem}(d_1, m_1) \\
&\qquad + E_{verif}(d_1, m_1, v_1) + E_{left}(v_1, p_1) \\
&\qquad\qquad + E(d_1, m_1, v_1, p_1, p_2, v_2)\big) \\
&\quad + (1 - p^f_{p_1,p_2})\Big(W_{p_1,p_2} + V + p^s\big(E_{verif}(d_1, m_1, v_1) + E_{left}(v_1, p_1) \\
&\qquad\qquad + E(d_1, m_1, v_1, p_1, p_2, v_2) + (1 - g)R_M \\
&\qquad\qquad + gE_{right}(d_1, m_1, v_1, p_2, v_2)\big)\Big).
\end{aligned} \tag{4}
$$

Substituting $T^{\text{lost}}_{p_1,p_2}$ into Equation (4) and simplifying, we obtain:

$$
\begin{aligned}
&E(d_1, m_1, v_1, p_1, p_2, v_2) = e^{\lambda_s W_{p_1,p_2}}\left(\frac{e^{\lambda_f W_{p_1,p_2}} - 1}{\lambda_f} + V\right) \\
&+ e^{\lambda_s W_{p_1,p_2}}\left(e^{\lambda_f W_{p_1,p_2}} - 1\right)(R_D + E_{mem}(d_1, m_1)) \\
&+ \left(e^{(\lambda_s+\lambda_f) W_{p_1,p_2}} - 1\right)\big(E_{verif}(d_1, m_1, v_1) + E_{left}(v_1, p_1)\big) \\
&+ \left(e^{\lambda_s W_{p_1,p_2}} - 1\right)\big((1 - g)R_M + gE_{right}(d_1, m_1, v_1, p_2, v_2)\big).
\end{aligned}
$$

Finally, because we do not know at this point how to compute $E_{left}(v_1, p_1)$, we remove the term

$$\left(e^{(\lambda_s+\lambda_f) W_{p_1,p_2}} - 1\right) E_{left}(v_1, p_1)$$

from $E(d_1, m_1, v_1, p_1, p_2, v_2)$. This corresponds to the amount of time needed to re-execute all the tasks from T_{v_1+1} to T_{p_1} when there is an error between T_{p_1+1} and T_{p_2}. This time will be added back when computing $E_{partial}$, as explained below. Therefore, we introduce the modified expression of E, denoted E^-, as follows:

$$
\begin{aligned}
&E^-(d_1, m_1, v_1, p_1, p_2, v_2) = e^{\lambda_s W_{p_1,p_2}}\left(\frac{e^{\lambda_f W_{p_1,p_2}} - 1}{\lambda_f} + V\right) \\
&+ e^{\lambda_s W_{p_1,p_2}}\left(e^{\lambda_f W_{p_1,p_2}} - 1\right)(R_D + E_{mem}(d_1, m_1)) \\
&+ \left(e^{(\lambda_s+\lambda_f) W_{p_1,p_2}} - 1\right)\big(E_{verif}(d_1, m_1, v_1)\big) \\
&+ \left(e^{\lambda_s W_{p_1,p_2}} - 1\right)\big((1 - g)R_M + gE_{right}(d_1, m_1, v_1, p_2, v_2)\big).
\end{aligned}
$$

Computing $E_{partial}(d_1, m_1, v_1, p_1, v_2)$, the expected time needed to execute all the tasks from T_{p_1+1} to T_{v_2} (and placing extra partial verifications). Finally, we need to compute $E_{partial}$ and to decide when to use additional partial verifications on tasks that are not yet verified. The function is first called from the third level between two guaranteed verifications, and p_1 is originally set to v_1. Therefore, $E_{partial}(d_1, m_1, v_1, p_1, v_2)$ denotes the expected time needed to execute all the tasks from T_{p_1+1} to T_{v_2}, where T_{p_1} is followed by a partial verification (with the exception of the first call) and T_{v_2} is followed by a guaranteed verification, knowing the position of the last disk checkpoint d_1, the last memory checkpoint m_1 and the last guaranteed verification v_1.

Contrarily to the expressions derived in Section 3.1, note that partial verifications are placed from the left to the right. We use the expression of E^-, trying all possible positions p_2 for the partial verification following p_1, and we account for the fact that tasks between T_{p_1+1} and T_{p_2} may be re-executed several times (because we removed E_{left} from E^-). In fact, for any number of partial verifications between p_2 and v_2, we can show that $E^-(d_1, m_1, v_1, p_1, p_2, v_2)$ is re-executed $e^{(\lambda_s+\lambda_f)W_{p_2,v_2}}$ times, and hence we obtain:

$$\begin{aligned}
&E_{partial}(d_1, m_1, v_1, p_1, v_2) = \\
&\quad \min_{p_1 < p_2 \le v_2} \Big\{ E^-(d_1, m_1, v_1, p_1, p_2, v_2) \times e^{(\lambda_s+\lambda_f)W_{p_2,v_2}} \\
&\qquad\qquad + E_{partial}(d_1, m_1, v_1, p_2, v_2) \Big\}.
\end{aligned}$$

The initialization case is for $p_2 = v_2$. Then, there is no re-execution of the interval induced by errors to the right of p_2 (within an E_{left}), and therefore we compute only once $E^-(d_1, m_1, v_1, p_1, v_2, v_2)$. Furthermore, the interval is ended by a guaranteed verification, and therefore we add the corresponding verification cost:

$$\begin{aligned}
&E_{partial}(d_1, m_1, v_1, v_2, v_2) = \\
&\quad E^-(d_1, m_1, v_1, p_1, v_2, v_2) + e^{(\lambda_s+\lambda_f)W_{p_1,v_2}}(V^* - V)
\end{aligned}$$

Because partial verifications are placed from the left to the right, when implementing the algorithm, we first compute all values of $E_{partial}$ on the right of the interval, which are needed to progress towards the left. This is why we always have the values of the next p_2 when computing E_{right}, which correspond to the minimum value selected by $E_{partial}$. However, it was not possible to derive the values of the re-execution for the left part of the interval, hence the trick to compute the number of times each interval is re-executed, due to a failure on the right (the E_{left} that is removed from the initial expression of E).

Accounting for re-executions *on the left*. Finally, let us show that for any number of partial verifications between p_2 and v_2, the function $E^-(d_1, m_1, v_1, p_1, p_2, v_2)$ re-executes an amount $e^{(\lambda_s+\lambda_f)W_{p_2,v_2}}$ of work. If there are no partial verifications after p_2, then it is executed once when progressing within the computation, and we also need to account for the $e^{(\lambda_s+\lambda_f)W_{p_2,v_2}} - 1$ re-executed work due to the E_{left} term that was suppressed from $E^-(d_1, m_1, v_1, p_2, v_2, v_2)$.

With one intermediate partial verification p_3 between p_2 and v_2, the same reasoning shows that there is an amount of $e^{(\lambda_s+\lambda_f)W_{p_2,p_3}}$ re-executed work coming from the initial execution and the E_{left} term suppressed from $E^-(d_1, m_1, v_1, p_2, p_3, v_2)$. Furthermore, there is an amount of $(e^{(\lambda_s+\lambda_f)W_{p_3,v_2}} - 1)$ re-executed work coming from the E_{left} term of $E^-(d_1, m_1, v_1, p_3, v_2, v_2)$. In turn, this re-executed work incurs $e^{(\lambda_s+\lambda_f)W_{p_2,p_3}}$ re-executed work (initial execution and re-executions due to the E_{left} term coming from $E^-(d_1, m_1, v_1, p_2, p_3, v_2)$). Overall, the number of re-executed work is finally

$$e^{(\lambda_s+\lambda_f)W_{p_2,p_3}} + \left(e^{(\lambda_s+\lambda_f)W_{p_2,p_3}}\right)\left(e^{(\lambda_s+\lambda_f)W_{p_3,v_2}} - 1\right),$$

and therefore there is a total amount of $e^{(\lambda_s+\lambda_f)W_{p_2,v_2}}$ re-executed work. It is easy to extend this reasoning to any number of intervals by induction, assuming that it is true for i intermediate partial verifications $pv_i, \ldots, pv_1$, and adding a partial verification pv_{i+1} between p_2 and pv_i. The same reasoning holds.

Complexity. Clearly, the complexity is now dominated by the computation of the table $E_{partial}(d_1, m_1, v_1, p_1, v_2)$, which contains $O(n^5)$ entries, and each entry depends on at most n other entries that are already computed. Hence, the overall complexity of the algorithm is $O(n^6)$.

4 Performance evaluation

In this section, we conduct a set of simulations to assess the relative efficiency of our approach under realistic scenarios. We instantiate the model with actual parameters from the literature and we compare the performance of three algorithms: (i) a single level algorithm A_{DV^*} with only disk checkpoints (and additional guaranteed verifications), (ii) a two-level algorithm combining memory and disk checkpoints A_{DMV^*} (as in Section 3.1), and (iii) the complete algorithm using additional partial verifications A_{DMV} (as in Section 3.2). The optimal positions of verifications and disk checkpoints can be easily derived for A_{DV^*}, using a simplification of the proposed dynamic programming algorithm in Section 3.1 with no additional memory checkpoints.

Simulation setup. We make several assumptions on the input parameters. First, we assume that the recovery cost is equivalent to the corresponding checkpointing cost, i.e., $R_D = C_D$ and $R_M = C_M$. This is reasonable because writing a checkpoint and reading one typically takes the same amount of time. Then, we assume that a guaranteed verification must check all the data in memory, making its cost in the same order as that of a memory checkpoint, i.e., $V^* = C_M$. Furthermore, we assume partial verifications similar to those proposed in [7, 2, 3], with very low cost while offering good recalls. In the following, we set $V = \frac{V^*}{100}$ and $r = 0.8$. Also, the total work is fixed to 25000 seconds and it is distributed uniformly between up to 50 tasks. All these choices are somewhat arbitrary and can easily be modified in the evaluations; we believe they represent reasonable values for current and next-generation HPC applications. The code is publicly available at `http://graal.ens-lyon.fr/~yrobert/chain2levels.zip` for the interested readers to experiment with their parameters.

Platform settings. Table 1 presents the four platforms used in the simulations and their main parameters. These platforms have been used to evaluate the Scalable Checkpoint/Restart (SCR) library by Moody et al. [10], who provide accurate measurements for λ_f, λ_s, C_D and C_M using real applications. Note that the Hera platform has the worst error rates, with a platform MTBF of 12.2 days for fail-stop errors and 3.4 days for silent errors. In comparison, and despite its higher number of nodes, the Coastal platform features a platform MTBF of 28.8 days for fail-stop errors and 5.8 days for silent errors. In addition, the last platform uses SSD technology for memory checkpointing, which provides more data space, at the cost of higher checkpointing costs.

platform	#nodes	λ_f	λ_s	C_D	C_M
Hera	256	9.46e-7	3.38e-6	$300s$	$15.4s$
Atlas	512	5.19e-7	7.78e-6	$439s$	$9.1s$
Coastal	1024	4.02e-7	2.01e-6	$1051s$	$4.5s$
Coastal SSD	1024	4.02e-7	2.01e-6	$2500s$	$180.0s$

Table 1: Platform parameters.

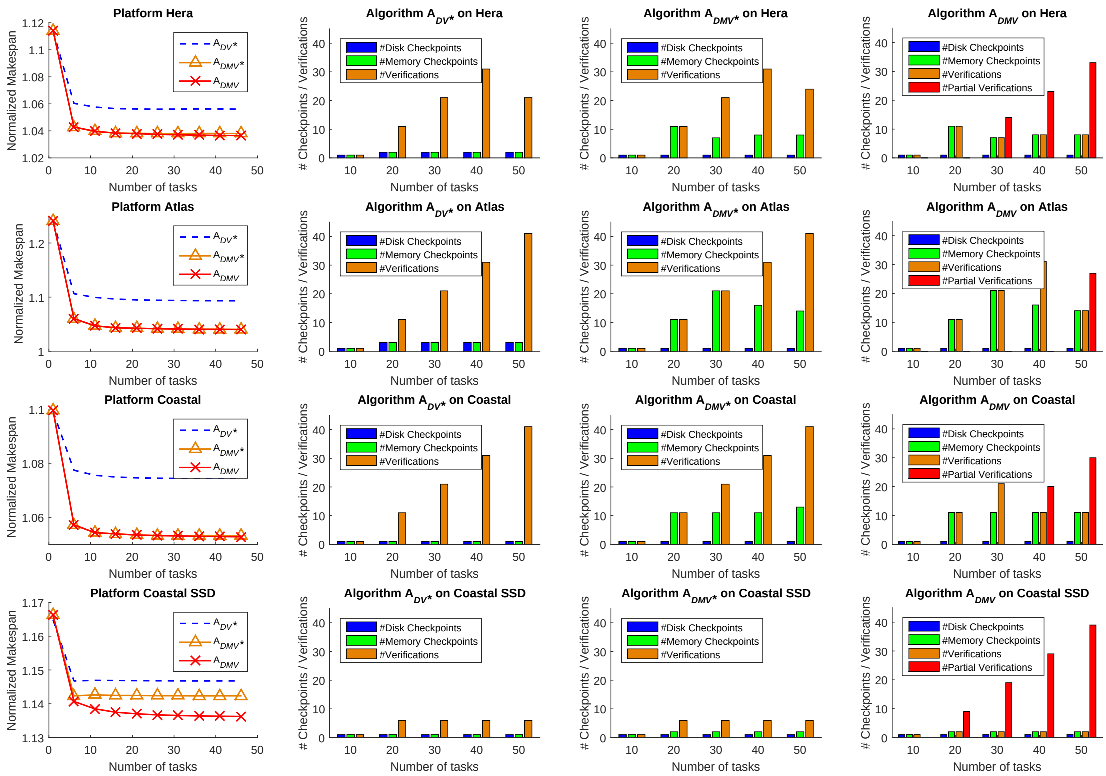

Figure 3: Performance of the three algorithms on the four platforms. Each line represents a platform.

Algorithm performance. The first column of Figure 3 presents, for each platform, the normalized makespan with respect to the execution time without error for different numbers of tasks. First, note that varying the number of tasks has an impact on both the size of the tasks and the maximum number of checkpoints and verifications that the algorithms can choose from. On the one hand, when the number of tasks is small (i.e., less than 5), the probability of having an error during the execution (either a fail-stop or a silent) increases quickly and reaches more than 10% on Hera for a single task. As a consequence, the application experiences more recoveries and re-executions (with significantly larger tasks), which increases the final overhead. However, when the number of tasks is large enough (i.e., more than 5), then tasks become small and the probability of having an error during the execution drops below 1% for one task, reducing recovery and re-execution costs at the same time.

Single level algorithm. The second column of Figure 3 shows the numbers of disk checkpoints (with associated memory checkpoints) and guaranteed verifications used by the A_{DV^*} algorithm on the four platforms and for different numbers of tasks. We observe that the number of guaranteed verifications is often set to the maximum (i.e., the number of tasks) while the number of checkpoints remains relatively small (i.e., less than 10 for all the platforms). This is because checkpoints are costly, and verifications help reducing the amount of time lost due to silent errors. Because they are cheap, the algorithm tends to place as many as possible. The algorithm limits their number only when the number of tasks is large enough (i.e., 50 on Hera) or the cost of the verification is too high, as it is on Coastal SSD.

Two-level algorithm. The third column of Figure 3 presents the numbers of disk checkpoints, memory checkpoints and guaranteed verifications used by the A_{DMV^*} algorithm on the four platforms and for different number of tasks. When using additional memory checkpoints, we observe that the number of guaranteed verifications remains similar to that of the number shown in the previous column concerning the A_{DV^*} algorithm. However, the algorithm now uses additional memory checkpoints, which drastically reduces the amount of time lost in re-execution when a silent error is detected. In particular, we observe that the two-level checkpointing algorithm A_{DMV^*} always lead to a better makespan compared to the single level algorithm A_{DV^*}, with 2% on Hera or 2.5% on Coastal, as shown in the first column, thus demonstrating the usefulness of our approach.

With partial verifications. The last column of Figure 3 presents the numbers of disk checkpoints, memory checkpoints, guaranteed verifications and additional partial verifications used by the A_{DMV} algorithm on the four platforms and for different numbers of tasks. With our settings, partial verifications are always more cost-effective than guaranteed verifications. But due to the smaller recall, they are only worth it if one can use a lot of them, which is only possible when the number of tasks is large enough. Therefore, the algorithm only starts to use partial verifications when the number of tasks is greater than 30 on Hera, 40 on Coastal and 50 on Atlas, where the silent error rate is the highest among the four platforms. Overall, adding partial verifications has a limited impact on the final makespan, with the exception of the Coastal SSD platform, where the cost of checkpoints and verifications are much higher than on the other platforms. Partial verifications, being 100 times cheaper than guaranteed verifications, remain the only affordable resilience tool on this platform. In this case, we observe an improved makespan (a little bit less than 1% with 50 tasks) compared to the A_{DMV^*} algorithm, as shown in the first column of Figure 3.

5 Conclusion

In this paper, we proposed a two-level checkpointing scheme to cope with both fail-stop errors and silent data corruptions on large-scale platforms. While numerous studies have dealt with either error source, few have dealt with both, while it is mandatory to address both sources simultaneously at scale. By combining standard disk checkpointing technique with in-memory checkpoints and verification mechanisms (partial or guaranteed), we have designed a multi-level dynamic programming algorithm that computes the optimal solution for a linear application workflow in polynomial time. Simulations based on realistic parameters on several platforms show consistent results, and confirm the benefit of the combined approach. While the most general algorithm has a high complexity in $O(n^6)$, where n is the number of tasks, it executes within a few seconds for $n = 50$, and therefore can be readily used for real-life linear workflows whose size rarely exceed ten or twenty tasks.

One interesting future direction is to assess the usefulness of this approach on general application workflows. The problem gets much more challenging, even in the simplified scenario where each task requires the entire platform to execute. In fact, in this simplified scenario, it is already NP-hard to decide which task to checkpoint in a simple join graph ($n - 1$ source tasks and a common sink task), with only fail-stop errors striking (hence a single level of checkpoint and no verification at all) [1]. Still, heuristics are urgently needed to address the same problem as in this paper, with two error sources, two checkpoint types and two verification types, if we are to deploy HPC workflows efficiently at scale.

References

[1] G. Aupy, A. Benoit, H. Casanova, and Y. Robert. Scheduling computational workflows on failure-prone platforms. In *17th Workshop on Advances in Parallel and Distributed Computational Models APDCM 2015*. IEEE Computer Society Press, 2015.

[2] L. Bautista Gomez and F. Cappello. Detecting silent data corruption through data dynamic monitoring for scientific applications. *SIGPLAN Notices*, 49(8):381–382, 2014.

[3] L. Bautista Gomez and F. Cappello. Detecting and correcting data corruption in stencil applications through multivariate interpolation. In *Proc.1st Int. Workshop on Fault Tolerant Systems (FTS)*, 2015.

[4] L. Bautista-Gomez, S. Tsuboi, D. Komatitsch, F. Cappello, N. Maruyama, and S. Matsuoka. FTI: High performance fault tolerance interface for hybrid systems. In *Proc. SC'11*, 2011.

[5] A. Benoit, A. Cavelan, Y. Robert, and H. Sun. Assessing general-purpose algorithms to cope with fail-stop and silent errors. In *Proc. 5th Int. Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*, 2014.

[6] A. Benoit, A. Cavelan, Y. Robert, and H. Sun. Optimal resilience patterns to cope with fail-stop and silent errors. Research report RR-8786, INRIA, 2015. Available at `graal.ens-lyon.fr/~yrobert/rr8786.pdf`.

[7] E. Berrocal, L. Bautista-Gomez, S. Di, Z. Lan, and F. Cappello. Lightweight silent data corruption detection based on runtime data analysis for HPC applications. In *Proc. HPDC*, 2015.

[8] K. M. Chandy and L. Lamport. Distributed snapshots: Determining global states of distributed systems. *ACM Transactions on Computer Systems*, 3(1):63–75, 1985.

[9] Z. Chen. Online-ABFT: An online algorithm based fault tolerance scheme for soft error detection in iterative methods. In *Proc. PPoPP*, pages 167–176, 2013.

[10] A. Moody, G. Bronevetsky, K. Mohror, and B. R. d. Supinski. Design, Modeling, and Evaluation of a Scalable Multi-level Checkpointing System. In *Proc. SC'10*. ACM/IEEE, 2010.

[11] T. O'Gorman. The effect of cosmic rays on the soft error rate of a DRAM at ground level. *IEEE Trans. Electron Devices*, 41(4):553–557, 1994.

[12] S. Toueg and O. Babaoğlu. On the optimum checkpoint selection problem. *SIAM J. Comput.*, 13(3), 1984.

[13] J. Ziegler, M. Nelson, J. Shell, R. Peterson, C. Gelderloos, H. Muhlfeld, and C. Montrose. Cosmic ray soft error rates of 16-Mb DRAM memory chips. *IEEE Journal of Solid-State Circuits*, 33(2):246–252, 1998.

[14] J. F. Ziegler, H. W. Curtis, H. P. Muhlfeld, C. J. Montrose, and B. Chin. IBM experiments in soft fails in computer electronics. *IBM J. Res. Dev.*, 40(1):3–18, 1996.

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399