Checkpointing strategies to protect parallel jobs from non-memoryless fail-stop errors

Anne Benoit, Lucas Perotin, Yves Robert, Frédéric Vivien

Checkpointing strategies to protect parallel jobs from non-memoryless fail-stop errors

Anne Benoit, Lucas Perotin, Yves Robert, Frederic Vivien

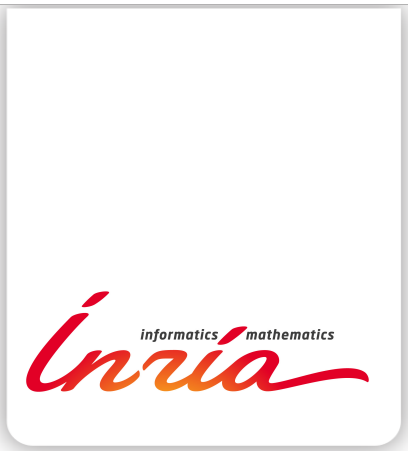
RESEARCH REPORT

N° 9465

March 2022

Project-Team ROMA

ISSN 0249-6399 ISRN INRIA/RR--9465--FR+ENG



Checkpointing strategies to protect parallel jobs from non-memoryless fail-stop errors

Anne Benoit*, Lucas Perotin*, Yves Robert*†, Frederic Vivien*

Project-Team ROMA

Research Report n° 9465 — March 2022 — 51 pages

Abstract: This paper studies checkpointing strategies for parallel jobs subject to fail-stop errors. The optimal strategy is well known when failure inter-arrival times obey an Exponential law, but it is unknown for non-memoryless failure distributions. We explain why the latter fact is misunderstood in recent literature. We propose a general strategy that maximizes the expected efficiency until the next failure, and we show that this strategy is asymptotically optimal for very long jobs. Through extensive simulations, we show that the new strategy is always at least as good as the Young/Daly strategy for various failure distributions. For distributions with a high infant mortality (such as LogNormal 2.51 or Weibull 0.5), the execution time is divided by a factor 1.9 on average, and up to a factor 4.2 for recently deployed platforms.

Key-words: checkpoint, failures, parallel jobs, non-memoryless

* Roma project-team, LIP laboratory, ENS Lyon, France
† Innovative Computing Laboratory, University of Tennessee Knoxville, USA

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

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

Stratégies de checkpoint pour protéger les tâches parallèles contre des erreurs ayant des distributions générales

Résumé : Cet article étudie les stratégies de checkpoint pour des tâches parallèles sujettes à des erreurs fatales. La stratégie optimale est bien connue lorsque les temps d'inter-arrivée des pannes obéissent à une loi exponentielle, mais elle est inconnue pour les distributions d'erreurs générales. Nous expliquons pourquoi ce dernier fait est mal compris dans la littérature récente. Nous proposons une stratégie générale qui maximise l'efficacité attendue jusqu'à la prochaine défaillance, et nous montrons que cette stratégie est asymptotiquement optimale pour les travaux très longs. Par des simulations extensives, nous montrons que la nouvelle stratégie est toujours au moins aussi bonne que la stratégie de Young/Daly pour diverses distributions de pannes. Pour les distributions avec une mortalité infantile élevée (comme LogNormal 2.51 ou Weibull 0.5), le temps d'exécution est divisé par un facteur 1.9 en moyenne, et jusqu'à un facteur 4.2 pour des plates-formes récemment déployées.

Mots-clés : checkpoint, fautes, tâches parallèles, distributions d'erreurs générales.

1 Introduction

Checkpointing is the standard technique to protect applications running on HPC (High Performance Computing) platforms. Every day, the platform experiences a few fail-stop errors (or failures, we use both terms indifferently). After each failure, the application executing on the faulty processor (and likely on many other processors for a large parallel application) is interrupted and must be restarted. Without checkpointing, all the work executed for the application is lost. With checkpointing, the execution can resume from the last checkpoint, after some downtime (enroll a spare to replace the faulty processor) and a recovery (read the checkpoint).

Consider a parallel application executing on a platform whose nodes are subject to fail-stop errors. How frequently should it be checkpointed so that its expected execution time is minimized? There is a well-known trade-off: taking too many checkpoints leads to a high overhead, especially when there are few failures, while taking too few checkpoints leads to a large re-execution time after each failure. However, the optimal strategy is known only when failure inter-arrival times obey an Exponential distribution. In that case, the optimal checkpointing period is (approximately) given by the Young/Daly formula as $W_{YD} = \sqrt{2\mu C}$ [43, 11], where μ is the application MTBF (Mean Time Between Failures) and C is the checkpoint duration.

This paper revisits checkpointing strategies for parallel jobs on platforms subject to failures that obey arbitrary probability distributions. This is a very important topic because the most accurate probability distributions to model processor failures are LogNormal [21] and Weibull [32, 33, 38, 39] instead of Exponential. For instance, LANL failure traces are best fit by Weibull distributions of different shapes [15]. However, dealing with non-memoryless distributions induces dramatic difficulties. Several recent papers mistakenly consider that if each processor experiences failures distributed according to some non-memoryless distribution, then the platform as a whole will experience failures distributed according to the same (scaled) distribution. This is wrong, unless we could rejuvenate thousands of processors each time a failure hits one single processor. Hence, it is important to provide a provenly correct strategy for arbitrary distributions.

The main contributions of this paper are the following:
• A synthetic overview of known results for Exponential distributions, some of which being frequently rediscovered;
• A detailed explanation of why non-memoryless distributions require a fully different approach;
• The design of a new checkpointing strategy, NextStep, which is asymptotically optimal for arbitrary distributions;
• A practical and fast implementation of NextStep through time discretization and numerical approximation;
• A detailed experimental comparison with the standard Young/Daly approach.

The paper is organized as follows. We first survey related work in Section 2. Then, we provide background on checkpointing parallel jobs with Exponential or non-memoryless distributions in Section 3. We detail the design of the checkpointing strategy NextStep in Section 4, and show that it is asymptotically optimal for arbitrary distributions. The experimental evaluation in Section 5 presents extensive simulation results comparing NextStep and the usual approach à la Young/Daly. Finally, we conclude in Section 6.

2 Related work

2.1 Checkpointing preemptible parallel jobs

Checkpoint-restart is one of the most widely used strategy to deal with fail-stop errors. Several variants of this policy have been studied; see [22] for an overview. The natural strategy is to checkpoint periodically, and one must decide how often to checkpoint, i.e., derive the optimal checkpointing period. An optimal strategy is defined as a strategy that minimizes the expectation of the execution time of the job. For a preemptible job, where one can checkpoint at any time, the classical formula due to Young [43] and Daly [11] states that the optimal checkpointing period is $W_{YD} = \sqrt{2\mu C}$, where μ is the job MTBF and C the checkpoint cost. This formula is a first-order approximation. For memoryless failures, Daly provides a second-order, more accurate, approximation in [11], while our previous work [7] provides the optimal value; both [11] and [7] use the Lambert function. The derivation in [7] is based on Equation (1) (see Section 3.2), a formula rediscovered ten years later, with a quite different proof based on a Markov model, in [35].

As explained in Section 3.6, non-memoryless failures are more difficult to deal with for parallel jobs. Several papers study non-periodic checkpointing strategies, with and without partial rejuvenation [29, 24, 31]. A recent paper [28] uses full rejuvenation while [38] wrongly assumes IID failures for a range a classic distributions, including Weibull and LogNormal, which are not memoryless (see Section 3.6). An unorthodox approach is used in [17], where it is assumed that the failures striking the whole platform obey a Weibull distribution; this is misleading for two reasons: (i) it is not clear what is the failure distribution on each individual processor; and (ii) after one processor is struck by a failure and rejuvenated, the platform failure distribution does not remain Weibull (see a more detailed discussion in Section 3.6).

In order to deal with non-memoryless failures, the NextFailure problem is studied in [7], where the goal is to maximize the expected amount of work completed before the next failure. This problem is solved using a dynamic programming algorithm, and it is used as a solution to the initial problem of makespan minimization. Simulations are done with Exponential and Weibull laws, showing that the proposed algorithm outperforms existing solutions with Weibull distributions. In this paper, we propose to maximize the expected efficiency rather than the expected work, with our new NextStep heuristic.

This requires a much more subtle approach. Nevertheless, we succeed in proving that NextStep is asymptotically optimal.

2.2 Checkpointing task-based applications

Going beyond preemptible applications, some works have studied task-based applications, using a model where checkpointing is only possible right after the completion of a task. The problem is then to determine which tasks should be checkpointed. This problem has been solved for linear workflows (where the task graph is a simple linear chain) by Toueg and Babaoglu [41], using a dynamic programming algorithm. This algorithm was later extended in [6] to cope with both fail-stop and silent errors simultaneously. Another special case is that of a workflow whose dependence graph is arbitrary but whose tasks are parallel tasks that each executes on the whole platform. In other words, the tasks have to be serialized. The problem of ordering the tasks and placing checkpoints is proven NP-complete for simple join graphs in [1], which also introduces several heuristics. For general workflows, deciding which tasks to checkpoint has been shown #P-complete [19], but several heuristics are proposed in [20].

2.3 Extensions: multi-criteria, hierarchical checkpointing, independence

In addition to the minimization of the expected job makespan, other optimization criteria have been considered in the literature. I/O is a scarce resource on modern platforms, and several works aim at minimzation I/O volume while enforcing an efficient checkpoint for makespan [25, 23]. Similarly, energy-makespan bi-criteria optimization has been addressed in [15, 18]. To reduce I/O overhead, various two-level checkpointing protocols have been studied [36, 14]. Some authors have also generalized two-level checkpointing to account for an arbitrary number of levels [30, 4, 12, 5].

As for failure independence, the standard model assumes IID failure inter-arrival times, or IATs, on each node, with a common distribution $\mathcal{D}$. While it is reasonable to assume that IATs are identically distributed on a given processor, because the faulty node is rejuvenated (replaced by a spare) after each failure, it is very questionable to assume that IATs are independent across the platform. As for *temporal* dependence, it has been observed many times that when a failure occurs, it may trigger other failures that will strike different system components [21, 40, 3]. As an example, a failing cooling system may cause a series of successive crashes of different nodes. Also, an outstanding error in the file system will likely be followed by several others [34, 27]. As for *spatial* dependence, it is clear that the overheating of some node in a cabinet is quite likely to be followed by the overheating of neighbor nodes (which comes atop of a temporal dependence as well!) Bautista-Gomez et al. [3] have studied nine systems, and they report periods of high failure density in all of them. They call these periods *cascade failures*. This observation has led them to revisit the temporal failure independence assumption, and to design bi-periodic checkpointing

algorithms that use different periods in normal (failure-free) and degraded (with failure cascades) modes. [40] introduces a dynamic strategy called *lazy checkpointing* to adjust to changes in the failure rate. Another approach has been proposed in [2], using quantiles of consecutive IAT pairs.

3 Framework

This section overviews known results for checkpoint strategies. We cover uniprocessor and multi-processor applications, either with Exponential failure distributions or with arbitrary failure distributions. Beforehand, we detail the platform and job model.

3.1 Model

3.1.1 Platform and jobs

We consider a large parallel platform with m identical processors, or nodes. These nodes are subject to fail-stop errors, or failures. A failure interrupts the execution of the node and provokes the loss of its whole memory. We consider parallel jobs that can be checkpointed at any time. In scheduling terminology, the jobs are preemptible. Consider a parallel job running on several nodes: when one of these nodes is struck by a failure, the state of the application is lost, and execution must restart from scratch, unless a fault-tolerance mechanism has been deployed. The classical technique to deal with failures makes use of a checkpoint-restart mechanism: the state of the application is periodically checkpointed, i.e., all participating nodes take a checkpoint simultaneously. This is the standard coordinated checkpointing protocol, which is routinely used on large-scale platforms [9], where each node writes its share of application data to stable storage (checkpoint of duration C). When a failure occurs, the platform is unavailable during a downtime D, which is the time to enroll a spare processor that will replace the faulty processor [11, 22]. Then, all application nodes (including the spare) recover from the last valid checkpoint in a coordinated manner, reading the checkpoint file from stable storage (recovery of duration R). Finally, the execution is resumed from that point on, rather than starting again from scratch. Note that failures can strike during checkpoint and recovery, but not during downtime (otherwise we can include the downtime in the recovery time). When a failure hits a processor, that processor is replaced by a spare. This amounts to start anew with a fresh processor. In the terminology of stochastic processes, the faulty processor is rejuvenated. However, all the other processors are not rejuvenated: this would be infeasible due to the multidinious spares needed!

3.1.2 Failures

We assume that each node experiences failures whose inter-arrival times follow Independent and Identically Distributed (IID) random variables obeying an ar-

bitrary probability distribution $\mathcal{D}$. We only assume that $\mathcal{D}$ is continuous and of finite expectation and variance, a condition satisfied by all standard distributions. We let μ_{ind} denote the expectation of $\mathcal{D}$, also known as the individual processor MTBF. Even if each node has an MTBF of several years, large-scale parallel platforms are composed of so many nodes that they will experience several failures per day [16, 8]. Hence, a parallel job using a significant fraction of the platform will typically experience a failure every few hours.

3.1.3 Checkpointing strategies

Given a parallel job of length T_{base} (base time without checkpoints nor failures), the optimization problem is to decide when and how often to checkpoint in order to minimize the expected execution time of the job. The job is divided into N_c segments of length W_i, $1 \leq i \leq N_c$, each followed by a checkpoint of length C. Of course $\sum_{i=1}^{N_c} W_i = T_{base}$. Throughout the paper, we add a final checkpoint at the end of the last segment, e.g., to write final outputs to stable storage. Symmetrically, we add an initial recovery when re-executing the first segment of a job (e.g., to read inputs from stable storage) if it has been struck by a failure before completing the checkpoint. Adding a last checkpoint and an initial recovery brings symmetry and simplifies formulas, but it is not at all mandatory: see Section 3.4 for an extension relaxing either or both assumptions.

3.2 Uni-processor job, and $\mathcal{D}$ Exponential

This is the simplest case. Consider a job J executing on a single processor experiencing failures whose inter-arrival times follow an Exponential distribution $\mathcal{D} = \textsc{Exp}(\lambda)$ of parameter $\lambda > 0$, whose PDF (Probability Density Function) is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. The processor MTBF is $\mu_{ind} = \frac{1}{\lambda}$. The optimal checkpointing strategy, i.e., the strategy minimizing the expected execution time, can be derived as shown below.

Lemma 1. *The expected time $\mathbb{E}(W, C, R)$ to execute a segment of W seconds of work followed by a checkpoint of C seconds and with recovery cost R seconds is*

$$\mathbb{E}(W, C, R) = \left(\frac{1}{\lambda} + D\right) e^{\lambda R} \left(e^{\lambda(W+C)} - 1\right). \tag{1}$$

Lemma 1 comes from [7, Theorem 1]. It also applies when the segment is not followed by a checkpoint (take $C=0$). The *slowdown function* is defined as $f(W, C, R) = \frac{\mathbb{E}(W,C,R)}{W}$. We have the following properties:

Lemma 2. *The slowdown function $W \mapsto f(W, C, R)$ has a unique minimum W_{opt} that does not depend on R, is decreasing in the interval $[0, W_{opt}]$ and is increasing in the interval $[W_{opt}, \infty)$.*

Proof. Again, see [7, Theorem 1]. The exact value of W_{opt} is obtained using the Lambert W function, but a first-order approximation is the Young/Daly formula $W_{YD} = \sqrt{\frac{2C}{\lambda}}$. □

Lemma 2 shows that infinite jobs should be partitioned into segments of size W_{opt} followed by a checkpoint. What about finite jobs? Back to our job J of duration T_{base}, we partition it into N_c segments of length W_i, $1 \leq i \leq N_c$, each followed by a checkpoint C. By linearity of the expectation, the expected time to execute job J is

$$\mathbb{E}(J) = \sum_{i=1}^{N_c} \mathbb{E}(W_i, C, R) = \left(\frac{1}{\lambda} + D\right) e^{\lambda R} \sum_{i=1}^{N_c} \left(e^{\lambda(W_i+C)} - 1\right),$$

where $\sum_{i=1}^{N_c} W_i = T_{base}$. By convexity of the Exponential function, or by using Lagrange multipliers, we see that $\mathbb{E}(J)$ is minimized when the W_i's take a constant value, i.e., all segments have same length. Thus, we obtain $W_i = \frac{T_{base}}{N_c}$ for all i, and we aim at finding N_c that minimizes

$$\mathbb{E}(J) = N_c\, \mathbb{E}\left(\frac{T_{base}}{N_c}, C, R\right) = f\left(\frac{T_{base}}{N_c}, C, R\right) \times T_{base}\ , \tag{2}$$

where f is the slowdown function. We let $N_{opt} = \frac{T_{base}}{W_{opt}}$, where W_{opt} achieves the minimum of the slowdown function. N_{opt} would be the optimal number of segments if we could have a non-integer number of segments. Lemma 2 shows that the optimal value N_{ME} of N_c is either $N_{ME} = \max(1, \lfloor N_{opt} \rfloor)$ or $N_{ME} = \lceil N_{opt} \rceil$, whichever leads to the smallest value of $\mathbb{E}(J)$. In the experiments, we use the simplified Young/Daly expression $N_{ME} = \left\lceil \frac{T_{base}}{W_{YD}} \right\rceil$.

3.3 Parallel job, and $\mathcal{D}$ Exponential

Because of the memoryless property of the Exponential distribution, the multi-processor case can be reduced to the uni-processor case. Everything holds by replacing the parameter λ by $p\lambda$, where p is the number of processors of job J. To see this formally, say the job J is executed on p processors $\{P_q\}_{1 \leq q \leq p}$. Let $X_i^q \sim \textsc{Exp}(\lambda)$, $i \geq 1$, $1 \leq q \leq p$, denote the IID fault inter-arrival times on processor P_q. Let Y_i, $i \geq 1$, denote the fault inter-arrival times for (the p processors executing) job J.

The assumption $X_i^q \sim \textsc{Exp}(\lambda)$ formally means that when processor P_q is rejuvenated (or when it is used for the first time), the next failure will strike according to a distribution $\textsc{Exp}(\lambda)$. If the job starts at time t, and the last failure stroke at time $t_1 < t$, what is the distribution of the probability of the next failure, knowing that P_q has been alive for $t - t_1$ seconds? The memoryless property of Exponential distributions is the key: it is still the same Exponential distribution. To keep notations simple, we let $X_i^q \sim \textsc{Exp}(\lambda)$, $i \geq 0$, denote the failure inter-arrival times on P_q once the job has started (and similarly for Y_i, $i \geq 0$).

First, we have $Y_1 = \min_q(X_1^q)$. Hence $Y_1 \sim \textsc{Exp}(p\lambda)$ (minimum of p $\textsc{Exp}(\lambda)$ distributions). Assume that the first failure for job J stroke at time t_2 (hence $Y_1 = t_2 - t$) on some processor P_{q_0}, which is then rejuvenated. Because of the memoryless property, knowing this information does not change the distribution of the next failure on any other processor, and $Y_2 \sim \textsc{Exp}(p\lambda)$ for the very same reason that $Y_1 \sim \textsc{Exp}(p\lambda)$.

3.4 Extension without final checkpoint nor initial recovery, and $\mathcal{D}$ Exponential

Consider a parallel job J with p processors, which is partitioned into segments. This section deals with the case where no checkpoint is enforced at the end of the last segment. By symmetry, we also deal with the case where no recovery is paid when re-executing the first segment after a failure. Let $p\lambda$ denote the failure rate for job J, assuming that job failures obey an Exponential law $\mathcal{D} = \textsc{Exp}(p\lambda)$.

The job is partitioned into N segments of length W_i, with checkpoint cost C_i and recovery cost R_i. Let $C_{tot} = \sum_{i=1}^{N} C_i$ and $R_{tot} = \sum_{i=1}^{N} R_i$. In the model of Sections 3.2 and 3.3, we had $C_i = C$, $R_i = R$ for $1 \leq i \leq N$, $C_{tot} = NC$, and $R_{tot} = NR$. If no checkpoint is taken after the last segment, $C_N = 0$ and $C_{tot} = (N-1)C$. If no recovery is paid when re-executing the first segment, $R_1 = 0$ and $R_{tot} = (N-1)R$.

What is the optimal strategy to minimize the expected execution time $\mathbb{E}_N$ of the job? From Lemma 1, we have

$$\mathbb{E}_N = \sum_{i=1}^{N} \mathbb{E}(W_i, C_i, R_i) = \left(\frac{1}{p\lambda} + D\right) \sum_{i=1}^{N} e^{p\lambda R_i} \left(e^{p\lambda(W_i+C_i)} - 1\right), \tag{3}$$

where $\sum_{i=1}^{N} W_i = T_{base}$. Given N, $\mathbb{E}_N$ is minimized when the sum $\sum_{i=1}^{N} e^{p\lambda(W_i+C_i+R_i)}$ is minimized. By convexity of the Exponential function, or by using Lagrange multipliers, we see that $\mathbb{E}_N$ is minimized when the $W_i+C_i+R_i$'s take a constant value W_{seg}. This value is given by

$$NW_{seg} = T_{base} + C_{tot} + R_{tot}, \tag{4}$$

and the length W_i of each segment is then computed as $W_i = W_{seg} - C_i - R_i$. If $C_N = 0$, the last segment involves an additional amount C of work duration. Similarly, if $R_1 = 0$, the first segment involves an additional amount R of work duration. Then, we can derive the optimal value of N and W_{seg} as follows: Equation (4) gives $N = \frac{T_{base}-R-C+R_1+C_N}{W_{seg}-R-C}$. Plugging this value into

$$\mathbb{E}_N = \left(\frac{1}{p\lambda} + D\right) \left[(N-1)e^{p\lambda R} + e^{p\lambda R_1} + Ne^{p\lambda W_{seg}}\right]$$

enables to solve for W_{seg}, using the Lambert function in a similar way as in [7].

While the derivation is painful, the conclusion is comforting: in the optimal solution, all segments have same length of work, up to an additional recovery for the first segment and an additional checkpoint for the last one. The Young/Daly approximation still holds, as well as all the results of this paper (whose presentation is much simpler with all segments followed by a checkpoint).

3.5 Uni-processor job, and $\mathcal{D}$ arbitrary

When failures inter-arrival times obey an arbitrary distribution $\mathcal{D}$, they are still IID, because the processor is rejuvenated (replaced by a spare) after each

failure. To the best of our knowledge, even the optimal value W_{opt} for the slowdown function is not known analytically. For some distributions, W_{opt} can be computed numerically, using partial moments for the expectation of the time lost due to failures. But note that W_{opt} does no give the optimal checkpointing period for infinite jobs, contrarily to the memoryless case. In fact, the optimal checkpointing strategy is not known for infinite jobs, let alone for finite jobs.

For instance, consider a Weibull distribution $\mathcal{D} \sim \textsc{Weibull}(k, \lambda)$ of shape k and scale λ; its PDF is $\mathbb{P}(X = t) = \frac{k}{\lambda}(\frac{t}{\lambda})^{k-1}e^{-(\frac{t}{\lambda})^k}$ for $t > 0$. If $k < 1$, the instantaneous failure rate of $\mathcal{D}$ is decreasing with time (infant mortality), checkpoints should be spaced more and more as time progresses since the last failure. On the contrary, if $k > 1$, the instantaneous failure rate of $\mathcal{D}$ is increasing with time (ageing) and, hence, checkpoints should be spaced less and less. This explains that the optimal checkpointing strategy is aperiodic. See [29, 24, 31] for more details.

3.6 Parallel job, and $\mathcal{D}$ arbitrary

When $\mathcal{D}$ is arbitrary, even though the failure inter-arrival times X_i^q are IID on each processor, they are not for (the p processors executing) job J. In other words, the Y_i are not IID, unless $\mathcal{D}$ is Exponential. However, owing to the theory on the superposition of renewal processes, whenever $\mathcal{D}$ is continuous and of finite expectation μ_{ind}, we know that the following limit exists:

$$\lim_{n\to\infty} \mathbb{E}\left(\frac{\sum_{i=1}^{n} Y_i}{n}\right) = \frac{\mu_{ind}}{p}. \tag{5}$$

This result is given as Formula 1.4 in [26]. See also [22] for an elementary proof using Wald's equation. Equation (5) is good news because we can define the job MTBF as $\frac{\mu_{ind}}{p}$: in average, a failure will strike the job every $\frac{\mu_{ind}}{p}$ seconds. Note that the MTBF is given a new definition here: the failures striking the parallel job J are not IID, so there is no longer a mean time before the next failure of the job. Instead, there is a limit on the average time between failures. At any time, the distribution of the next failure is complicated because it must account for the history of the $p-1$ processors that have not been rejuvenated when the last failure stroke. Indeed, assume that the execution of job J was started on p fresh processors $\{P_q\}_{1\leq q\leq p}$, and that the last failure stroke on processor P_q at time t_q (where $t_q = 0$ if P_q has never been struck). Let $i(q)$ be the index of the last failure on P_q (where $i(q) = 0$ if P_q has never been struck). To simplify notations, say that the last failure stroke processor P_1, meaning that $t_q < t_1$ for $q \geq 2$. The probability that the next failure strikes at time t on P_q (it will be failure number $i(q) + 1$) is

$$\mathbb{P}(X^q_{i(q)+1} = t | X^q_{i(q)+1} \geq t - t_q).$$

In other words, only X_1^q follows the distribution $\mathcal{D}$, while each X_i^q, $q \geq 2$, is shifted. To compute the distribution of the next failure of job J, we need to

compute the distribution of the minimum of all the $X^q_{i(q)+1}$'s, which are not identical because of their history.

There is a theoretical approach that simplifies the problem, namely rejuvenating all the p processors of the job after each failure (and before starting the execution of the job). Of course, this is impossible in practice when p exceeds a small number, but it is nice from a theory perspective: with total rejuvenation, each failure becomes a renewal point for the whole job, and the failure inter-arrival times that strike the job become IID: their distribution is the minimum of p IID distributions $\mathcal{D}$. Even better, there are a few failure distributions $\mathcal{D}$ such that the minimum of p IID instances also obey the same distribution $\mathcal{D}$ (with scaled parameters). For instance, consider a Weibull distribution $\mathcal{D} \sim \textsc{Weibull}(k, \lambda)$ of shape k and scale λ, whose expectation is $\mu_{ind} = \lambda\Gamma(1 + \frac{1}{k})$, where Γ denotes the Gamma function $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$ for $t > 0$. Then the minimum Y of p IID $\textsc{Weibull}(k, \lambda)$ is also a Weibull distribution $Y \sim \textsc{Weibull}(k, \frac{\lambda}{p^{1/k}})$. We observe that the MTBF does not scale linearly with p, unless $k = 1$: the expectation of Y is $\mu = \frac{\mu_{ind}}{p^{1/k}}$. This discussion explains the errors in [17]: (i) the platform cannot obey a Weibull distribution, unless total rejuvenation is used; and (ii) assuming total rejuvenation, the MTBF of a job is not inversely proportional to its number of processors.

A realistic approach to cope with the not-IID problem is to discretize time into small quanta, and to use dynamic programming to compute the best checkpoint strategy for job J up to the next failure [7]. Obviously, the smaller the quanta, the more accurate the results, but the more costly the dynamic programming algorithm. The approach in [7] greedily uses this strategy from one failure to another, up to the completion of the job. However, optimizing checkpoints up to the next failure, while optimal from one failure to the next (up to the precision of the quanta), may well be sub-optimal for the whole job. A main contribution of this paper is to introduce a new greedy strategy and to prove an approximation bound for its performance. To the best of our knowledge, this is the first theoretical result for parallel jobs with non-memoryless failures.

4 The NextStep heuristic

In this section, we present the NextStep heuristic to checkpoint parallel jobs under any failure probability distribution. The main idea of NextStep is the following: after each failure, NextStep is able to find the checkpointing strategy that maximizes the expected efficiency (see below) before the next failure or the end of the job. Intuitively, optimizing the expected efficiency on a *failure-by-failure* basis should lead to a good approximation on the optimal solution, at least for large job sizes. One major contribution of this work is to show that NextStep is asymptotically optimal for arbitrary failure distributions.

We first introduce notations in Section 4.1, before formally describing NextStep in Section 4.2. Finally, we prove the asymptotic optimality in Section 4.3.

4.1 Preliminaries

Consider a parallel job J of length T_{base} executing on p processors, with checkpoint time C. Assume that the job just experienced a failure, and it is ready to resume execution of the remaining W work units (or the job is just starting, and then $W = T_{base}$). For any processor P_j, $1 \leq j \leq p$, let τ_j be the time elapsed since its last failure. In particular, if P_j has been hit by the last failure, then $\tau_j = 0$; note also that we do not assume fresh processors when starting the job. We call $\vec{\tau} = (\tau_1, \tau_2, \ldots, \tau_p)$ the *history* vector.

Given a checkpointing strategy, a job with W remaining work units and a history vector $\vec{\tau}$, the function $\mathit{first}(W|\vec{\tau})$ returns the size W_1 of the segment preceding the first checkpoint.

Work. Let $\mathcal{W}(W|\vec{\tau})$ be the random variable that quantifies the amount of work successfully executed before the next failure. We have the following recursion:

$$\begin{aligned} \mathcal{W}(0|\vec{\tau}) &= 0 \\ \mathcal{W}(W|\vec{\tau}) &= \begin{cases} W_1 + \mathcal{W}(W - W_1|\vec{\tau} + \overrightarrow{W_1 + C}) \\ \quad \text{if the processor does not fail during} \\ \quad \text{the next } W_1 + C \text{ units of time,} \\ 0 \quad \text{otherwise.} \end{cases} \end{aligned} \tag{6}$$

In Equation (6), given a scalar quantity x, $\overrightarrow{x} = (x, x, \ldots, x)$ denotes the vector with p identical components equal to x. Weighting the two cases in Equation (6) by their probabilities of occurrence, we obtain the expected amount of work successfully computed before the next failure:

$$\mathbb{E}(\mathcal{W}(W|\vec{\tau})) = \mathbb{P}_{suc}(W_1 + C|\vec{\tau})(W_1 + \mathbb{E}(\mathcal{W}(W - W_1|\vec{\tau} + \overrightarrow{W_1 + C}))), \tag{7}$$

where the probability of success $\mathbb{P}_{suc}$ is computed as

$$\mathbb{P}_{suc}(x|\vec{\tau}) = \prod_{i=1}^{p} \mathbb{P}(X \geq x + \tau_i | X \geq \tau_i). \tag{8}$$

X is a generic random variable following the probability distribution $\mathcal{D}$, the failure inter-arrival time on each processor. Given any such distribution $\mathcal{D}$, $\mathbb{P}_{suc}(x|\vec{\tau})$ can be computed in time $O(p)$.

Efficiency. Rather than focusing solely on the work done, we aim at maximizing the expected efficiency, which also depends on the number of checkpoints that have been taken. This is particularly crucial at the end of the job, where maximizing the amount of work until the next failure may not be the best strategy if the job is about to complete. Indeed, the efficiency also depends on the time spent computing; if no failures occur, it depends on the number of checkpoints that are taken. Hence, we define $\mathbb{E}_W(W, \vec{\tau}, N_c)$ as the maximum expected work until the next failure (or the completion of the job if no failure occurs) using N_c checkpoints; similarly, $\mathbb{E}_{T_{next}}(W, \vec{\tau}, N_c)$ is the expected time until the next failure, or before the completion of the job if no failure occurs. Note that

the number N_c of checkpoints only matters in the latter case where the job has completed.

Finally, if a job still needs to be processed for W units of time, with a history $\vec{\tau}$, we define the maximum possible efficiency among all possible numbers of checkpoints N_c:

$$\mathbb{E}_e(W, \vec{\tau}) = \max_{N_c} \frac{\mathbb{E}_W(W, \vec{\tau}, N_c)}{\mathbb{E}_{T_{next}}(W, \vec{\tau}, N_c)}. \tag{9}$$

Time discretization. We introduce a time quantum u, and discretize time into quanta. This means that all quantities (segment sizes, checkpoint and recovery times) are integer multiples of u, and that failures strike at the end of a quantum. More precisely, the probability that a failure happens at the end of quantum i is $\int_{(i-1)u}^{iu} f(x|\vec{\tau})dx$, where $f(x|\vec{\tau})$ is the probability density function of the platform failure distribution $\mathcal{D}$ in the continuous case conditioned by the history. This discretization restricts the search for an optimal execution to a finite set of possible executions. The trade-off is that a smaller value of u leads to a more accurate solution, but also to a higher number of states in the algorithm, hence, to a higher computing time.

In what follows, if a variable y is defined in seconds (or work units), $y^* = y/u$ is the corresponding number of quanta, which we always suppose integer. For instance, the job size becomes $W^* = W/u$ and the checkpoint size $C^* = C/u$. Similarly, we let $\mathbb{P}^*_{suc}(x^*|\vec{\tau^*}) \triangleq \mathbb{P}_{suc}(ux|u\vec{\tau})$ be the probability that the next x^* quanta succeed given the history $\vec{\tau^*}$, expressed in number of quanta.

4.2 NextStep

We define NextStep formally as: find a function returning the size of the first segment to be checkpointed, such that $\mathbb{E}_e(W, \vec{\tau_0})$ is maximized. Here, $\vec{\tau_0}$ corresponds to the initial history of the platform when the execution starts. Solving NextStep analytically seems out of reach, but the recursive definition of $\mathbb{E}(\mathcal{W}(W|\vec{\tau}))$ (see Equation (7)), together with time discretization, allows us to compute the maximum efficiency. Indeed, there is no need to keep the time elapsed since the last failure of each processor as a parameter of the recursive calls. This is because the τ variables of all processors evolve identically: recursive calls only correspond to cases where no failure has occurred. The algorithm is called again each time a failure occurs, to decide where checkpoints should be taken.

Thanks to the discretization, all the $\mathbb{E}_W(W, \vec{\tau_0}, N_c)$ values can be computed with a time quantum u. We let x^* be the number of quanta that remain to proceed (where initially, $x^* = W^*$). We need to find and store the best solutions for any possible values of x^* and N_c in the recursive call. Hence, we further consider N_p, the number of checkpoints already taken, and N_f, the number of checkpoints that can still be taken (where $N_p + N_f = N_c$). This corresponds to Algorithm 1: the *compE* procedure fills a table *solve* that contains, for any triple (x^*, N_p, N_f), the maximum expected work until the next failure for these parameters, and the best segment size W_1^* that achieves this. For $N_f = 1$, the

Algorithm 1: $compE(x^*,N_p,N_f)$

```
1  if x* = 0 then return 0;
2  if N_f = 1 then
3      τ* ← τ*_0 + overrightarrow(W* − x* + N_p C*);
4      best ← x* P*_suc(x* + C* | τ*) ;
5      solve[x*][N_p][N_f] ← (best, x*);
6      return best;
7  if solve[x*][N_p][N_f] = (best, W*_1) then return best;
8  else
9      best ← −∞;
10     τ* ← τ*_0 + overrightarrow(W* − x* + N_p C*);
11     for i = 1 to x* do
12         work ← compE(x* − i, N_p + 1, N_f − 1);
13         cur ← P*_suc(i + C* | τ*) × (i + work);
14         if cur > best then
15             best ← cur;
16             W*_1 ← i;
17     solve[x*][N_p][N_f] ← (best, W*_1);
18     return best;
```

only possibility is to compute x^* in its entirety and then checkpoint. Otherwise, we try all possible places for the first checkpoint, and recursively call *compE*. If a value with a given (x^*, N_p, N_f) had been computed before, we retrieve the corresponding result line 7.

There remains to compute $\mathbb{E}_{T_{next}}(W^*, \vec{\tau^*}, N_c)$, i.e., the expected time until next failure or job completion. The following lemma helps us compute these values efficiently with discrete segments:

Lemma 3. *Using discrete quanta of size u, the expectation of the time before the next failure or the completion of the job, expressed in number of quanta, is the following:*

$$\mathbb{E}_{T_{next}}(W^*, \vec{\tau_0^*}, N_c) = \sum_{i=0}^{W^*+N_cC^*-1} \mathbb{P}^*_{suc}(i|\vec{\tau_0^*}).$$

Proof. Let X denote the random variable of the number of quanta executed before the next failure (or the completion of the job) given the history $\vec{\tau_0^*}$, the total number of quanta of the job W^* and the number of checkpoints N_c. Clearly, X is taking integer values in $[1, W^* + N_cC^*]$, thus

$$\begin{aligned}\mathbb{E}(X) &= \sum_{i=1}^{W^*+N_cC^*} i\mathbb{P}\{X = i\} = \sum_{i=1}^{W^*+N_cC^*} \mathbb{P}\{X \geq i\} \\ &= \sum_{i=1}^{W^*+N_cC^*} \mathbb{P}^*_{suc}(i-1|\vec{\tau_0^*}) = \sum_{i=0}^{W^*+N_cC^*-1} \mathbb{P}^*_{suc}(i|\vec{\tau_0^*}).\end{aligned}$$

Algorithm 2: NEXTSTEP (W^*)

```
   /* Compute E_{T_next}(W*, τ⃗*_0, n_c) for n_c ∈ [1, W*]                    */
 1 S ← 0;
 2 for i = 0 to W* − 1 do S ← S + P*_suc(i|τ⃗*_0);
 3 for n_c = 1 to W* do
 4 |   for i = 1 to C* do
 5 |   |   S ← S + P*_suc(W* + (n_c − 1)C* + i|τ⃗*_0);
 6 |   E_{T_next}(W*, τ⃗*_0, n_c) ← S;
   /* Compute E_W(W*, τ⃗*_0, n_c) (array solve)                              */
 7 for n_c = 1 to W* do compE(W*, 0, n_c);
   /* Solution of NEXTSTEP                                                   */
 8 best ← −∞; N_c ← 0; W*_1 ← 0;
 9 for n_c = 1 to W* do
10 |   cur ← first(solve[W*][0][n_c])/E_{T_next}(W*, τ⃗*_0, n_c);
11 |   cursegment ← second(solve[W*][0][n_c]);
12 |   if cur > best then
13 |   |   best ← cur; N_c ← n_c; W*_1 ← cursegment;
14 return (N_c, W*_1);
```

□

From Lemma 3, we derive that $\mathbb{E}_{T_{next}}(W^*, \vec{\tau_0^*}, n_c+1) = \mathbb{E}_{T_{next}}(W^*, \vec{\tau_0^*}, n_c) + \sum_{i=W^*+n_cC^*}^{W^*+(n_c+1)C^*-1} \mathbb{P}^*_{suc}(i|\vec{\tau_0^*})$. This is used in Algorithm 2 to compute all the $\mathbb{E}_{T_{next}}$ values more efficiently on lines 1–6. Algorithm 1 is called to fill the *solve* table with all values of $\mathbb{E}_W$ on line 7. We obtain the efficiency $\mathbb{E}_e(W^*, \vec{\tau_0^*})$ for all possible number of checkpoints and keep the maximum, see lines 8–13. Finally, the algorithm returns the values for N_c and W_1^* corresponding to the maximum efficiency, which allows us to rebuild completely the corresponding solution using the table *solve*.

Proposition 1. *Using a time quantum u, and for any failure inter-arrival time distribution, Algorithm 2 computes the solution to* NEXTSTEP *(maximizing efficiency) in time $O(p(W^*)^4)$.*

Proof. The NEXTSTEP algorithm consists of three steps. In the first step, it computes all values of $\mathbb{E}_{T_{next}}(W^*, \vec{\tau_0^*}, n_c)$ for $n_c \in [1, W^*]$. To do so, two loops are executed. The first one has W^* steps, where each step computes a single addition. The value of $\mathbb{P}^*_{suc}(W^* + (n_c-1)C^* + i|\vec{\tau_0^*})$ is the product of the individual probability of failure of the p processors (as in Equation (8)). We assume that the individual probabilities of failure can be computed in $O(1)$, thus the loop takes a time $O(pW^*)$. The second loop is similar, with two nested loops, and its complexity is $O(pW^*C^*)$. We can safely assume that $C^* \leq W^*$, otherwise doing any checkpoint is straightforwardly bad (if we succeed the checkpoint, we would have succeeded the entire job), and the complexity is at most $O(p(W^*)^2)$.

In the second step, the algorithm fills the table *solve* and calls $compE(W^*, 0, n_c)$ for $n_c \in [1, W^*]$. The function $compE(x, y, z)$ fills the table entry corresponding to its parameters if necessary, with eventual recursive calls to *compE* where $y + z$ is constant and x decreases. Given the initial calls with $x = W^*$ and $y + z \in [1, W^*]$ the number of entries written in the table is at most $\frac{W^{*3}}{2}$. To upper bound the overall complexity of this step, we first note that an entry may only be written in the table if the *compE* function is called with the same parameters. In the sub-case $N_f = 1$, this takes few operations and involves a call to $\mathbb{P}^*_{suc}$, thus a time $O(p)$. Otherwise, this means *compE* has been called with parameters corresponding to the last sub-case. If so, a loop is executed $x^* \leq W^*$ times, and each iteration requires a call to *compE*, which either takes time $O(1)$ or fills another entry of the table (therefore the complexity is taken into account for this other entry), and the other operations are in $O(1)$, except the computation of $\mathbb{P}^*_{suc}$ in $O(p)$. Overall, the individual cost of each entry of the table is at most in $O(pW^*)$. Finally the calls to *compE* that do not fill the table may only be made recursively and were taken into account in the analysis. Given the size of the table in $O((W^*)^3)$, this second step is in $O(p(W^*)^4)$.

Finally, the last step returning the solution consists of a loop with W^* iterations, where each iteration is done in $O(1)$, which gives a complexity in $O(W^*)$. The complexity is dominated by the second step; hence, the result. □

4.3 Asymptotic analysis

We proceed in two steps. First, we show that for an infinite job and for any integer n, the expected work completed by NextStep before the n-th failure (which happens at a random time) is larger than or equal to the one of any other strategy. Then, we show that the expected work processed within T units of time is asymptotically optimal with T (Theorem 1).

4.3.1 Expected work completed before the n-th failure

We compare NextStep to any other strategy for an infinite job that has an infinite number of failures. We assume that the job starts at time $t_0 = 0$. For $k \geq 1$, let t_k be the random variable representing the date of the k-th failure, and for $k \geq 0$, let $\vec{\tau}_k$ be the random variable representing the history of the machine at time t_k. Note that neither t_k nor $\vec{\tau}_k$ depend on the checkpointing strategy. For any $n \geq 1$, let WF_n be the random variable corresponding to the expected work executed from the start up to failure n by NextStep, and let OPT_n denote the same variable for an optimal strategy, both depending on the initial history $\vec{\tau}_0$. We show that $\mathbb{E}(WF_n) = \mathbb{E}(OPT_n)$.

We define wf_k (resp. opt_k) the random variable representing the work executed by NextStep (resp. an optimal strategy) between times t_{k-1} and t_k. At any decision point, i.e., after each failure, the expected time before the next failure or the completion is the expected time before the next failure, because the job is infinite; hence, this time does not depend upon the number of checkpoints. From Equation (9), we deduce that, for any history, NextStep maximizes the

expected work accomplished before the next failure. Hence, we have for any $k \geq 1$ and any possible history $\vec{\tau}$ at the $(k-1)$-th failure (or $\vec{\tau} = \vec{\tau_0}$ if $k = 1$):

$$\mathbb{E}(wf_k | \vec{\tau}_{k-1} = \vec{\tau}) \geq \mathbb{E}(opt_k | \vec{\tau}_{k-1} = \vec{\tau}).$$

This inequality holds for any possible history $\vec{\tau}$, i.e., for any possible event $\{\vec{\tau}_{k-1} = \vec{\tau}\}$, and $\vec{\tau}_{k-1}$ does not depend on the strategy. Therefore, this inequality can be directly extended to $\mathbb{E}(wf_k | \vec{\tau}_{k-1})$ and $\mathbb{E}(opt_k | \vec{\tau}_{k-1})$. Note that these expectations are conditioned by a random variable instead of an event, thus are random variables themselves (constant for $k = 1$ if we consider $\vec{\tau_0}$ as a constant random variable) which always verify $\mathbb{E}(wf_k | \vec{\tau}_{k-1}) \geq \mathbb{E}(opt_k | \vec{\tau}_{k-1})$. In particular:

$$\mathbb{E}(\mathbb{E}(wf_k | \vec{\tau}_{k-1})) \geq \mathbb{E}(\mathbb{E}(opt_k | \vec{\tau}_{k-1})). \tag{10}$$

This result can be combined with the property $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$ (Law of Total Expectation) whenever both sides exist [37, p. 179] to obtain, for all $k \geq 1$,

$$\mathbb{E}(wf_k) = \mathbb{E}(\mathbb{E}(wf_k | \vec{\tau}_{k-1})) \geq \mathbb{E}(\mathbb{E}(opt_k | \vec{\tau}_{k-1})) = \mathbb{E}(opt_k).$$

Finally, we obtain:

$$\begin{aligned} \mathbb{E}(WF_n) &= \mathbb{E}\left(\sum_{k=1}^{n} wf_k\right) = \sum_{k=1}^{n} \mathbb{E}(wf_k) \\ &\geq \sum_{k=1}^{n} \mathbb{E}(opt_k) = \mathbb{E}\left(\sum_{k=1}^{n} opt_k\right) = \mathbb{E}(OPT_n). \end{aligned}$$

This shows that the expected work completed by NextStep before the n-th failure is larger than or equal to the one of any other strategy.

4.3.2 NextStep is asymptotically optimal

For any T, we show how to define $n(T)$, the index of a failure striking at a time close enough to T, so that the relative work difference performed between T and $t_{n(T)}$ (whichever comes first) is negligible:

Lemma 4. *Let* $n(T) = p\lfloor \frac{T}{\mathbb{E}(X)} \rfloor$, *then* $\lim_{T \to \infty} \frac{\mathbb{E}(|T - t_{n(T)}|)}{T} = 0$.

Proof. Consider an infinitely long job executing on p processors P_i, $1 \leq i \leq p$. Let X denote the random variable for failure inter-arrival times on each processor if there is no history. For $T > 0$, we fix $K(T) = \left\lfloor \frac{T}{\mathbb{E}(X)} \right\rfloor$, thus $n(T) = pK(T)$.

Let $t_{i,j}$ be the random variable representing the time when processor P_i fails for the j-th time. Clearly, for all i and $k > 0$, $t_{i,k+1} - t_{i,k}$ follows the distribution X, because P_i is rejuvenated after each failure. Therefore, $\mathbb{E}(t_{i,j}) = \mathbb{E}(X_{i,0}) + (j-1)\mathbb{E}(X)$, where $\mathbb{E}(X_{i,0})$ depends on the initial state of processor

P_i. We then use a variant of the strong law of large numbers [10, Ex. 8 p. 137]: If $(X_1, X_2, \ldots, X_j)$ are identically distributed with finite expectations and $S_j = \sum_{k=1}^{j} X_k$, then $\frac{S_j}{j} \to \mathbb{E}(X_1)$ in L^1, i.e., $\lim_{j\to\infty} \mathbb{E}\left(\left|\frac{S_j}{j} - \mathbb{E}(X_1)\right|\right) = 0$. Applying this result with $X_k = t_{i,k+1} - t_{i,k}$, we obtain $S_{j-1} = t_{i,j} - t_{i,1}$ and

$$\forall i, \lim_{j\to\infty} \mathbb{E}\left(\left|\frac{t_{i,j} - t_{i,1}}{j-1} - \mathbb{E}(X)\right|\right) = 0.$$

For any given j, since the $t_{i,j} - t_{i,1}$'s are identically distributed for all i, we can define a function $\epsilon(j)$ verifying $\lim_{j\to\infty} \epsilon(j) = 0$ and such that, using triangular inequalities:

$$\mathbb{E}(|t_{i,j} - j\mathbb{E}(X)|) \leq j\epsilon(j) + \mathbb{E}(X) + t_{i,1}$$

for all i and j. Finally, $\min_{1\leq i\leq p} t_{i,K(T)} \leq t_{n(T)} \leq \max_{1\leq i\leq p} t_{i,K(T)}$, because the total number of failures is the sum of the number of failures of each processor and $n(T) = pK(T)$. Hence,

$$\begin{aligned}
\mathbb{E}(|t_{n(T)} - K(T)\mathbb{E}(X)|) &\leq \mathbb{E}(\max_{1\leq i\leq p}(|t_{i,K(T)} - K(T)\mathbb{E}(X)|)) \\
&\leq \mathbb{E}(\sum_{i=1}^{p}(|t_{i,K(T)} - K(T)\mathbb{E}(X)|)) \\
&\leq pK(T)\epsilon(K(T)) + p\mathbb{E}(X) + \sum_{i=1}^{p} t_{i,1}.
\end{aligned}$$

By definition, $K(T)\mathbb{E}(X) \leq T \leq (K(T)+1)\mathbb{E}(X)$, and, because of the triangular inequality, we have:

$$\begin{aligned}
\frac{\mathbb{E}(|T - t_{n(T)}|)}{T} &= \frac{\mathbb{E}(|(T - K(T)\mathbb{E}(X)) + (K(T)\mathbb{E}(X) - t_{n(T)})|)}{T} \\
&\leq \frac{\mathbb{E}(X)}{K(T)\mathbb{E}(X)} + \frac{K(T)p\epsilon K(T)}{K(T)\mathbb{E}(X)} + \frac{p\mathbb{E}(X)}{K(T)\mathbb{E}(X)} + \frac{\sum_{i=1}^{p} t_{i,1}}{K(T)\mathbb{E}(X)} \\
&= \frac{1}{K(T)} + \frac{p}{\mathbb{E}(X)}\epsilon(K(T)) + \frac{p}{K(T)} + \frac{\sum_{i=1}^{p} t_{i,1}}{K(T)\mathbb{E}(X)}.
\end{aligned}$$

Here, p, $\mathbb{E}(X)$, and $\sum_{i=1}^{p} t_{i,1}$ are fixed, while $\lim_{T\to\infty} K(T) = \infty$. Hence the result, using $\lim_{K(T)\to\infty} \epsilon(K(T)) = 0$. □

Theorem 1. *For any T, with $n(T) = p\lfloor\frac{T}{\mathbb{E}(X)}\rfloor$, let $w^*_{n(T)}$ be the optimal expected work done up to time $t_{n(T)}$ (from the start to the $n(T)$-th failure). Then, for any checkpointing strategy, we have $\frac{\mathbb{E}_W([0,T])}{T} \leq \frac{w^*_{n(T)}}{T} + o(1)$. Furthermore, with* NextStep*, we have $\frac{\mathbb{E}_W([0,T])}{T} \geq \frac{w^*_{n(T)}}{T} - o(1)$. Hence,* NextStep *is asymptotically optimal.*

Proof. Assuming $\mathbb{E}_W([a,b]) = 0$ if $a > b$, thanks to Lemma 4, we obtain that, for any strategy,

$$\begin{aligned}\frac{\mathbb{E}_W(T)}{T} &\leq \frac{\mathbb{E}_W([0,t_{n(T)}])}{T} + \frac{\mathbb{E}_W([t_{n(T)},T])}{T} \\ &\leq \frac{w^*_{n(T)}}{T} + \frac{\mathbb{E}(|T-t_{n(T)}|)}{T} = \frac{w^*_{n(T)}}{T} + o(1).\end{aligned}$$

Furthermore, for NextStep, we have:

$$\begin{aligned}\frac{\mathbb{E}_W(T)}{T} &\geq \frac{\mathbb{E}_W([0,t_{n(T)}])}{T} - \frac{\mathbb{E}_W([T,t_{n(T)}])}{T} \\ &\geq \frac{w^*_{n(T)}}{T} - \frac{\mathbb{E}(|T-t_{n(T)}|)}{T} = \frac{w^*_{n(T)}}{T} - o(1),\end{aligned}$$

which concludes the proof. □

4.3.3 Counter-example to optimality

The NextStep heuristic is asymptotically optimal but not always optimal. This is because, for short jobs, maximizing the efficiency until the next failure is not exactly equivalent to minimizing the makespan. In Appendix C, we give an example where NextStep is not optimal, for an Exponential law. The example is designed as a worst-case scenario and shows that the number of checkpoints may differ between NextStep and the optimal.

4.3.4 A note on the optimal solution for an Exponential law

Sections 3.2 and 3.3 have shown how to statically compute the optimal strategy to minimize the expected makespan of a job when the failures obey an Exponential distribution. This optimal strategy is *static*, meaning that we compute the number and length of the job segments once and for all, before starting the execution. On the contrary, the NextStep strategy is dynamic, since it is called after each failure. One may envision a dynamic version of the optimal static strategy, where one would recompute the number and length of the job segments after each failure (and maybe after each checkpoint too), as a function of the remaining size of the job. We show in Appendix B that this dynamic approach is identical to the static one. This new result demonstrates the fairness of comparing NextStep with a static approach.

5 Performance Evaluation

In this section, we evaluate and compare the performance of NextStep with the Young/Daly periodic checkpointing heuristic, using simulations on synthetic jobs with various parameters, and subject to failures that are sampled from a wide range of probability distributions. Section 5.1 details job parameters and failure distributions. Section 5.2 presents all simulation results.

5.1 Simulation Setup

Algorithms

We compare the performance of NextStep with YoungDaly, the Young/Daly periodic checkpointing strategy (see Sections 3.2 and 3.3). For YoungDaly, a job J of length T_{base} and using p processors is divided into $N_{ME} = \left\lceil \frac{T_{base}}{W_{YD}} \right\rceil$ equal-size segments, each followed by a checkpoint. Here, $W_{YD} = \sqrt{\frac{2C\mu_{ind}}{p}}$. By default, we use $\mu_{ind} = 10$ years in the simulations.

Because YoungDaly is a *periodic* strategy, the size of its checkpointed segments are defined once and for all. On the contrary, NextStep adapts its checkpointing strategy to the failure history. Hence, after each failure, NextStep must recompute the size of its checkpointed segments. We take this recomputation time into account in the simulation, and conservatively add it to the recovery time[1]. To keep the recomputation time as low as possible, we introduce two optimizations to Algorithm 2. The goal is to dramatically shorten its execution time while maintaining the quality of the produced solution.

The first optimization is about the loop at Line 9. In Algorithm 2, the loop is over all possible numbers of checkpoints, ranging from a single checkpoint to one checkpoint per time quantum (i.e., W^* checkpoints). This latter solution would lead to a huge number of checkpoints. A natural conjecture is that the expected performance of NextStep would be a bell-shaped function of the number of checkpoints that are taken, first increasing and then decreasing after a threshold number has been reached. Therefore, in our implementation, we have replaced the `for` loop at Line 9 by a `while` loop that continues to look for a solution involving one additional checkpoint only if at least one of the five prior attempts leads to the best solution overall.

The second optimization is about the computation of the probability of success in Algorithm 1, Lines 4 and 13; and Algorithm 2, Line 2 and 5. This probability is the product of the individual probabilities of the processors. Hence, the execution time of this step is linear in the number of processors, while we want to consider platforms with tens of thousands of processors. Furthermore, this probability is computed many times, for many different values of i and n_c. We replace the exact computation by the following approximation. We first sort the values in τ_0^* and we retain the smallest ten and largest ten values; then we approximate the remaining values using 100 quantiles, according to the distribution. When i and n_c vary, they add an additive term to the history, which does not change the ranking of the values. We can thus replace the exact computation by one that uses the 10 smallest and 10 largest values of the history, and the 100 quantiles with their frequency of occurrences (if there are k values for a quantile, we compute a single probability and its exponentiation rather than k probabilities that we multiply). Hence, for a pre-processing cost of $O(p \log p)$ we approximate in constant time the probability of success, since the processors

[1] An alternative would be to perform this recomputation on dedicated resources, and in parallel to the recovery. We study the most costly scenario for NextStep.

that define the 120 history values remain the same up to the next failure.

Probability Distributions

We experiment with a wide range of probability distributions:
• **The Exponential distribution**, with probability density function $f(t) = \frac{e^{-t/\mu_{ind}}}{\mu_{ind}}$;
• **The Weibull distribution**, with probability density function $f(t;k,\lambda) = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1} e^{-(t/\lambda)^k}$, where k is the shape parameter and λ is the scale parameter. To obtain an MTBF of μ_{ind}, we chose $\lambda = \frac{\mu_{ind}}{\Gamma(1+\frac{1}{k})}$, and therefore the probability density function becomes $f(t,k) = \frac{k\Gamma(1+\frac{1}{k})}{\mu_{ind}}\left(\frac{t\Gamma(1+\frac{1}{k})}{\mu_{ind}}\right)^{k-1} e^{-\left(\frac{t\Gamma(1+\frac{1}{k})}{\mu_{ind}}\right)^k}$. In the experiments, k is varied in $\{0.5, 0.7, 1.5\}$. The first two shape values are realistic values taken from [32, 33, 15]; for $k < 1$, processors are more likely to fail if the processor is recent (infant mortality). The last shape value $k = 1.5$ provides an example of a distribution whose instantaneous failure rate increases with time. Note that $k = 1$ corresponds to the Exponential distribution;
• **The Gamma distribution**, with probability density function $f(t,k,\Theta) = \frac{t^{k-1}e^{-\frac{t}{\Theta}}}{\Gamma(k)\Theta^k}$, where k is the shape parameter and Θ is the scale parameter. To obtain an MTBF of μ_{ind}, we scale it using $\Theta = \frac{\mu_{ind}}{k}$ and obtain $f(t,k) = \frac{k^k t^{k-1} e^{-\frac{kt}{\mu_{ind}}}}{\Gamma(k)\mu_{ind}^k}$, where k is the shape parameter and Γ is the Gamma function. In the experiments, k is varied in $\{0.5, 0.7\}$. Note that $k = 1$ corresponds again to the exponential distribution;
• **The Lognormal distribution**, with probability density function $f(t,\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} e^{\left(-\frac{(\ln t-\mu)^2}{2\sigma^2}\right)}$ and with expectation $e^{\mu+\sigma^2/2}$, where μ and σ are respectively the expectation and the standard deviation of the variable's natural logarithm. We tested with two sets of (μ,σ), used in [13] and [42]: $(\mu_1 = 6.6025, \sigma_1 = 1.6206)$ and $(\mu_2 = 10.89, \sigma_2 = 1.08)$. In order to harmonize with the other probability distributions, we aim at having $\mu_{ind} = e^{\mu+\sigma^2/2} = 10$ years. To achieve this without altering the shape of the probability distribution, we fix a parameter $k = \mu/\sigma^2$, in order to express the probability density function with (μ_{ind}, k). After scaling, we obtain two sets $(\mu_{ind} = 10, k = 2.51)$ and $(\mu_{ind} = 10, k = 9.34)$ that we consider in the experiments. We retrieve the probability density function with:

$$\mu = \frac{\ln(\mu_{ind})}{1+\frac{1}{2k}}; \qquad \sigma = \sqrt{\frac{\mu}{k}} = \sqrt{\frac{\ln(\mu_{ind})}{k+\frac{k}{2}}}.$$

Traces

We generate a failure trace for each failure distribution and for each processor. In that trace, failure inter-arrival times obey the distribution, and the last failure happens after time h, where h is the horizon of the failure trace. The horizon

Table 1: Ratio of the execution time achieved by YoungDaly to that of NextStep for the 8 failure distributions, when $T_{base} = 48$ and $T_{plat} = 100$, and when aggregating all results.

	LogNormal 2.51		Weibull 0.5		Gamma 0.5		Weibull 0.7		Gamma 0.7		Exponential		Weibull 1.5		LogNormal 9.34	
$T_{base}=48$, $T_{plat}=100$	1.89	(2.02)	1.15	(1.34)	1.04	(1.17)	1.04	(1.14)	1	(1.1)	1.01	(1.06)	1.03	(1.06)	1.02	(1.11)
Aggregated	2.48	(2.26)	1.44	(1.6)	1.24	(1.43)	1.13	(1.28)	1.07	(1.21)	1.01	(1.07)	1.04	(1.07)	1.03	(1.09)

is set to two years ($h = 730$ days) for all the traces. The different heuristics are then evaluated using the trace, thereby making sure that all heuristics are evaluated using the very same failure scenario. If during a simulation, a checkpointing strategy reaches time h before the completion of the job, the simulation is said to fail.

Simulation Parameters

In the experiments, we compare both checkpointing strategies under the following parameter settings:

• The number of processors p, logarithmically varied in the range 10^3 to 10^5. These values represent mid-size to large parallel platforms.

• The checkpoint/recovery/downtime $C = R = 10D$, in seconds, varied in $\{60, 600\}$. In practice, the small value of C is optimistic while the later is pessimistic; and the low value of D assumes that spares are immediately available.

• The duration of the job T_{base}, in hours, varied in $\{1, 3, 10, 48\}$. T_{base} corresponds to the total length of the job, excluding checkpoints, if no failure occurs, when it is run on p processors (weak scaling). This corresponds to the duration range of typical HPC jobs, lasting from one hour up to two days.

• The age of the platform T_{plat}, in days, varied in $\{0, 10, 30, 100, 365\}$. This is the time from which we start using the failure traces: either from their very beginning if $T_{plat} = 0$, or from a later time if $T_{plat} > 0$. The age of the platform plays an important role for non-memoryless failure distributions. At the creation of the platform, all the processors are new and without any failure history. After a failure, the processor that failed is replaced/rejuvenated, but the other processors are not and keep their history. For instance, if the processors experience infant mortality, we expect the number of failures to be much higher with $T_{plat} = 0$, when all processors are new, than after a year of service ($T_{plat} = 365$).

Evaluation Methodology

For each possible choice of parameters, we generate 50 different failure scenarios. For each failure scenario, the simulated makespan (duration of the whole execution) of both heuristics is computed. We include the time spent to compute the segment sizes of NextStep. On the plots, we report the average makespan over these 50 instances, together with the tenth and the ninetieth percentiles, as a function of the number of processors. The YoungDaly heuristic is shown in red, and NextStep in blue. In all figures, the y-axis is the makespan in hours,

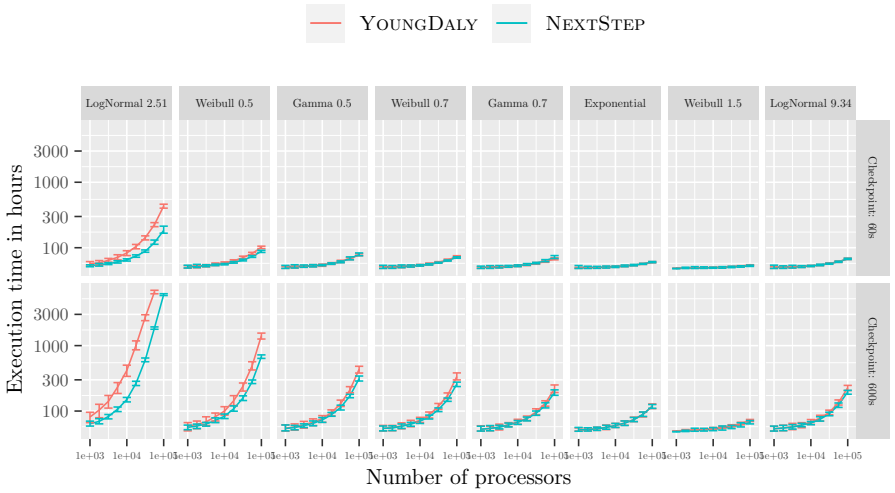


Figure 1: Expected performance of both heuristics under all failure distributions for a 100 days old platform with a workflow of $T_{base} = 48$ hours.

and the x-axis corresponds in most cases to the number of processors; both axes are in log-scale.

In the tables, we report the relative performance of YoungDaly and NextStep. More precisely, for each failure scenario, we compute the ratio of the makespan achieved by YoungDaly divided by that of NextStep. Hence, NextStep achieves a better makespan when the ratio is greater than 1; the larger the ratio, the higher the benefice of using NextStep. To produce meaningful statistics on these ratios, we compute and report their geometric mean and geometric standard deviation (in parentheses). For a few configurations, YoungDaly does not succeed to complete the job before the trace horizon. For these cases, in order to be able to compute statistics, we take for the execution time of YoungDaly a lower bound, namely the time at which the execution was stopped: $h - T_{plat}$. We checked that using this lower-bound or computing the statistics while just discarding these configurations leads to almost identical results (differences below 1%). The simulation code for all experiments is publicly available at `http://perso.ens-lyon.fr/frederic.vivien/resilience/non-memoryless-checkpoint`.

5.2 Results

We first compare the behavior of both checkpointing heuristics with the different probability distributions on a particular set of parameters, before studying the impact of the different parameters. Only a selection of results is presented here

Table 2: Ratio of the execution time achieved by YoungDaly to that of NextStep for the 8 failure distributions for the different platform sizes and when averaging over all the other parameters.

Platform size	LogNormal 2.51		Weibull 0.5		Gamma 0.5		Weibull 0.7		Gamma 0.7		Exponential		Weibull 1.5		LogNormal 9.34	
1000	1.34	(1.6)	1.14	(1.33)	1.08	(1.22)	1.03	(1.11)	1.01	(1.08)	1	(1.04)	1.01	(1.04)	1.01	(1.04)
1778	1.53	(1.82)	1.18	(1.41)	1.11	(1.3)	1.03	(1.12)	1.02	(1.1)	1	(1.04)	1.01	(1.04)	1.01	(1.04)
3162	1.88	(2.07)	1.26	(1.49)	1.16	(1.37)	1.05	(1.16)	1.02	(1.13)	1.01	(1.04)	1.02	(1.04)	1.01	(1.05)
5623	2.22	(2.28)	1.33	(1.54)	1.19	(1.38)	1.07	(1.19)	1.04	(1.14)	1.01	(1.06)	1.03	(1.04)	1.02	(1.06)
10000	2.72	(2.33)	1.38	(1.58)	1.23	(1.42)	1.09	(1.24)	1.06	(1.18)	1.01	(1.05)	1.03	(1.05)	1.02	(1.06)
17783	3.2	(2.33)	1.51	(1.66)	1.26	(1.44)	1.14	(1.29)	1.08	(1.22)	1.01	(1.07)	1.04	(1.06)	1.03	(1.07)
31623	3.74	(2.18)	1.72	(1.72)	1.36	(1.5)	1.22	(1.36)	1.12	(1.27)	1.02	(1.08)	1.07	(1.08)	1.05	(1.12)
56234	3.65	(2.02)	1.76	(1.63)	1.37	(1.5)	1.24	(1.35)	1.14	(1.3)	1.01	(1.1)	1.08	(1.09)	1.05	(1.13)
100000	3.5	(1.92)	1.85	(1.55)	1.41	(1.48)	1.33	(1.41)	1.15	(1.32)	1.01	(1.09)	1.08	(1.1)	1.07	(1.16)

due to lack of space, but exhaustive results for all combinations of parameters can be found in Apendix C.

5.2.1 Comparison of Probability Distributions

Figure 1 compares the two heuristics for the different failure distributions, with a checkpoint length of one or ten minutes, where the job length is 48 hours and the platform is 100 days old. In this case, although the platform is not new, we see that the NextStep heuristic is performing either better than or similarly to YoungDaly. Moreover, the difference tends to be more important when the checkpoint length is higher (bottom graphs). Recall that the lower is the better, since we plot execution times.

Although the MTBF of any individual processor is the same for all failure laws ($\mu_{ind} = 10$ years), the shape of these laws significantly impacts the number of failures that occur during the processing of the job, as well as the distribution of the failures. For instance, if the processors tend to have infant mortality (which corresponds to distributions on the left of the figure), and if the platform is not very old, then jobs may actually experience more failures than expected. This is the case for the Lognormal distribution with $k = 2.51$ or Weibull with $k = 0.5$ in Figure 1. This explains the higher execution times of YoungDaly for both heuristics.

Furthermore, YoungDaly does not checkpoint often enough, as it considers the global long term MTBF of the platform instead of its actual instantaneous failure rate. This is because YoungDaly does not take the failure history into account. On the contrary, NextStep does take that history into account. Therefore, it correctly estimates the instantaneous failure rate. This results in a makespan that can be up to two times lower.

There are some distributions for which processors tend to be more robust at the beginning because of their young age (distributions on the right of the figure). In this case, when the platform is rather young, the number of failures is lower than what would be expected regarding the MTBF of the platform. A good example is the Weibull distribution with $k = 1.5$ in Figure 1. In that case, the actual instantaneous failure rate of the platform is lower than expected, YoungDaly tends to over-checkpoint because it does not take into account

Table 3: Ratio of the execution time achieved by YoungDaly to that of NextStep for the 8 failure distributions as a function of platform age, with $p = 56234$ and $T_{base} = 48$ and when averaging over the two checkpoint sizes.

Platform age	LogNormal 2.51	Weibull 0.5	Gamma 0.5	Weibull 0.7	Gamma 0.7	Exponential	Weibull 1.5	LogNormal 9.34	**Average**
0	4.17 (2.06)	2.33 (1.48)	1.85 (1.44)	1.42 (1.38)	1.28 (1.32)	1.03 (1.08)	1.08 (1.07)	1.08 (1.07)	**1.58** (1.78)
10	3.27 (2.26)	1.61 (1.66)	1.29 (1.46)	1.13 (1.29)	1.06 (1.2)	1.01 (1.06)	1.04 (1.06)	1.02 (1.06)	**1.31** (1.71)
30	2.57 (2.17)	1.36 (1.55)	1.15 (1.32)	1.08 (1.21)	1.03 (1.16)	1 (1.06)	1.03 (1.06)	1 (1.06)	**1.21** (1.58)
100	1.89 (2.02)	1.15 (1.34)	1.04 (1.17)	1.04 (1.14)	1 (1.1)	1.01 (1.06)	1.03 (1.06)	1.02 (1.11)	**1.12** (1.42)
365	1.42 (1.72)	1.05 (1.17)	1.01 (1.1)	1.02 (1.11)	1 (1.08)	1 (1.06)	1.01 (1.07)	1.03 (1.13)	**1.06** (1.27)

Table 4: Ratio of the execution time achieved by YoungDaly to that of NextStep for the 8 failure distributions for the two checkpoint durations (in seconds) when averaging over all the other parameters.

Checkpoint duration	LogNormal 2.51	Weibull 0.5	Gamma 0.5	Weibull 0.7	Gamma 0.7	Exponential	Weibull 1.5	LogNormal 9.34
60	2.18 (2.27)	1.33 (1.52)	1.17 (1.37)	1.08 (1.2)	1.03 (1.14)	1 (1.03)	1.02 (1.04)	1.01 (1.04)
600	2.83 (2.2)	1.56 (1.66)	1.3 (1.47)	1.18 (1.34)	1.11 (1.26)	1.02 (1.09)	1.06 (1.08)	1.05 (1.12)

this actual failure rate, whereas NextStep adapts its checkpointing strategy according to this history, showing once again its versatility. Yet this time, the difference between heuristics is low, because the overall checkpointing cost remains small in both cases.

Finally, if the platform actual instantaneous failure rate is in accordance with the expected MTBF, as is the case for an Exponential law, YoungDaly is optimal. We check that the performance of NextStep and YoungDaly are similar in this setting.

Altogether, these results show that NextStep always adapts to the actual instantaneous failure rate, because it accounts for the failure history of processors. Its versatility makes it a better strategy in all the cases: Table 1 summarizes the results, reporting the ratio of the execution time achieved by YoungDaly to that of NextStep (geometric average, geometric standard deviation). We point out that the difference is more significant for the realistic distribution laws that have been advocated in the literature: namely Weibull with a shape parameter smaller than one [32, 33, 15], and Lognormal [13, 42].

The previous study was for long jobs lasting 48 hours and $T_{plat} = 100$. We also present the aggregated results over all job lengths and platform ages in Table 1. We observe that NextStep achieves even larger gains. For instance, for Lognormal 2.51, the average ratio becomes 2.48, instead of 1.89 for the scenario with $T_{plat} = 100$.

5.2.2 Impact of the Different Parameters

Impact of the number of processors. It can also be observed on Figure 1: the more processors, the more failures, and the larger the makespan for both heuristics, as one could have foretold. In most settings, the performance of YoungDaly worsens relatively to that of NextStep when the number of processors increases (recall that the y-axis is in log-scale). Again, this can be explained as follows: the difference between the estimated failure rate and

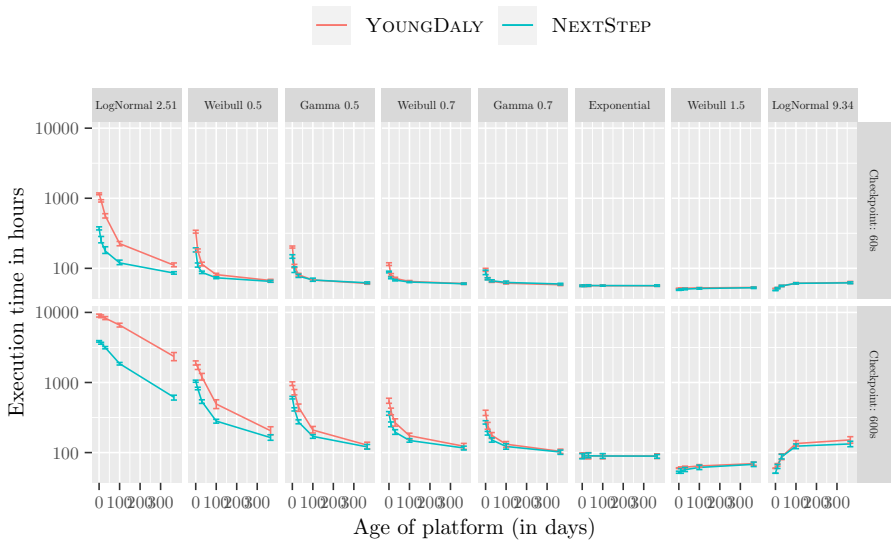


Figure 2: Expected performance of both heuristics under all failure distributions, with $p = 56234$ and $T_{base} = 48$ hours.

the instantaneous failure rate increases with the number of processors; hence, worse results for YoungDaly. On the contrary, NextStep adapts to the instantaneous failure rate. Table 2 provides a comprehensive summary of results for each platform size, averaging over all other parameters. The table confirms this observation.

Impact of the age of the platform. The age of the platform has a great impact on the performance of both heuristics, because the instantaneous failure rate of the platform highly depends on it. When processors have a high infant mortality, a more recent platform leads to more errors and thus to a higher makespan for both heuristics. This can be observed in Figure 2, especially on the leftmost graphs. On this figure, the x-axis is now the age of the platform (in a linear scale). The number of processors is fixed to $p = 56234$ and the job execution time is $T_{base} = 48$ hours.

For all distributions to the left of the Exponential, the newer the platform, the higher the difference between the heuristics. Indeed, for younger platforms, processors are more likely to fail due to infant mortality; and the older the platform, the more the instantaneous failure rate resembles an Exponential. The same observation can be made for Weibull 1.5, although in this case, this is due to low infant mortality. Indeed, with this distribution, less failures occur for younger platforms.

Lognormal 9.34 behaves slightly differently. For this distribution, once again, YoungDaly does not adapt to the instantaneous failure rate: either it overesti-

Table 5: Ratio of the execution time achieved by YoungDaly to that of NextStep for the 8 failure distributions for the different job lengths (in hours) and when averaging over all the other parameters.

Job length	LogNormal 2.51		Weibull 0.5		Gamma 0.5		Weibull 0.7		Gamma 0.7		Exponential		Weibull 1.5		LogNormal 9.34	
1	2.36	(2.84)	1.42	(1.81)	1.25	(1.58)	1.12	(1.39)	1.07	(1.29)	1.02	(1.09)	1.05	(1.08)	1.04	(1.12)
3	2.83	(2.47)	1.52	(1.71)	1.29	(1.49)	1.15	(1.32)	1.09	(1.25)	1.01	(1.07)	1.04	(1.08)	1.03	(1.1)
10	2.63	(2)	1.47	(1.51)	1.25	(1.35)	1.14	(1.23)	1.07	(1.17)	1	(1.05)	1.04	(1.06)	1.03	(1.08)
48	2.16	(1.6)	1.35	(1.32)	1.16	(1.21)	1.11	(1.15)	1.05	(1.11)	0.997	(1.03)	1.03	(1.05)	1.02	(1.06)

mates the instantaneous failure rate of a new platform and does not checkpoint enough, or it underestimates the instantaneous failure rate of an old platform and checkpoints too much. On the contrary, NextStep adjusts the checkpointing strategy for both cases. For intermediate platform ages, both heuristics have close performance because this is where the instantaneous failure rate is the closest to what is expected ($\frac{\mu_{ind}}{p}$) by YoungDaly. Nevertheless, the variance is different from that of the Exponential law, and NextStep achieves slightly better performance.

Table 3 summarizes these results. The last column provides an average over all distributions. Most gains are obtained for young platforms. NextStep always achieves a performance at least similar to YoungDaly, and much better in many cases.

Impact of the checkpoint time. As expected, the larger the checkpoint cost, the larger the execution time for both heuristics, as shown in Figure 1. Having a larger checkpoint cost exacerbates the differences between both heuristics. Indeed, when checkpoints cost more, both heuristics execute fewer checkpoints and thus lose more time at each failure. In the end, this increases the relative errors due to a bad checkpointing strategy. Table 4 summarizes the results for the two checkpoint costs (one minute or ten minutes).

Impact of the job length. Again, the larger the job length, the larger the execution time for both heuristics, as shown on Figure 3. Moreover, the error bars are much wider for a small workload, because having larger jobs will smooth the impact of each individual failure. Table 5 summarizes results for the four job lengths by aggregating all results. Overall, more gain can be achieved with smaller job lengths. Indeed, relatively to the lengths of jobs, checkpoints are more expensive for small jobs. This conclusion is similar to that on the impact of the cost of checkpoints. This phenomenon can be observed by comparing Tables 4 and 5, where the impact of increasing the checkpoint cost is similar to the impact of decreasing the job length.

6 Conclusion

In this paper, we have investigated checkpointing strategies to protect parallel jobs from non-memoryless fail-stop errors. Indeed, the optimal strategy has been well studied when failure inter-arrival times obey an Exponential law, but not well understood for non-memoryless failure distributions. We have designed

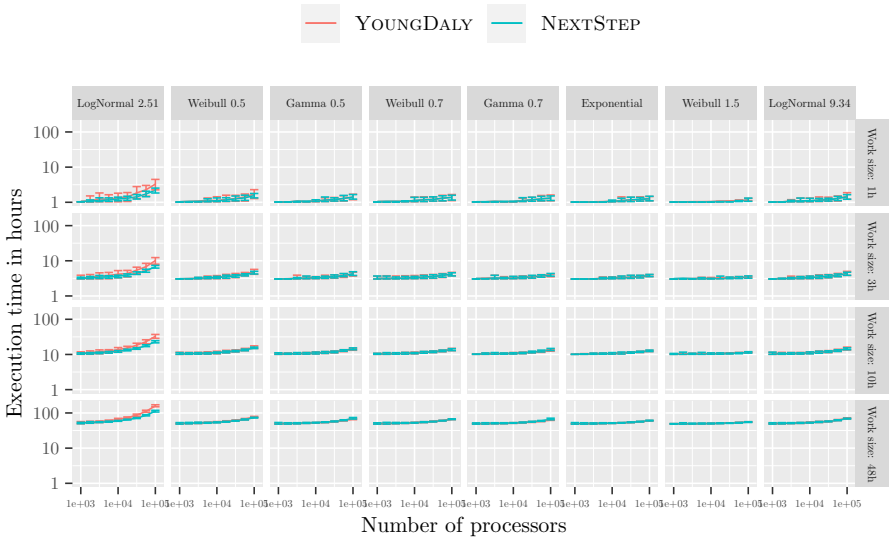


Figure 3: Expected performance of both heuristics under all failure distributions on a 365 day old platform, with $C = 60$ seconds.

a general strategy, NextStep, which maximizes the expected efficiency until the next failure. While it may not be optimal because of side-effects towards the end of the job, we proved that this strategy is asymptotically optimal for very long jobs.

Instead of maximizing the expected efficiency until the next failure, traditional solutions consist in checkpointing periodically according to the platform MTBF (YoungDaly strategy). Our extensive simulation results show that this strategy works well for Exponential laws, but not for the other laws, because it either underestimates or overestimates the actual instantaneous failure rate. On the contrary, NextStep is always at least as good as YoungDaly for any failure distribution, and significantly outperforms it in many cases. Overall, our study demonstrates the interest of always using NextStep instead of YoungDaly.

In particular, the difference between NextStep and YoungDaly is very important for distributions whose infant mortality of the distribution is high, e.g. LogNormal 2.51 or Weibull 0.5. The latter distributions have been advocated to model failures on real-life platforms, which further evidences the impact and significance of NextStep.

Future work will focus on checkpointing strategies for workflows composed of parallel jobs with dependencies, instead of independent jobs as in this study. The criticality of some jobs in the workflow may lead to checkpoint them more often than prescribed by the NextStep strategy tuned for a given non-memoryless

failure distribution.

References

[1] G. Aupy, A. Benoit, H. Casanova, and Y. Robert. Scheduling computational workflows on failure-prone platforms. *Int. J. of Networking and Computing*, 6(1):2–26, 2016.

[2] G. Aupy, Y. Robert, and F. Vivien. Assuming failure independence: are we right to be wrong? In *FTS'2017*, 2017.

[3] L. Bautista-Gomez, A. Gainaru, S. Perarnau, D. Tiwari, S. Gupta, C. Engelmann, F. Cappello, and M. Snir. Reducing waste in extreme scale systems through introspective analysis. In *IPDPS*, pages 212–221. IEEE, 2016.

[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, V. Le Fèvre, Y. Robert, and H. Sun. Towards optimal multi-level checkpointing. *IEEE Trans. Computers*, 66(7):1212–1226, 2017.

[6] A. Benoit, A. Cavelan, Y. Robert, and H. Sun. Assessing general-purpose algorithms to cope with fail-stop and silent errors. *ACM Trans. Parallel Computing*, 3(2), 2016.

[7] M. Bougeret, H. Casanova, M. Rabie, Y. Robert, and F. Vivien. Checkpointing strategies for parallel jobs. In *Proc. of SC'11*, 2011.

[8] F. Cappello, A. Geist, W. Gropp, S. Kale, B. Kramer, and M. Snir. Toward exascale resilience: 2014 update. *Supercomputing frontiers and innovations*, 1(1), 2014.

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

[10] K. L. Chung. *A Course in Probability Theory.* Stanford University, 3 edition, 2000.

[11] J. T. Daly. A higher order estimate of the optimum checkpoint interval for restart dumps. *FGCS*, 22(3):303–312, 2006.

[12] S. Di, M. S. Bouguerra, L. Bautista-Gomez, and F. Cappello. Optimization of multi-level checkpoint model for large scale HPC applications. In *IPDPS*. IEEE, 2014.

[13] S. Di, H. Guo, R. Gupta, E. R. Pershey, M. Snir, and F. Cappello. exploring properties and correlations of fatal events in a large-scale hpc system. *Trans. on Parallel and Distributed Systems*, 2018.

[14] S. Di, Y. Robert, F. Vivien, and F. Cappello. Toward an optimal online checkpoint solution under a two-level HPC checkpoint model. *IEEE Trans. Parallel & Distributed Systems*, 2016.

[15] N. El-Sayed and B. Schroeder. To checkpoint or not to checkpoint: Understanding energy-performance-i/o tradeoffs in hpc checkpointing. In *CLUSTER*, pages 93–102, 2014.

[16] K. Ferreira, J. Stearley, J. H. I. Laros, R. Oldfield, K. Pedretti, R. Brightwell, R. Riesen, P. G. Bridges, and D. Arnold. Evaluating the Viability of Process Replication Reliability for Exascale Systems. In *SC'11*. ACM, 2011.

[17] A. Frank, M. Baumgartner, R. Salkhordeh, and A. Brinkmann. Improving checkpointing intervals by considering individual job failure probabilities. In *IPDPS*, pages 299–309, 2021.

[18] E. Gelenbe, P. Boryszko, M. Siavvas, and J. Domanska. Optimum checkpoints for time and energy. In *28th MASCOTS*, pages 1–8. IEEE, 2020.

[19] L. Han, L.-C. Canon, H. Casanova, Y. Robert, and F. Vivien. Checkpointing workflows for fail-stop errors. *IEEE Trans. Computers*, 67(8):1105–1120, 2018.

[20] L. Han, V. Le Fèvre, L.-C. Canon, Y. Robert, and F. Vivien. A generic approach to scheduling and checkpointing workflows. In *ICPP'2018, the 47th Int. Conf. on Parallel Processing*, 2018.

[21] E. Heien, D. Kondo, A. Gainaru, D. LaPine, B. Kramer, and F. Cappello. Modeling and tolerating heterogeneous failures in large parallel systems. In *Proc. SC'11*, 2011.

[22] T. Herault and Y. Robert, editors. *Fault-Tolerance Techniques for High-Performance Computing*, Computer Communications and Networks. Springer Verlag, 2015.

[23] T. Herault, Y. Robert, A. Bouteiller, D. Arnold, K. B. Ferreira, G. Bosilca, and J. Dongarra. Checkpointing strategies for shared high-performance computing platforms. *International Journal of Networking and Computing*, 9(1):28–52, 2019.

[24] S. Hiroyama, T. Dohi, and H. Okamura. Aperiodic checkpoint placement algorithms—survey and comparison. *Journal of Software Engineering and Applications*, 6(4A):41–53, 2013.

[25] W. Jones, J. Daly, and N. DeBardeleben. Impact of sub-optimal checkpoint intervals on application efficiency in computational clusters. In *HPDC'10*, pages 276–279. ACM, 2010.

[26] O. Kella and W. Stadje. Superposition of renewal processes and an application to multi-server queues. *Statistics & probability letters*, 76(17):1914–1924, 2006.

[27] S. Y. Ko, I. Hoque, B. Cho, and I. Gupta. Making cloud intermediate data fault-tolerant. In *Proc. 1st ACM Symposium on Cloud Computing*, SoCC '10. ACM, 2010.

[28] S. Levy and K. B. Ferreira. An examination of the impact of failure distribution on coordinated checkpoint/restart. In *FTXS Workshop*, pages 35–42. ACM, 2016.

[29] Y. Ling, J. Mi, and X. Lin. A variational calculus approach to optimal checkpoint placement. *IEEE Trans. Computers*, pages 699–708, 2001.

[30] 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*, 2010.

[31] H. Okamura and T. Dohi. Comprehensive evaluation of aperiodic checkpointing and rejuvenation schemes in operational software system. *Journal of Systems and Software*, 83(9):1591–1604, 2010.

[32] B. Schroeder and G. A. Gibson. A large-scale study of failures in high-performance computing systems. In *Proc. of DSN*, pages 249–258, 2006.

[33] B. Schroeder and G. A. Gibson. Understanding failures in petascale computers. *Journal of Physics: Conference Series*, 78(1), 2007.

[34] K. Schroiff, P. Gemsjaeger, and C. Bolik. Cascading failover of a data management application for shared disk file systems in loosely coupled node clusters, 2006. US Patent 6,990,606.

[35] P. Sigdel, X. Yuan, and N. Tzeng. Realizing best checkpointing control in computing systems. *IEEE TPDS*, 32(2):315–329, 2021.

[36] L. Silva and J. Silva. Using two-level stable storage for efficient checkpointing. *IEE Proceedings - Software*, 145(6):198–202, 1998.

[37] D. Stirzaker. *Elementary Probability*. Cambridge University Press, 2 edition, 2003.

[38] O. Subasi, G. Kestor, and S. Krishnamoorthy. Toward a general theory of optimal checkpoint placement. In *CLUSTER*, pages 464–474. IEEE, 2017.

[39] O. Subasi, T. Martsinkevich, F. Zyulkyarov, O. Unsal, J. Labarta, and F. Cappello. Unified fault-tolerance framework for hybrid task-parallel message-passing applications. *IJHPCA*, 32(5):641–657, 2018.

[40] D. Tiwari, S. Gupta, and S. S. Vazhkudai. Lazy checkpointing: Exploiting temporal locality in failures to mitigate checkpointing overheads on extreme-scale systems. In *44th Int. Conf. on Dependable Systems and Networks*, pages 25–36. IEEE, 2014.

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

[42] N. Yigitbasi, M. Gallet, D. Kondo, A. Iosup, and D. Epema. Analysis and modeling of time-correlated failures in large-scale distributed systems. *Parallel and Distrib. Syst. Report Series*, 2010.

[43] J. W. Young. A first order approximation to the optimum checkpoint interval. *Comm. of the ACM*, 17(9):530–531, 1974.

A NextStep is not optimal for Exponential laws

A.1 Introduction to the example

Consider a single-processor job of length $T_{base} = T = 0.062249$, which we need to checkpoint at the end. The duration of a checkpoint is $C = 0.001$ and we have $R = D = 0$. The failure probability density function is $f(x) = e^{-x}$, an Exponential law with parameter $\lambda = 1$. We show that NextStep is not optimal for this example.

We start in Section A.2 with the computation of a fundamental function: if a failure strikes between instants a and b, what is the expected time $\mathbb{E}_{lost}(a, b)$ elapsed between a and the time of the failure? Here, a will always be either equal to 0 ($a = 0$), or to the time the last checkpoint was completed. Then, b can be any instant before the end of the next checkpoint. Thus, $\mathbb{E}_{lost}(a, b)$ corresponds to the expected work lost after the last checkpoint.

Next, we compute in Section A.3 the expected makespan when a single checkpoint is taken $\mathbb{E}^*_{stat}(T, N_c = 1)$. We show in Section A.4 that this is the best static strategy, by proving that the optimal makespan of a static strategy which takes two checkpoints, $\mathbb{E}^*_{stat}(T, N_c = 2)$, is strictly greater than $\mathbb{E}^*_{stat}(T, N_c = 1)$. This enables us to conclude using the results recalled in Section 3.2 of the main paper. Indeed, with the notations of Section 3.2, if taking two checkpoints is worse than taking one, then the optimal number of checkpoints N_{opt} is not greater than 2. Hence, $N_{ME} \leq 2$.

Finally, we lower-bound the expected makespan of NextStep. To do so, we first find in Section A.5 the optimal expected efficiency $\mathbb{E}^*_e(T, N_c)$ achievable by NextStep for $N_c = 1$ and $N_c = 2$, and we show that the efficiency is better for $N_c = 2$, implying that the initial strategy includes at least two checkpoints. Then, in Section A.6, for any $N_c \geq 2$, we suppose that NextStep

does N_c checkpoints in the first call and we lower-bound the expected makespan $\hat{\mathbb{E}}_{NS}(T, N_c)$ in this case. Rephrased differently, $\hat{\mathbb{E}}_{NS}(T, N_c)$ corresponds to the expected makespan of the algorithm that optimizes the efficiency up to the next failure starting with N_c checkpoints, and applies NextStep afterwards. We check that for all $N_c \geq 2$, $\hat{\mathbb{E}}_{NS}(T, N_c) > \mathbb{E}^*_{stat}(T, 1)$. In particular, if the actual number of checkpoints of NextStep planned in the first call is $N_c^* \geq 2$, its expected makespan $\mathbb{E}_{NS}(T)$ verifies $\mathbb{E}_{NS}(T) = \hat{\mathbb{E}}_{NS}(T, N_c^*) > \mathbb{E}^*_{stat}(T, 1)$, which shows that it is not optimal.

A.2 Computation of $\mathbb{E}_{lost}(a, b)$

The expected time lost if we have a failure between a and b, when a checkpoint was completed exactly at time a or when $a = 0$, is computed as follows:

$$\begin{aligned}
\mathbb{E}_{lost}(a,b) &= \int_a^b (x-a)\mathbb{P}\{X = x | a < X < b\}dx \\
&= \frac{1}{\mathbb{P}\{a < X < b\}} \int_a^b (x-a) f(x) dx \\
&= \frac{1}{\mathbb{P}\{X < b\} - \mathbb{P}\{X < a\}} \int_a^b (x-a) e^{-x} dx \\
&= \frac{1}{e^{-a} - e^{-b}} \int_a^b (x-a) e^{-x} dx
\end{aligned}$$

$$\begin{aligned}
\int_a^b (x-a)e^{-x}dx &= \left[-(x-a)e^{-x}\right]_a^b - \int_a^b -e^{-x}dx \\
&= -(b-a)e^{-b} - \left[e^{-x}\right]_a^b \\
&= -(b-a)e^{-b} + e^{-a} - e^{-b}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{lost}(a,b) &= \frac{-(b-a)e^{-b} + e^{-a} - e^{-b}}{e^{-a} - e^{-b}} \\
&= 1 - \frac{(b-a)e^{-b}}{e^{-a} - e^{-b}} \\
&= 1 - \frac{b-a}{e^{b-a} - 1}
\end{aligned}$$

We check that $\mathbb{E}_{lost}(a, b)$ only depends on $b - a$, which is comforting because the law is memoryless. In the following, we will use the difference $b - a$ as single parameter x, i.e.,

$$\mathbb{E}_{lost}(x) = 1 - \frac{x}{e^x - 1} \tag{11}$$

A.3 Computation of $\mathbb{E}^*_{stat}(T, N_c = 1)$

With a single checkpoint, the makespan is $T+C$ if there are no failures. Otherwise, we lose the time elapsed before the first failure occurs, $\mathbb{E}_{lost}(T+C)$, and we need to start again. Therefore, $\mathbb{E}^*_{stat}(T, N_c = 1)$ can be expressed as follows:

$$\begin{aligned}\mathbb{E}^*_{stat}(T,1) &= \mathbb{P}\{X > T+C\}(T+C) \\ &\quad + \mathbb{P}\{X < T+C\}(\mathbb{E}^*_{stat}(T,1) + \mathbb{E}_{lost}(T+C))\end{aligned}$$

$$\mathbb{E}^*_{stat}(T,1) = T + C + \frac{1-e^{-(T+C)}}{e^{-(T+C)}}\mathbb{E}_{lost}(T+C) \tag{12}$$

With $T = 0.062249$ and $C = 0.001$, using Equation (11), we obtain $\mathbb{E}^*_{stat}(T,1) \approx 0.06529206$.

A.4 Computation of $\mathbb{E}^*_{stat}(T, N_c = 2)$

In this case, recall from Section 3.2 (main paper) that the best solution is to have two equal-sized segments. Therefore, we can use Equation (12) with $\mathbb{E}^*_{stat}\left(\frac{T}{2}, 1\right)$ and obtain:

$$\begin{aligned}\mathbb{E}^*_{stat}\left(T, N_c = 2\right) &= 2 \times \mathbb{E}^*_{stat}\left(\frac{T}{2}, 1\right) \\ &= 2\left(\frac{T}{2} + C + \frac{1-e^{-\left(\frac{T}{2}+C\right)}}{e^{-\left(\frac{T}{2}+C\right)}}\mathbb{E}_{lost}\left(\frac{T}{2}+C\right)\right)\end{aligned}$$

Using Equation (11), we get $\mathbb{E}^*_{stat}(T,2) \approx 0.06529212 > \mathbb{E}^*_{stat}(T,1)$ with $T = 0.062249$ and $C = 0.001$. As stated in Section A.1, this result shows that the optimal checkpointing heuristic uses a single checkpoint at the end.

A.5 Computation of $\mathbb{E}^*_e(T, N_c = 1)$ and $\mathbb{E}^*_e(T, N_c = 2)$

We now show that NextStep does at least two checkpoints in the first call, by computing the best expected efficiency $\mathbb{E}^*_e(T, N_c)$ using one or two checkpoints, and checking that indeed $\mathbb{E}^*_e(T,2) > \mathbb{E}^*_e(T,1)$.

With a single checkpoint at the end, we can either process 0 units of work if we have a failure, or process the whole job. Therefore,

$$\begin{aligned}\mathbb{E}^*_W(T,1) &= 0 \times \mathbb{P}\{X < T+C\} + T\mathbb{P}\{X > T+C\} \\ &= Te^{-(T+C)}\end{aligned}$$

Furthermore, the expected time before the next failure or the end of the job can be computed as follows:

$$\begin{aligned}\mathbb{E}_{T_{next}}(T,1) &= \mathbb{P}\{X > T+C\}(T+C) \\ &\quad + \mathbb{P}\{X < T+C\}\mathbb{E}_{lost}(T+C) \\ &= e^{-(T+C)}(T+C) \\ &\quad + (1-e^{-(T+C)})\left(1 - \frac{T+C}{e^{T+C}-1}\right)\end{aligned}$$

Hence, with no additional checkpoint,

$$\mathbb{E}_e^*(T,1) = \frac{Te^{-(T+C)}}{e^{-(T+C)}(T+C) + (1-e^{-(T+C)})\left(1-\frac{T+C}{e^{T+C}-1}\right)} \quad (13)$$

With $T = 0.062249$ and $C = 0.001$, we obtain $\mathbb{E}_e^*(T,0,1) \approx 0.95339305$.

If we add one checkpoint, we could place it anywhere; in that case, if we checkpoint after $w \in (0,T)$ units of work, the work done before the next failure or the end of the job can either be 0 if a failure occurs in $[0, w+C]$, w if a failure occurs in $(w+C, T+2C)$ or T if no failure occurs before $T+2C$. Thus,

$$\begin{aligned}
\mathbb{E}_W^*(T, N_c = 2) &= \max_w \mathbb{E}_W(T,2,w) \\
&= \max_w (w\mathbb{P}\{w+c < X < T+2C\} \\
&\qquad + T\mathbb{P}\{X > T+2C\}) \\
&= \max_w (w(e^{-(w+C)} - e^{-(T+2C)}) + Te^{-(T+2C)}) \\
&= \max_w ((T-w)e^{-(T+2C)} + we^{-(w+C)})
\end{aligned}$$

The expected time before the next failure or the end of the job is:

$$\begin{aligned}
\mathbb{E}_{T_{next}}(T, N_c = 2) &= \mathbb{P}\{X > T+2C\}(T+2C) \\
&\quad + \mathbb{P}\{X < T+2C\}\mathbb{E}_{lost}(0, T+2C) \\
&= e^{-(T+2C)}(T+2C) \\
&\qquad + (1-e^{-(T+2C)})\left(1 - \frac{T+2C}{e^{T+2C}-1}\right)
\end{aligned}$$

Altogether, these results give the formula for $\mathbb{E}_e^*(T, N_c = 2)$:

$$\mathbb{E}_e^*(T,2) = \frac{\max_w((T-w)e^{-(T+2C)} + we^{-(w+C)})}{e^{-(T+2C)}(T+2C) + (1-e^{-(T+2C)})\left(1-\frac{T+2C}{e^{T+2C}-1}\right)}$$

The best choice for w is hard to express analytically, but with $T = 0.062249$ and $C = 0.001$, we numerically obtain $w^* \approx 0.0313732$ and $\mathbb{E}_e^*(T,2) > 0.95339312 > \mathbb{E}_e^*(T,1)$. This shows that NextStep will not execute a single checkpoint at the end. We will finally show that for all $N_c \geq 2$, $\hat{\mathbb{E}}_{NS}(T, N_c) > \mathbb{E}_{stat}^*(T,1)$.

A.6 Lower bound on $\hat{\mathbb{E}}_{NS}(T, N_c = 2)$

We have shown that NextStep plans to take at least two checkpoints in its first call. We now move to evaluating the expected makespan achieved by NextStep.

We have already shown that if NextStep does two checkpoints in the first call, the first one is placed after w^* units of work executed and, by definition, the second one is taken at the end. To compute $\hat{\mathbb{E}}_{NS}(T, 2)$, we note that if no failures happen, the makespan is $T + 2C$, whereas if the failure strikes in the second segment, we lose $\mathbb{E}_{lost}(T - w^* + C)$ and we need to reprocess the remaining $T - w^*$ units of time, which takes a time $\mathbb{E}_{NS}(T - w^*)$ (because the law is memoryless and $R = D = 0$). Finally, if a failure strikes in the first segment, we lose $\mathbb{E}_{lost}(w^* + C)$ and we need to retry from the beginning. Overall, $\hat{\mathbb{E}}_{NS}(T, 2)$ can be expressed as follows:

$$\begin{aligned}\hat{\mathbb{E}}_{NS}(T,2) &= \mathbb{P}\{X > T + 2C\}(T + 2C)\\ &\quad + (\mathbb{P}\{X > w^* + C\} - \mathbb{P}\{X > T + 2C\})\\ &\quad \times (w^* + C + \mathbb{E}_{lost}(T + C - w^*) + \mathbb{E}_{NS}(T - w^*))\\ &\quad + (1 - \mathbb{P}\{X > w^* + C\})(\mathbb{E}_{lost}(w^* + C) + \hat{\mathbb{E}}_{NS}(T, 2))\end{aligned}$$

We isolate $\hat{\mathbb{E}}_{NS}(T, 2)$ on the left side of the equation and divide by $\mathbb{P}\{X > w^* + C\}$ to obtain:

$$\begin{aligned}\hat{\mathbb{E}}_{NS}(T,2) &= \frac{1}{e^{-(w^*+C)}}\\ &\quad \times ((T + 2C)e^{-(T+2C)} + (1 - e^{-(w^*+C)})\mathbb{E}_{lost}(w^* + C))\\ &\quad + \frac{e^{-(w^*+C)} - e^{-(T+2C)}}{e^{-(w^*+C)}}\\ &\quad \times (w^* + C + \mathbb{E}_{lost}(T + C - w^*) + \mathbb{E}_{NS}(T - w^*)))\end{aligned}$$

The only unknown value is $\mathbb{E}_{NS}(T - w^*)$, which we may lower-bound by the optimal strategy given a segment of length $T - w^*$. If we choose to do a single checkpoint (at the end of the segment) and stick with this strategy, the expected makespan of the segment is $\mathbb{E}^*_{stat}(T - w^*, 1) \approx 0.0324$. However, if we choose to do at least one additional checkpoint, the expected makespan is larger than $T - w^* + 2C$. Indeed, this will be the case if there is no failure before the end of the first checkpoint because, in this case, we will have to take at least one more checkpoint after the initial one. If, otherwise, a failure strikes before the completion of this first checkpoint, the situation is exactly the same as before (the law being memoryless). Therefore, the strategy will not change and, whatever the failure scenario, we will eventually execute two checkpoints. Since $T - w^* + 2C \approx 0.0329 > \mathbb{E}^*_{stat}(T - w^*, 1)$, an optimal strategy is initially to do a single checkpoint.

Clearly, if a failure occurs, changing the strategy can only result in an increase in expected makespan, so we have $\mathbb{E}_{NS}(T - w^*) \geq \mathbb{E}^*_{stat}(T - w^*, 1)$. Finally,

$$\begin{aligned}\hat{\mathbb{E}}_{NS}(T,2) \geq & \frac{1}{e^{-(w^*+C)}} \\ & \times ((T+2C)e^{-(T+2C)} + (1-e^{-(w^*+C)})\mathbb{E}_{lost}(w^*+C)) \\ & + \frac{e^{-(w^*+C)} - e^{-(T+2C)}}{e^{-(w^*+C)}} \\ & \times (w^* + C + \mathbb{E}_{lost}(T+C-w^*) + \mathbb{E}^*_{stat}(T-w^*,1))\end{aligned}$$

Using Equations (11) and (12), with $T = 0.062249$ and $C = 0.001$, we get $\hat{\mathbb{E}}_{NS}(T,2) > 0.06529218 > \mathbb{E}^*_{stat}(T,1)$.

A.7 Lower bound on $\hat{\mathbb{E}}_{NS}(T, N_c = 3)$

We now assume that NextStep initially splits the job into three segments. We consider the longest segment, of length $L \geq \frac{T}{3}$, which starts after w units of work and s segments. In particular, $w = 0$ if $s = 0$, $w \in (0, T-L)$ if $s = 1$, and $w = T - L$ if $s = 2$.

To lower-bound the expected makespan in that scenario, we consider several cases:

- If no failure occurs, the makespan is $T+3C$. This happens with probability $e^{-(T+3C)}$.
- If a failure occurs in the longest segment, the conditional expected makespan is at least $T + C + \mathbb{E}_{lost}\left(\frac{T}{3} + C\right)$, because we have to process at least $T + C$ for the job in total, and we lose an additional time $\mathbb{E}_{lost}(L + C) \geq \mathbb{E}_{lost}\left(\frac{T}{3} + C\right)$ because a failure occurred in the longest segment. We denote by $f(w, s, L)$ the probability of this case.
- Finally, in all the other cases, whose total probability is $1 - e^{-(T+3C)} - f(w, s, L)$, we will have at least to process the whole work and take a checkpoint for a time of $T + C$.

We are always in one and only one of these three scenarios, therefore we can lower-bound the expected makespan using previous bounds on the conditional expected makespan combined with their probability of occurrence:

$$\begin{aligned}\hat{\mathbb{E}}_{NS}(T, N_c = 3) \geq & (T+3C)e^{-(T+3C)} \\ & + \left(T + C + \mathbb{E}_{lost}\left(\frac{T}{3} + C\right)\right) f(w,s,L) \\ & + (T+C)\left(1 - e^{-(T+3C)} - f(w,s,L)\right) \\ \hat{\mathbb{E}}_{NS}(T, N_c = 3) \geq & (T+3C)e^{-(T+3C)} \\ & + \mathbb{E}_{lost}\left(\frac{T}{3} + C\right) f(w,s,L) + (T+C)\left(1 - e^{-(T+3C)}\right)\end{aligned}$$

Clearly, the bound increases when $f(w, s, L)$ increases. Therefore, we need to find a lower bound for $f(w, s, L)$ to be able to conclude. $f(w, s, L)$ is the probability that a failure strikes the longest segment the first time we try to process it. This is the conjunction of the success of the segments before and the failure of the longest segment. Therefore,

$$f(w, s, L) = e^{-(w+sC)}(1 - e^{-(L+C)}).$$

This function is clearly increasing with L and decreasing with w and s. The largest possible s is 2. As $L \geq \frac{T}{3}$ and $w \leq T - L$, the smallest possible L is $\frac{T}{3}$ whereas the largest possible w is $\frac{2T}{3}$. Clearly, $f(w, s, L) \geq f\left(\frac{2T}{3}, 2, \frac{T}{3}\right)$, and the lower bound on $\hat{\mathbb{E}}_{NS}(T, N_c = 3)$ becomes:

$$\begin{aligned}\hat{\mathbb{E}}_{NS}(T, 3) \geq (T + 3C)e^{-(T+3C)} + (T + C)\left(1 - e^{-(T+3C)}\right) \\ + \mathbb{E}_{lost}\left(\frac{T}{3} + C\right)e^{-\left(\frac{2T}{3}+2C\right)}\left(1 - e^{-\left(\frac{T}{3}+C\right)}\right)\end{aligned}$$

Using Equation (11), with $T = 0.062249$ and $C = 0.001$, we get $\hat{\mathbb{E}}_{NS}(T, 3) > 0.06534 > \mathbb{E}^*_{stat}(T, 1)$.

A.8 Lower bound of $\hat{\mathbb{E}}_{NS}(T, N_c)$ for $4 \leq N_c \leq 1001$

If NextStep initially chooses to do N_c checkpoints, the resulting makespan is $T + N_cC$ if no failures occur, and at least $T + C$ otherwise. Therefore,

$$\begin{aligned}\hat{\mathbb{E}}_{NS}(T, N_c) &\geq (T + N_cC)e^{-(T+N_cC)} + (1 - e^{-(T+N_cC)})(T + C) \\ &= T + C + (N_c - 1)Ce^{-(T+N_cC)} \\ &= T + C + e^{-(T+C)} \times (N_c - 1)Ce^{-(N_c-1)C}\end{aligned}$$

The function $f(x) = xe^{-x}$ is increasing for $0 \leq x \leq 1$. Hence, because $(N_c - 1)C \leq 1000 \times 0.001 = 1$, we use $N_C = 4$ to get

$$\hat{\mathbb{E}}_{NS}(T, N_c) \geq \hat{\mathbb{E}}_{NS}(T, 4) = T + C + 3Ce^{-(T+4C)}.$$

With $T = 0.062249$ and $C = 0.001$, we get $\hat{\mathbb{E}}_{NS}(T, N_c) > 0.066 > \mathbb{E}^*_{stat}(T, 1)$ in this case.

A.9 Lower bound of $\hat{\mathbb{E}}_{NS}(T, N_c)$ for $N_c > 1001$

Finally, if NextStep initially chooses to do N_c checkpoints, the resulting makespan is at least $\mathbb{E}_{lost}(T + N_cC)$ if a failure occurs. Thus,

$$\begin{aligned}
\hat{\mathbb{E}}_{NS}(T, N_c) &\geq (1 - e^{-(T+N_cC)})\mathbb{E}_{lost}\left(T + N_cC\right) \\
&= (1 - e^{-(T+N_cC)})\left(1 - \frac{T + N_cC}{e^{T+N_cC} - 1}\right) \\
&= (1 - e^{-(T+N_cC)})\left(1 - \frac{T + N_cC}{1 - e^{-(T+N_cC)}}e^{-(T+N_cC)}\right) \\
&= 1 - e^{-(T+N_cC)} - (T + N_cC)e^{-(T+N_cC)} \\
&= 1 - (1 + T + N_cC)e^{-(T+N_cC)}
\end{aligned}$$

Let $f(x) = 1 - (1 + x)e^{-x}$. Then, $\hat{\mathbb{E}}_{NS}(T, N_c) = f(T + N_CC)$ and

$$f'(x) = -e^{-x} + (1 + x)e^{-x} = xe^{-x}.$$

For $x > 0$, f is therefore increasing with x and we have

$$f(T + N_CC) \geq f(T + 1001C).$$

With $T = 0.062249$ and $C = 0.001$, we obtain $\hat{\mathbb{E}}_{NS}(T, N_c) \geq f(T + N_CC) \geq f(T + 1001C) > 0.2 > \mathbb{E}^*_{stat}(T, 1)$.

We have shown that for all $N_c \geq 2$, $\hat{\mathbb{E}}_{NS}(T, N_c) > \mathbb{E}^*_{stat}(T, 1)$; therefore NextStep is strictly worse than the best static strategy, which shows that it is not optimal. This concludes the analysis of the counter-example.

B On the dynamic version of the optimal static strategy for an Exponential law

In this section, we prove that the dynamic version of the optimal static strategy is identical to the static version when failures obey an Exponential law. We start with a few notations before formally stating this result.

B.1 Notations

In the following, we consider a sequential or parallel job of length T_{base}. A checkpointing strategy $\mathcal{S}$ is defined as $\mathcal{S} = \{c_1, c_2, \ldots, c_m\}$, where each $c_k \in (0, T_{base})$ denotes the amount of the work executed before checkpoint number k. Note that we assume that there is a checkpoint at the end, i.e., $c_m = T_{base}$.

When a failure occurs, let $\mathbb{E}(R)$ be the expected time before the processors are ready to work again. This includes the downtime and a recovery time, but may be longer if we encounter another failure during the recovery time. For a given checkpointing strategy $\mathcal{S}$ and a work $w \in \mathcal{S} \cup \{0\}$, we denote by $\mathbb{E}([0, w], \mathcal{S})$ the expected time between the start of the job and the completion of the checkpoint corresponding to w units of work. Similarly, we denote by $\mathbb{E}([w, T_{base}], \mathcal{S})$

the expected time between the moment the checkpoint corresponding to w units of work is completed (or the start of the job if $w = 0$) and the moment the job completes, including the last checkpoint. If we do not have $w \in \mathcal{S} \cup \{0\}$, both expectations are considered infinite. With these definitions, we clearly have:

$$\forall w \in \mathcal{S}, \mathbb{E}([0, T_{base}], \mathcal{S}) = \mathbb{E}([0, w], \mathcal{S}) + \mathbb{E}([w, T_{base}], \mathcal{S}).$$

Finally, given a work $w \in [0, T_{base}]$, we let $\mathcal{S}^*_w$ be a checkpointing strategy such that for all $\mathcal{S}$, we have $\mathbb{E}([w, T_{base}], \mathcal{S}) \geq \mathbb{E}([w, T_{base}], \mathcal{S}^*_w)$. Although multiple checkpointing strategies may minimize this expectation, the value of this expectation $\mathbb{E}^*_w \triangleq \mathbb{E}([w, T_{base}], \mathcal{S}^*_w)$ is unique and well defined. Intuitively, $\mathcal{S}^*_w$ is an optimal checkpointing strategy for the end of the job after w units have been processed and checkpointed.

B.2 Main result

Theorem 2. *For a job of length T_{base}, consider the following two approaches:*

- *(A) Static Strategy: Find an optimal checkpointing heuristic $\mathcal{S}^*_0$ that minimizes the total expected makespan $\mathbb{E}^*_0$ and does not update the strategy until the job is completed.*
- *(B) Dynamic Strategy: Start with the best static strategy $\mathcal{S}^*_0$, then whenever an **event** occurs, i.e., a segment is completed or a failure happens, find an optimal static checkpointing strategy minimizing the remaining expected makespan. If the remaining expected makespan is strictly smaller with the new strategy, update the checkpointing strategy accordingly.*

The static strategy (A) and the dynamic strategy (B) are identical.

The optimal static strategy (A) is well-known and uses N_{ME} segments, where N_{ME} is given in Section 3.2 (main paper). The value of N_{ME} depends upon the length of the job that remains to be processed, so strategy (B) could compute a different value when called after the first checkpoint or the first failure. The proof shows that this is never the case.

Proof. Initially, both strategies are identical by definition. We assume that strategy (B) and strategy (A) are not always identical and obtain a contradiction. Suppose that both strategies are different. Then, there exists a failure scenario in which both strategies diverge. Consider such a scenario and let W be the total work executed and checkpointed when the first *event* e occurs, after which strategy (B) becomes different from strategy (A).

After this event e, the expected resulting makespan of strategy (A) is $\mathbb{E}^A_{remain} = t(e) + \mathbb{E}([W, T_{base}], \mathcal{S}^*_0)$, where $t(e) = 0$ if the event is the end of a segment, $t(e) = D$ if the event is a failure in the first segment for the model in which a recovery is not necessary for the first segment, and $t(e) = \mathbb{E}(R)$ otherwise. In any case, $t(e)$ is a duration independent of the checkpointing heuristic. We must finish the processing of the job as planned with strategy $\mathcal{S}^*_0$. The latter is

identical to $\mathbb{E}([W, T_{base}], \mathcal{S}_0^*)$, because the law is memoryless. Therefore, we are exactly at the same point after event e as we were when we first succeeded the checkpoint corresponding to W units of work.

Strategy (B) also needs to spend $t(e)$ units of time to deal with event e. Then, by hypothesis, strategy B finds a new checkpointing strategy such that the expected makespan for the remaining processing is reduced. As before, because the law is memoryless, an optimal strategy is $\mathcal{S}_W^*$. By assumption, the new strategy reduces the total expected makespan and we have:

$$\begin{aligned}
&t(e) + \mathbb{E}([W, T_{base}], \mathcal{S}_W^*) < \mathbb{E}_{remain}^A && \Leftrightarrow \\
&t(e) + \mathbb{E}([W, T_{base}], \mathcal{S}_W^*) < t(e) + \mathbb{E}([W, T_{base}], \mathcal{S}_0^*) && \Leftrightarrow \\
&\quad \mathbb{E}([W, T_{base}], \mathcal{S}_W^*) < \mathbb{E}([W, T_{base}], \mathcal{S}_0^*).
\end{aligned}$$

Now, suppose that we had applied the strategy $S_2 = (\mathcal{S}_0^* \setminus [W, T_{base}]) \cup (\mathcal{S}_W^* \setminus (0, W))$. The total expectation would have been:

$$\begin{aligned}
\mathbb{E}([0, T_{base}], \mathcal{S}_2) &= \mathbb{E}([0, W], \mathcal{S}_2) + \mathbb{E}([W, T_{base}], \mathcal{S}_2) \\
&= \mathbb{E}([0, W], \mathcal{S}_0^*) + \mathbb{E}([W, T_{base}], \mathcal{S}_W^*) \\
\mathbb{E}([0, T_{base}], \mathcal{S}_2) &< \mathbb{E}([0, W], \mathcal{S}_0^*) + \mathbb{E}([W, T_{base}], \mathcal{S}_0^*) \\
\mathbb{E}([0, T_{base}], \mathcal{S}_2) &< \mathbb{E}([0, T_{base}], \mathcal{S}_0^*) \\
\mathbb{E}([0, T_{base}], \mathcal{S}_2) &< \mathbb{E}_0^*
\end{aligned}$$

This contradicts the definition of $\mathcal{S}_0^*$. □

C All simulation results

The following figures are the results of the simulations with all combinations of parameters.

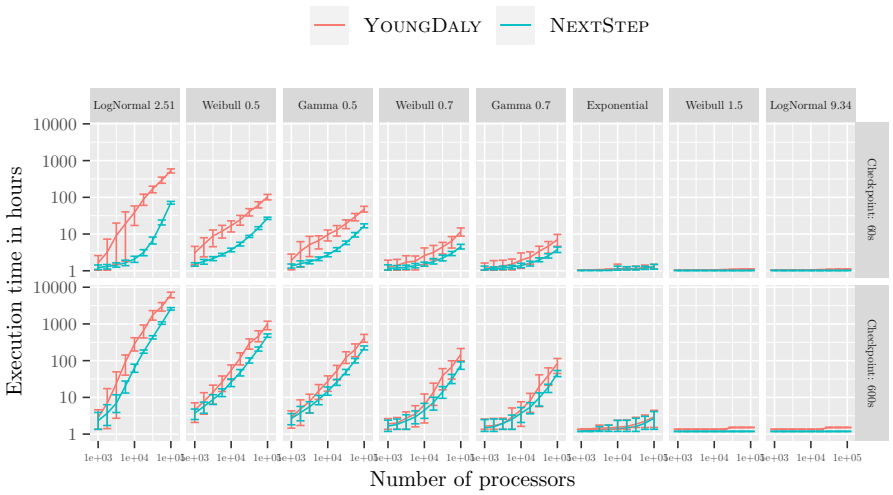


Figure 4: Expected performance of the two heuristics under all failure distributions on a 0 day old platform and with a workflow W of 1 hour. Top: $C = 60$, bottom: $C = 600$.

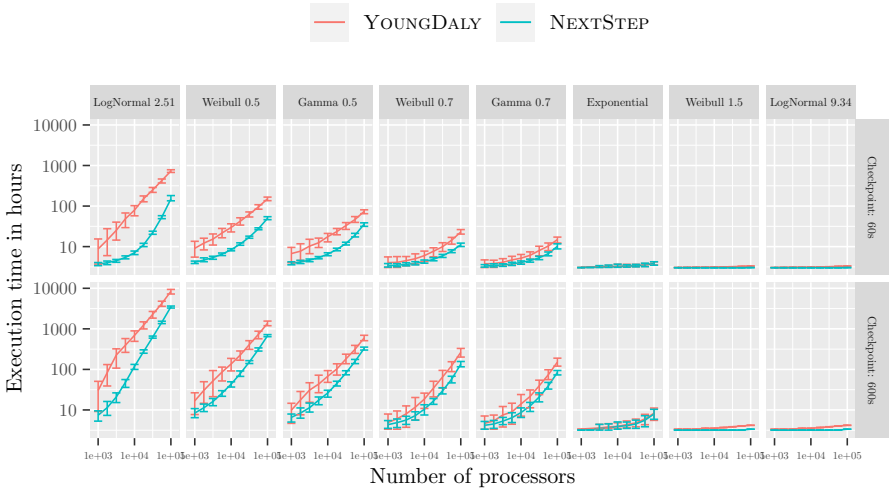


Figure 5: Expected performance of the two heuristics under all failure distributions on a 0 day old platform and with a workflow W of 3 hours. Top: $C = 60$, bottom: $C = 600$.

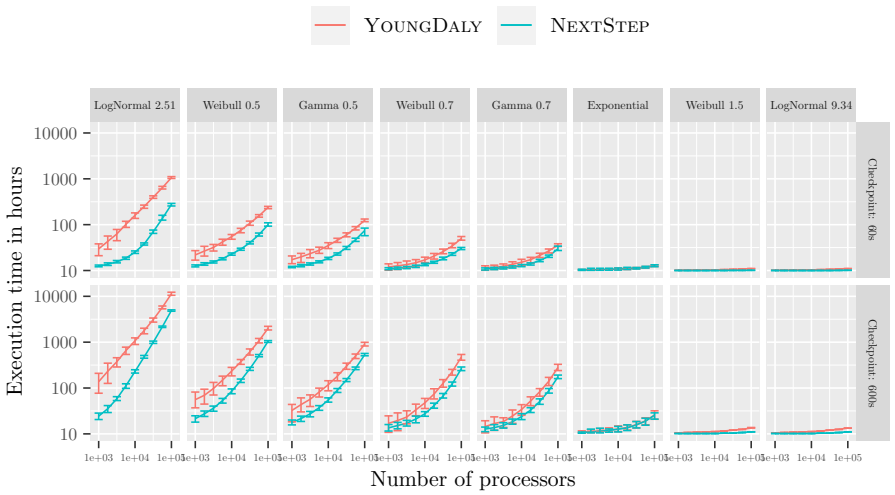


Figure 6: Expected performance of the two heuristics under all failure distributions on a 0 day old platform and with a workflow W of 10 hours. Top: $C = 60$, bottom: $C = 600$.

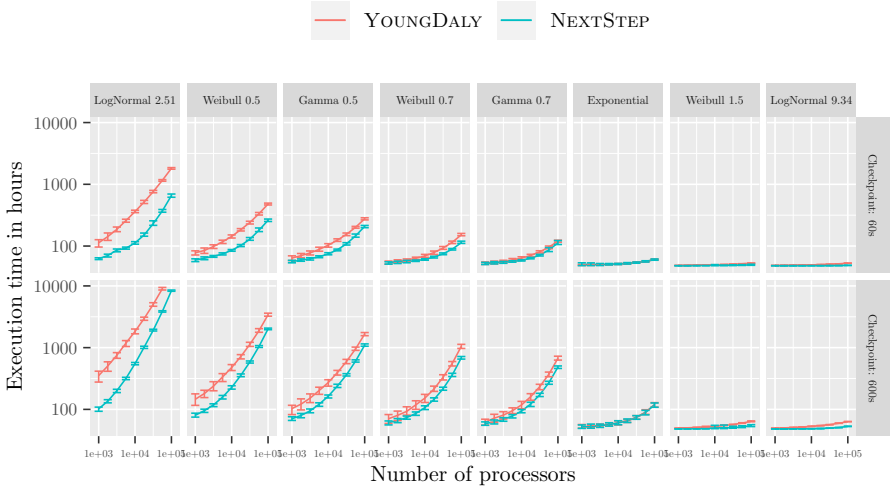


Figure 7: Expected performance of the two heuristics under all failure distributions on a 0 day old platform and with a workflow W of 48 hours. Top: $C = 60$, bottom: $C = 600$.

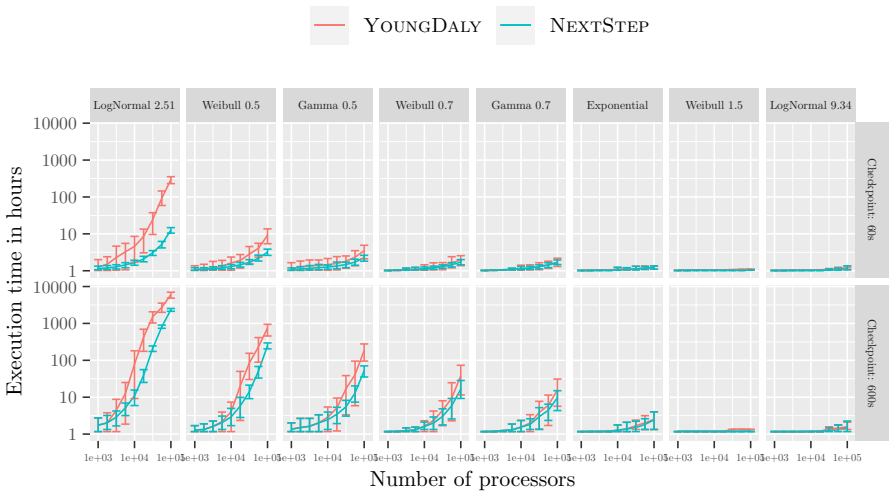


Figure 8: Expected performance of the two heuristics under all failure distributions on a 10 day old platform and with a workflow W of 1 hour. Top: $C = 60$, bottom: $C = 600$.

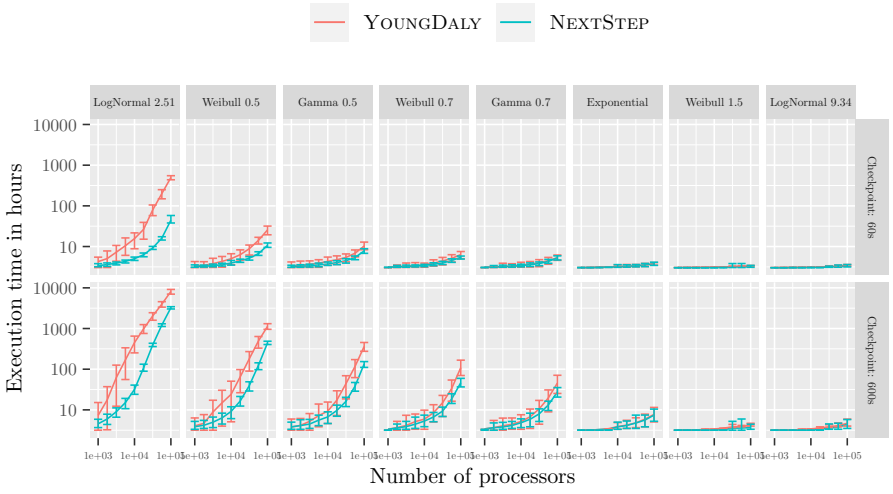


Figure 9: Expected performance of the two heuristics under all failure distributions on a 10 day old platform and with a workflow W of 3 hours. Top: $C = 60$, bottom: $C = 600$.

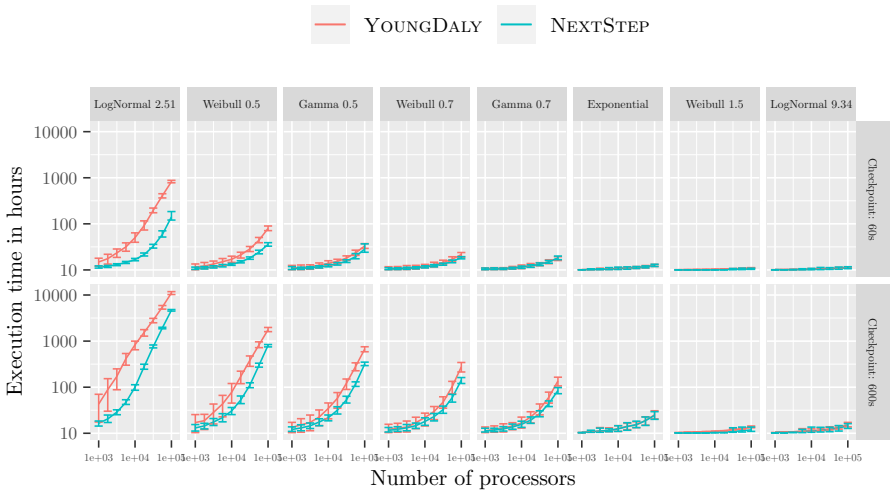


Figure 10: Expected performance of the two heuristics under all failure distributions on a 10 day old platform and with a workflow W of 10 hours. Top: $C = 60$, bottom: $C = 600$.

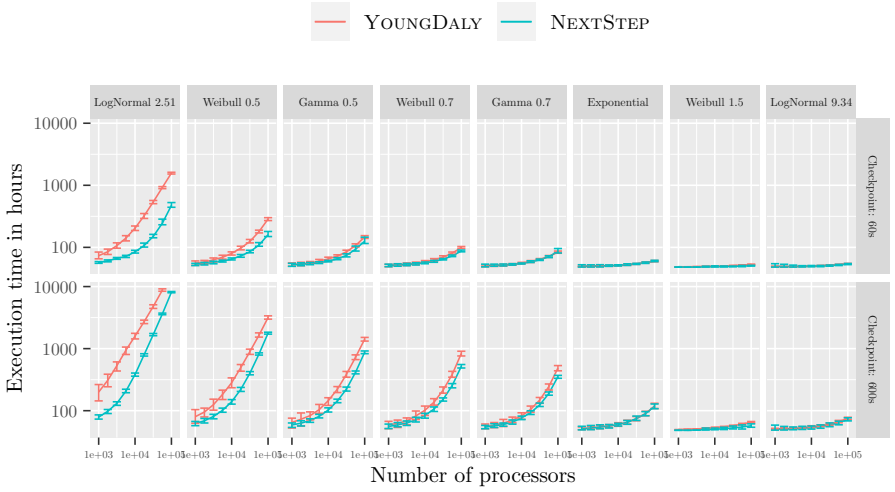


Figure 11: Expected performance of the two heuristics under all failure distributions on a 10 day old platform and with a workflow W of 48 hours. Top: $C = 60$, bottom: $C = 600$.

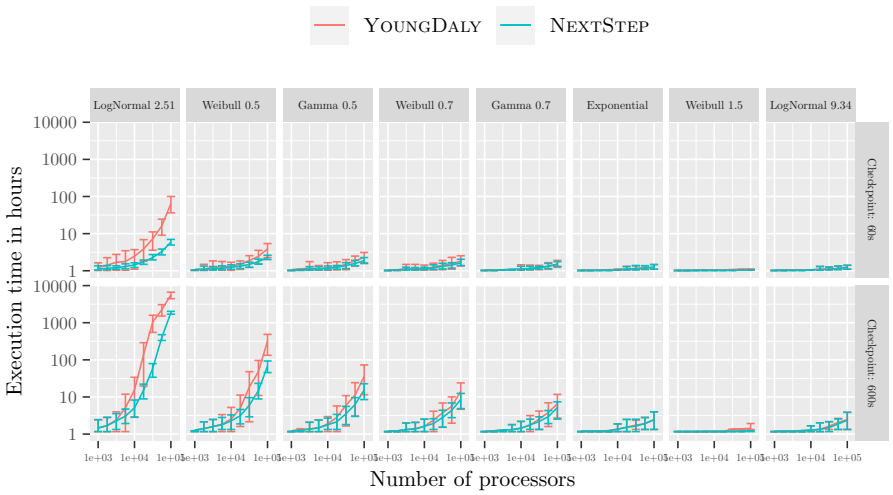


Figure 12: Expected performance of the two heuristics under all failure distributions on a 30 day old platform and with a workflow W of 1 hour. Top: $C = 60$, bottom: $C = 600$.

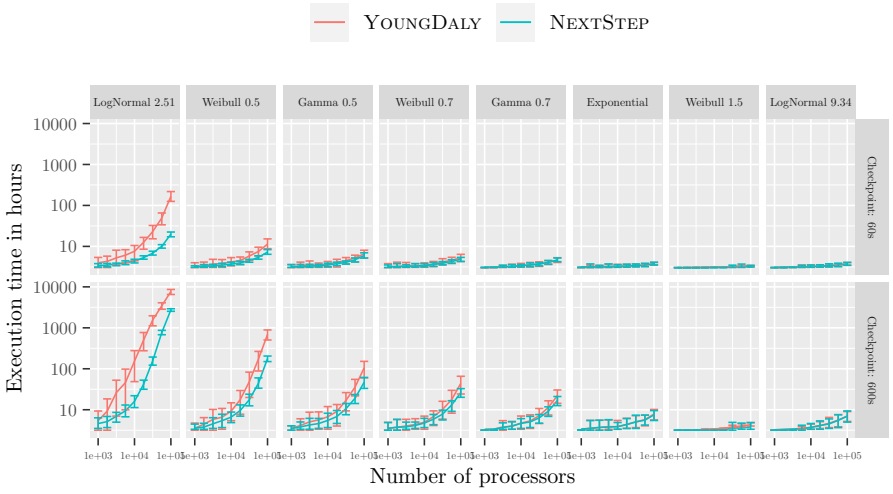


Figure 13: Expected performance of the two heuristics under all failure distributions on a 30 day old platform and with a workflow W of 3 hours. Top: $C = 60$, bottom: $C = 600$.

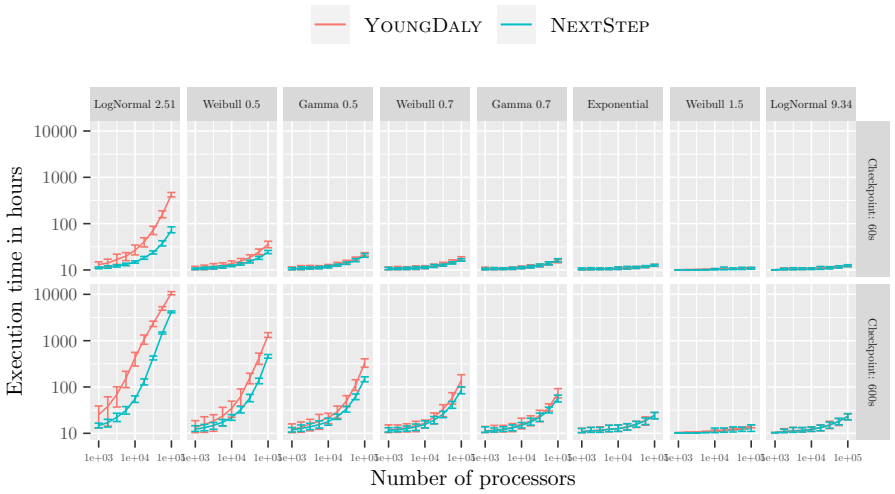


Figure 14: Expected performance of the two heuristics under all failure distributions on a 30 day old platform and with a workflow W of 10 hours. Top: $C = 60$, bottom: $C = 600$.

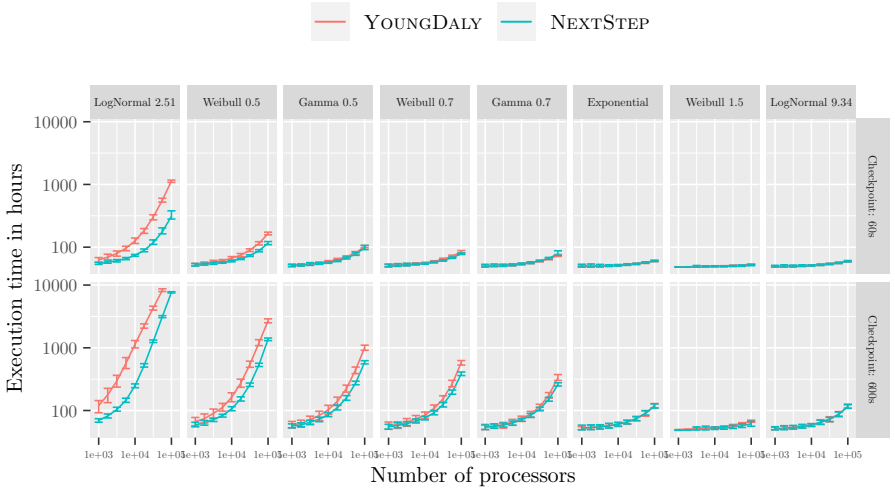


Figure 15: Expected performance of the two heuristics under all failure distributions on a 30 day old platform and with a workflow W of 48 hours. Top: $C = 60$, bottom: $C = 600$.

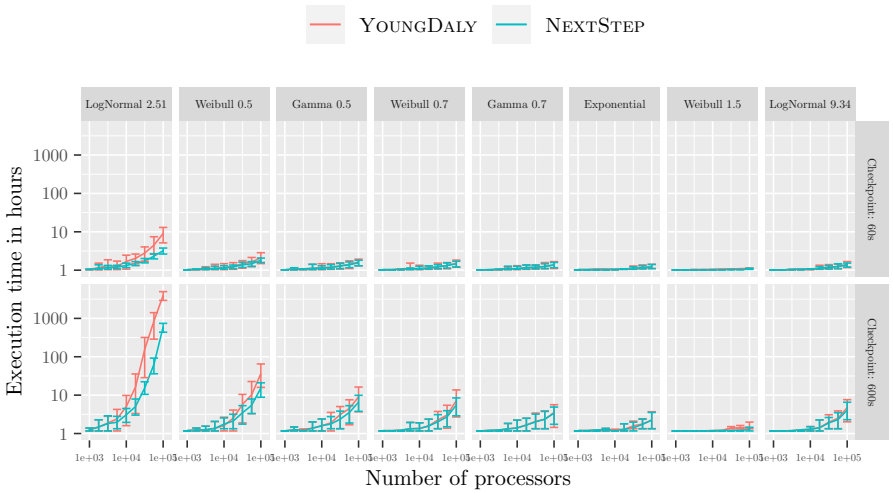


Figure 16: Expected performance of the two heuristics under all failure distributions on a 100 day old platform and with a workflow W of 1 hour. Top: $C = 60$, bottom: $C = 600$.

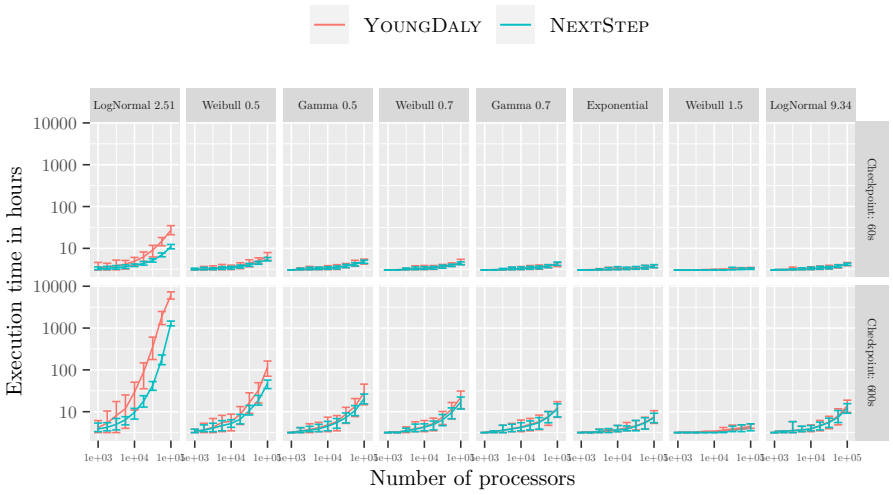


Figure 17: Expected performance of the two heuristics under all failure distributions on a 100 day old platform and with a workflow W of 3 hours. Top: $C = 60$, bottom: $C = 600$.

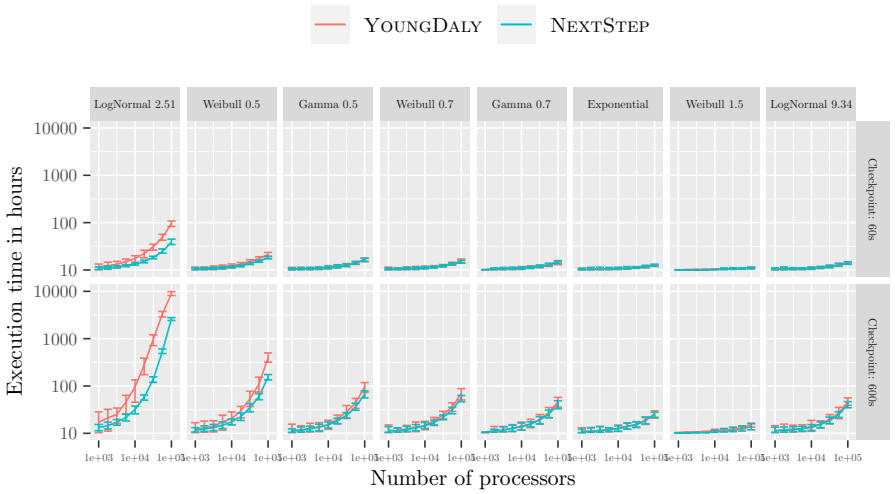


Figure 18: Expected performance of the two heuristics under all failure distributions on a 100 day old platform and with a workflow W of 10 hours. Top: $C = 60$, bottom: $C = 600$.

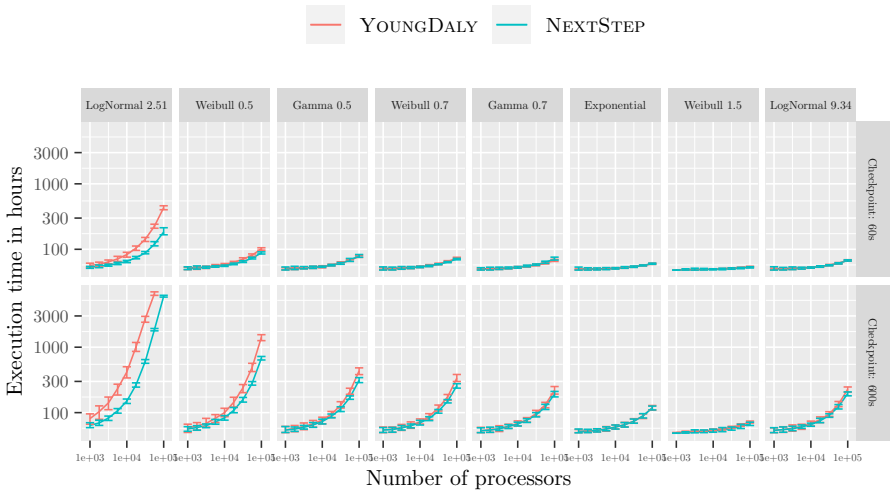


Figure 19: Expected performance of the two heuristics under all failure distributions on a 100 day old platform and with a workflow W of 48 hours. Top: $C = 60$, bottom: $C = 600$.

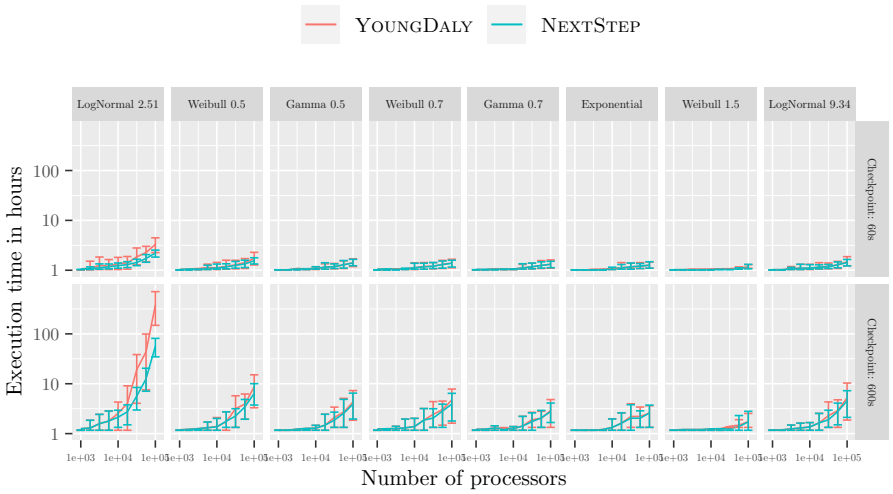


Figure 20: Expected performance of the two heuristics under all failure distributions on a 365 day old platform and with a workflow W of 1 hour. Top: $C = 60$, bottom: $C = 600$.

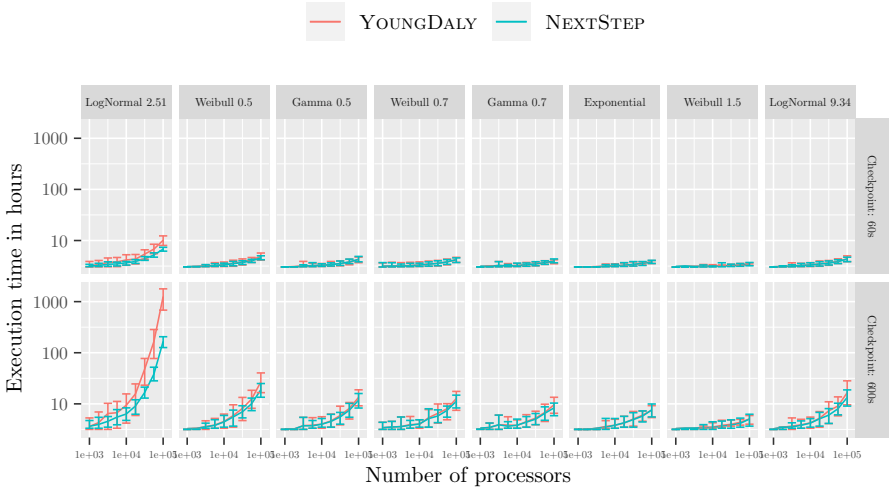


Figure 21: Expected performance of the two heuristics under all failure distributions on a 365 day old platform and with a workflow W of 3 hours. Top: $C = 60$, bottom: $C = 600$.

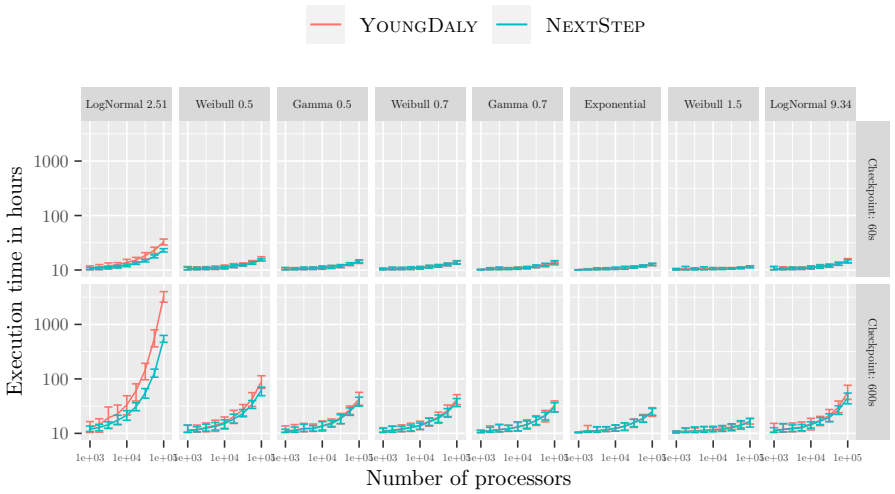


Figure 22: Expected performance of the two heuristics under all failure distributions on a 365 day old platform and with a workflow W of 10 hours. Top: $C = 60$, bottom: $C = 600$.

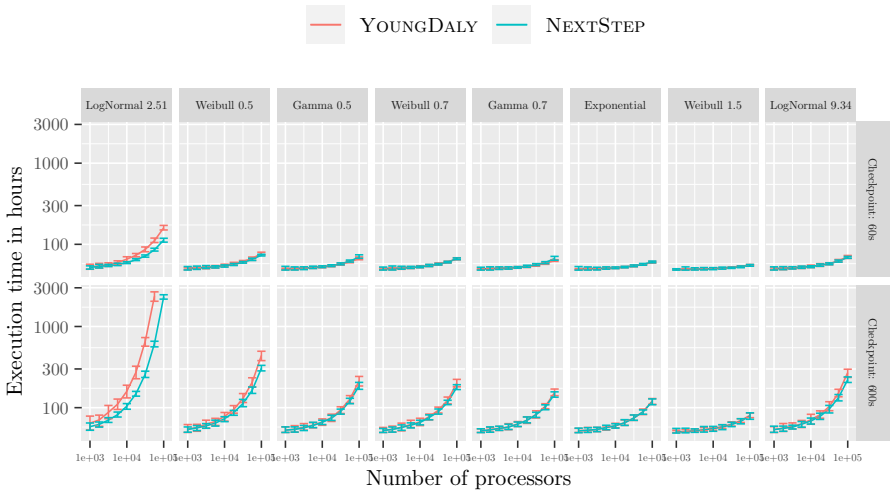
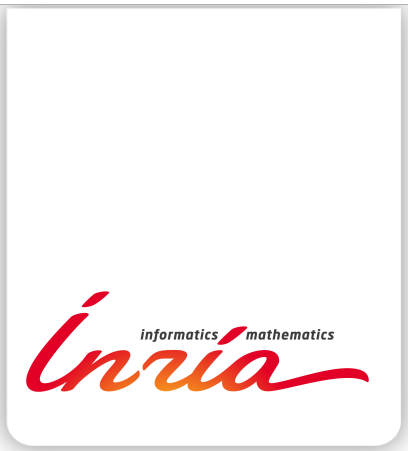


Figure 23: Expected performance of the two heuristics under all failure distributions on a 365 day old platform and with a workflow W of 48 hours. Top: $C = 60$, bottom: $C = 600$.



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