HAL open science

Design and Comparison of Resilient Scheduling Heuristics for Parallel Jobs

Anne Benoit, Valentin Le Fèvre, Padma Raghavan, Yves Robert, Hongyang Sun

To cite this version:

Anne Benoit, Valentin Le Fèvre, Padma Raghavan, Yves Robert, Hongyang Sun. Design and Comparison of Resilient Scheduling Heuristics for Parallel Jobs. [Research Report] RR-9296, Inria - Research Centre Grenoble – Rhône-Alpes. 2019, pp.31. hal-02317464v2

HAL Id: hal-02317464

https://inria.hal.science/hal-02317464v2

Submitted on 21 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Design and Comparison of Resilient Scheduling Heuristics for Parallel Jobs

Anne Benoit, Valentin Le Fèvre, Padma Raghavan, Yves Robert, Hongyang Sun

RESEARCH REPORT
N° 9296
February 2020
Project-Team ROMA

ISSN 0249-6399 ISRN INRIA/RR--9296--FR+ENG

Design and Comparison of Resilient Scheduling Heuristics for Parallel Jobs

Anne Benoit*, Valentin Le Fèvre*, Padma Raghavan†, Yves Robert*‡, Hongyang Sun†

Project-Team ROMA

Research Report n° 9296 — February 2020 — 31 pages

Abstract: This paper focuses on the resilient scheduling of parallel jobs on high-performance computing (HPC) platforms to minimize the overall completion time, or the makespan. We revisit the classical problem while assuming that jobs are subject to transient or silent errors, and hence may need to be re-executed each time they fail to complete successfully. This work generalizes the classical framework where jobs are known offline and do not fail: in this framework, list scheduling that gives priority to longest jobs is known to be a 3-approximation when imposing to use shelves, and a 2-approximation without this restriction. We show that when jobs can fail, using shelves can be arbitrarily bad, but unrestricted list scheduling remains a 2-approximation. The paper focuses on the design of several heuristics, some list-based and some shelf-based, along with different priority rules and backfilling strategies. We assess and compare their performance through an extensive set of simulations, using both synthetic jobs and log traces from the Mira supercomputer.

Key-words: Resilient scheduling, parallel jobs, silent errors, list schedules, shelf schedules, approximation ratios.

* LIP, École Normale Supérieure de Lyon, CNRS & Inria, France
† Vanderbilt University, Nashville, TN, USA
‡ University of Tennessee Knoxville, TN, USA

RESEARCH CENTRE GRENOBLE – RHÔNE-ALPES

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

Heuristiques d'ordonnancement avec tolérance aux pannes pour des tâches parallèles

Résumé : Ce rapport s'intéresse à l'ordonnancement résilient de tâches parallèles sur des plates-formes de calcul haute performance (HPC), avec pour but de minimiser le temps totsal d'exécution. Nous considérons que les tâches sont confrontées à des erreurs silencieuses, et il faut donc les ré-exécuter après chaque faute afin d'avoir une exécution correcte. Ces travaux généralisent le cadre classique où les tâches sont connues avant l'exécution et ne sont pas confrontées à des erreurs. Dans le cas sans erreurs, les ordonnancements de liste donnant la priorité aux plus longues tâches sont connus pour être une 3-approximation pour un ordonnancement par étagères, et une 2-approximation sans la restriction des étagères. Nous montrons qu'avec des erreurs, l'utilisation d'étagères peut aboutir à un ordonnancement arbitrairement loin de l'optimal, alors que l'ordonnancement de liste classique reste une 2-approximation. Nous concevons plusieurs heuristiques, à base de listes ou d'étagères, avec différentes règles de priorité et de réservation. Nous évaluons et comparons leur performance dans une campagne de simulations, en utilisant à la fois des tâches générées aléatoirement, et des tâches d'applications exécutées sur le super-calculateur Mira.

Mots-clés : Ordonnancement tolérant aux pannes, tâches parallèles, erreurs silencieuses, ordonnancement de liste, ordonnancement par étagères, facteurs d'approximation.

1 Introduction

One of the main challenges faced by today's HPC platforms is resilience, since such platforms are confronted with many failures or errors due to their large scale [32]. Indeed, the number of failures is known to grow proportionally with the number of nodes on a platform [22], and the largest supercomputers today experience several failures per day. There are two main classes of errors that can cause failures in an application's execution, namely, fail-stop and silent errors. While fail-stop errors cause the execution to terminate (e.g., due to hardware fault), large-scale platforms are also confronted with *silent errors*, or *silent data corruptions (SDCs)*. Such errors are caused by cosmic radiation or packaging pollution, striking either the cache or memory units (bit flips), or the CPU operations [36, 49]. Even though any bit can be corrupted, the execution continues (unlike fail-stop errors), hence the error is transient, but it may dramatically impact the result of a running application. Many silent errors can be accurately detected by verifying the data using dedicated, lightweight detectors (e.g., [23, 46, 9, 4, 20, 8]). In this work, we focus on job failures caused by silent errors, and we aim to design resilient scheduling heuristics while assuming the availability of ad-hoc detectors to detect such errors.

The problem of scheduling a set of independent jobs on parallel platforms with the goal of minimizing the total completion time, or the *makespan*, has been extensively studied (see Section 2). Jobs may be parallel and should be executed on a given number of processors for a certain duration; both the processor requirement and the execution time of each job are known at the beginning. Such jobs are called *rigid* jobs, contrarily to moldable or malleable jobs, whose processor allocations can vary at launch time or during execution [13]. While moldable or malleable jobs offer more flexibility in the execution, rigid jobs remain the most prevalent form of parallel jobs submitted on today's HPC systems, and we focus on rigid jobs in this paper.

Unlike the classical scheduling problem without job failures, we consider *failure-prone platforms* subject to silent errors. Hence, at the end of each job's execution, an SDC detector will flag if a silent error has occurred during its execution. In this case, the job must be re-executed until it has been successfully completed without errors. For a set of jobs, each execution may lead to a different failure scenario, depending upon the jobs that have experienced failures as well as the number of such failures. The objective is to minimize the makespan under any failure scenario, as well as the *expected makespan*, averaged over all possible failure scenarios, where each scenario is weighted by a probability that governs its occurrence under certain failure assumptions. Since a failure scenario is unknown a priori, the scheduling decisions must be made *dynamically* on-the-fly, whenever an error has been detected. As a result, even for the same set of jobs, different schedules may be produced, depending on the failure scenario that occurred in a particular

execution.

Building upon the existing framework for scheduling parallel jobs without failures, we propose two scheduling strategies, namely, a *list-based* strategy and a *shelf-based* strategy. While list-based schedules have no restrictions on the starting times of the jobs, shelf-based schedules group all jobs into subsets of jobs having the same starting time (called shelves); a shelf of jobs can start its execution once the longest job from the previous shelf has completed. For list-based scheduling, practical systems also employ a combination of reservation and backfilling strategies with different job priority rules to increase the system utilization. On platforms with no failures, variants for all of these strategies exist that could achieve constant approximations for the makespan (see Section 2 for details). The main focus of this paper is to extend these existing heuristics to execution scenarios with job failures, and to experimentally compare their performance using a variety of job and platform configurations.

Our main contributions are the following:

- We propose a formal model for the problem of resilient scheduling of parallel jobs on failure-prone platforms. The model formulates the performance of an algorithm under both worst-case and expected executions.
- We design a resilient list-based strategy, and prove that its greedy variant achieves $(2 - \frac{1}{P})$-approximation, and its reservation variant is $(3 - \frac{4}{P+1})$-approximation, where P is the total number of processors. These results apply to both worst-case and expected makespans.
- We design a resilient shelf-based strategy, but we show that, under some failure scenarios, any shelf-based algorithm has an unbounded approximation ratio, thus having a makespan that is arbitrarily higher than the optimal makespan in the worst case.
- We conduct an extensive set of simulations to evaluate and compare different variants of these heuristics using both synthetic jobs and log traces from the Mira supercomputer. The results show that the performance of these resilient scheduling heuristics is close to the optimal in practice, even when confronted with failures.

The rest of this paper is organized as follows: Section 2 describes the background of parallel job scheduling and presents some related work. The formal models and the problem statement are presented in Section 3. The key contributions of the paper are presented in Section 4, where we describe both list-based and shelf-based strategies, and analyze their performance. Section 5 presents an extensive set of simulation results and highlights the main findings. Finally, Section 6 concludes the paper and discusses future directions.

2 Background and related work

This section describes the background of scheduling rigid parallel jobs and reviews some related work. We start with a brief description of the different scheduling flavors and strategies in Section 2.0.1. In Section 2.0.2, we discuss the simpler offline problem, where all jobs are known statically and available initially. Taking job failures into account calls for a dynamic schedule, because re-executions are decided on-the-fly after the completion of each job. We then review the online problem, where jobs are presented dynamically to the scheduler in Section 2.1. Our problem with job failures is harder than the offline problem, and is different from the online problem with jobs submitted at arbitrary but fixed release times. Practical schedulers often use reservation and backfilling, and we review related work in this area in Section 2.2. Finally, with failures, job execution times are no longer deterministic, and we review scheduling strategies for stochastic jobs in Section 2.3.

2.0.1 Different scheduling flavors and strategies

Historically, scheduling parallel jobs comes in two flavors: if a job requests p processors, either any subset of p processors can be assigned, or only subsets of p *contiguous* processors can be chosen. In the latter case, processors are organized as a linear array and labeled from 1 to P, where P is the total number of processors; then only neighboring processors (whose labels differ by one) can be assigned to a job. The *contiguous* variant is equivalent to the *rectangle strip packing* problem, where rectangles are to be stacked (without rotation) within a strip of width P: rectangle widths represent processor numbers, and rectangle heights represent execution times.

Most scheduling strategies also come in two flavors: either the schedule is restricted to building *shelves* (also referred to as *levels* in some literature), or it is unrestricted, in which case the jobs are often scheduled based on an ordered *list*. Shelves are subsets of jobs with the same starting time, and for which each of the P processors is used at most once: the height of a shelf is the length of its longest job; when the shorter jobs complete, their processors become idle, but these processors are not reassigned to other jobs until the completion of the longest job of the shelf. Thus, a shelf resembles a bookshelf, hence the name. Shelf-based schedules play an important role in HPC, because they correspond to batched execution scenarios, where jobs are grouped into batches that are scheduled one after another. Note that for shelf-based algorithms, the contiguous and non-contiguous variants collapse.

2.0.2 Offline scheduling of rigid jobs

To minimize the makespan for a set of rigid jobs that are known statically and available initially (i.e., offline), the problem is obviously NP-complete, as

it generalizes the problem of scheduling independent jobs on two processors, a variant of the 2-PARTITION problem [17]. Coffman et al. [10] showed that the Next-Fit Decreasing Height (NFDH) algorithm is 3-approximation, and the First-Fit Decreasing-Height (FFDH) algorithm is 2.7-approximation. Both algorithms are shelf-based. See the survey by Lodi et al. [31] for more results and lower bounds on the best possible approximation ratio for shelf-based algorithms, and see Han et al. [21] for the intricate relationship between strip packing and bin packing.

For list-based scheduling, Baker et al. [3] showed that the Bottom-up Left-justified (BL) heuristic while ordering the jobs in decreasing processor requirement achieves 3-approximation. Turek et al. [42] showed that ordering jobs in decreasing execution time is also 3-approximation. Moreover, both algorithms guarantee contiguous processor allocations for all jobs. Without the contiguous processor constraint, several works [42, 16, 15] showed that the greedy list-scheduling heuristic achieves 2-approximation. Finally, Jansen [26] presented a $(3/2 + \epsilon)$-approximation algorithm for any fixed $\epsilon > 0$. This is the best result possible, since a lower bound on the approximation ratio is $3/2$, which holds even when considering asymptotic performance [27].

When jobs have precedence constraints among them, list scheduling is shown to be P-approximation in the worst case, which holds for many commonly used job-ordering rules [29, 14]. However, if jobs require no more than qP processors for any $0 < q < 1$, then the approximation ratio of greedy list scheduling is $\frac{(2-q)}{(1-q)P+1}$ [29, 14]. In our problem, a particular failure instance can be regarded as a special case of the general precedence constraint, where each job forms a linear chain, but the failure instance is unknown to the scheduler beforehand.

2.1 Online scheduling of rigid jobs

In the Online problem, a set of rigid jobs arrive dynamically over time and information of a job is not known until the job has arrived. In this case, the list-based greedy algorithm maintains a competitive ratio of 2 [34, 27]. Chen and Vestjens [7] showed a 1.3473 lower bound on the competitive ratio of any deterministic online algorithm even when all jobs are sequential. Shmoys et al. [38] showed that by collecting all jobs that arrive during a batch and then scheduling them together in the next batch, one can convert any c-approximation offline algorithm to a $2c$-competitive online algorithm.

In another online model referred to as One-By-One, jobs, although all arrive initially, are presented one at a time to the online scheduler and must be irrevocably scheduled before the next job can be revealed. Johannes [27] showed that list-based greedy algorithm is P-competitive in the worst case, and presented a 12-competitive algorithm. Baker and Schwarz [2] extended the two shelf-based algorithms presented in [10] and showed that

Next-Fit is 7.46-competitive and First-Fit is 6.99-competitive. The surveys by Csirik and Woeginger [11, 12] describe more results and lower bounds that use shelf-based algorithms in this model. The best known competitive ratio for One-By-One is 6.6623, obtained by Hurink and Paulus [24] and independently by Ye et al. [47].

The problem studied in this paper can be considered as semi-online, since all jobs are known to the scheduler initially but not their failure scenarios. We point out that the technique by Shmoys et al. [38] to obtain $2c$-competitiveness is not applicable here, because it relies on jobs having fixed, although unknown, release times, whereas the "new job arrival" times in our problem (corresponding to failed jobs restarting) depends on the decisions made on-the-fly by the schedulers.

2.2 Batch schedulers in practical systems

In practical systems, parallel jobs are often scheduled by batch schedulers [25, 48, 41] that use a combination of *reservation* and *backfilling* strategies: while the high-priority jobs are scheduled by reserving processors in advance, the low-priority ones are used to fill in the "holes" to improve system utilization. Two popular backfilling strategies are *conservative* [33] and *aggressive* (a.k.a. *EASY*) [30, 39]. The former gives a reservation for every job in the queue and a lower-priority job is moved forward as long as it does not delay the reservation for any higher-priority job. The latter only gives reservation to the job at the head of the queue (i.e., the one with the highest priority) and backfilling is allowed without delaying this highest-priority job. Note that the list-based greedy scheduling can be considered as an even more aggressive strategy, where no job receives a reservation and all jobs are scheduled using backfilling. As jobs arrive over time, most practical schedulers use First-Come First-Serve (FCFS) in conjunction with these strategies to prevent job starvation, but no worst-case performance guarantee is known for such schedulers. Various priority rules have been evaluated to characterize and tune their performance for different performance metrics (see, e.g., [40, 18, 45]).

2.3 Scheduling stochastic jobs

When a job could fail during execution and has to be restarted, it can be regarded as a stochastic job, whose execution time depends on the number of failures. Most prior works on stochastic scheduling have considered *sequential* jobs whose execution times follow a known probability distribution. The book by Pinedo [37] and the survey by Niño-Mora [35] discuss many relevant results on stochastic scheduling. For offline problems (i.e., no new job arrival), the literature has focused on two models. In the *static* model, all scheduling decisions (i.e., job assignments to processors) are made

beforehand, whereas in the *dynamic* model, scheduling decisions are made dynamically on the fly. While both models coincide when job execution times are deterministic, they lead to different results for stochastic jobs. Under the static model, Kleinberg et al. [28] showed an $O(1)$-approximation algorithm for jobs with arbitrary distributions. Goel and Indyk [19] obtained a 2-approximation for jobs with Poisson distribution and a PTAS for exponential distribution. Under the dynamic model, the *Longest Expected Processing Time* first (LEPT) algorithm is known to achieve the optimal expected makespan for jobs with exponential distributions [5, 44] or when all jobs follow a common distribution with a non-increasing hazard rate function [43]. For jobs with arbitrary distributions, a straightforward extension of the classical online list scheduling yields a 2-approximation [6].

In this paper, we adopt the dynamic stochastic scheduling model to handle parallel jobs with failures. However, there are two main differences: job execution times follow a discrete distribution, and a failure does not require the job to be immediately re-executed. We prove a 2-approximation for a greedy algorithm in terms of expected makespan, and experimentally evaluate several list-based and shelf-based heuristics with different priority rules and backfilling options.

3 Models

In this section, we formally present the models and the problem statement. We also state the main assumptions we make in this paper.

3.1 Job model

We consider a set $\mathcal{J} = \{J_1, J_2, \ldots, J_n\}$ of n parallel jobs to be executed on a platform consisting of P identical processors. All jobs are released at the same time, corresponding to the batch scheduling scenario in an HPC environment. In this paper, we focus on *rigid* jobs, which must be executed with a fixed number of processors that is usually set by the user when the job is submitted[1]. For each job $J_j \in \mathcal{J}$, let $p_j \in \{1, 2, \ldots, P\}$ denote its fixed (integral) processor allocation, and let t_j denote its error-free execution time. The *area* of the job is defined as $a_j = p_j \times t_j$.

3.2 Error model

We consider job failures that manifest as *silent errors* or *silent data corruptions (SDCs)* [32] that could corrupt a job during execution. A silent error detector is assumed to be available for each job, which is triggered at

[1]Other parallel job models include *moldable* and *malleable* models, which allow a job's processor allocation to vary at launch time or during execution [13]. Considering alternative job models will be part of our future work.

the end of the job's execution. If an error is detected, the job needs to be re-executed, followed by another error detection. This process repeats until the job completes successfully without errors. Current state-of-the-art SDC detectors are typically lightweighted (e.g., ABFT for matrix computations [23, 46, 9], or data analytics for scientific applications [4, 20, 8]), and hence incur a negligible cost compared to the overall execution time of the job.

All the list-based and shelf-based scheduling heuristics introduced and compared in this paper are agnostic of the probability of each job to fail any given number of times. Specifically, for a job J_j, consider a particular run where it fails f_j times before succeeding on the $(f_j + 1)$-th execution. The probability that this happens is denoted as $q_j(f_j)$. Let $\mathbf{f} = (f_1, f_2, \ldots, f_n)$ denote a *failure scenario*, i.e., a vector of the number of failed execution attempts for all jobs, during a particular run. Assuming that errors occur independently for different jobs, the probability that this combined failure scenario happens can be computed as $Q(\mathbf{f}) = \prod_{j=1\ldots n} q_j(f_j)$. The failure scenario $\mathbf{f}$, as well as the associated probabilities $q_j(f_j)$ and $Q(\mathbf{f})$ may be unknown to the scheduler.

3.3 Problem statement

We study the following resilient scheduling problem: Given a set $\mathcal{J}$ of parallel jobs, find a schedule for $\mathcal{J}$ on P identical processors under any failure scenario $\mathbf{f} = (f_1, f_2, \ldots, f_n)$. Here, a *schedule* for $\mathbf{f}$ is defined by a collection $\mathbf{s} = (\vec{s}_1, \vec{s}_2, \ldots, \vec{s}_n)$ of starting time vectors for all jobs, where vector $\vec{s}_j = (s_j^{(1)}, s_j^{(2)}, \ldots, s_j^{(f_j+1)})$ specifies the starting times for job J_j at different execution attempts until success.

The objective is to minimize the overall completion time of all jobs, or the *makespan*. Suppose an algorithm ALG makes scheduling decision $\mathbf{s}$ during the failure scenario $\mathbf{f}$, then the makespan of the algorithm for this scenario is defined as:

$$T_{\text{ALG}}(\mathbf{f}, \mathbf{s}) = \max_{j=1\ldots n} \left(s_j^{(f_j+1)} + t_j \right) . \tag{1}$$

All scheduling decisions should be made while satisfying the following two constraints:

1. The number of processors utilized at any time t by the set $\mathcal{J}_t$ of running jobs should not exceed the total number P of available processors on the platform, i.e.,

$$\sum\nolimits_{J_j \in \mathcal{J}_t} p_j \leq P, \quad \forall t. \tag{2}$$

2. A job cannot be re-executed if its previous execution attempt has not yet been completed, i.e.,

$$s_j^{(i+1)} \geq s_j^{(i)} + t_j, \quad \forall j = 1 \ldots n, \quad \forall i \geq 1. \tag{3}$$

This scheduling problem, encompassing the failure-free problem as a special case, is clearly NP-hard. A scheduling algorithm ALG is said to be *c-approximation* if its makespan is at most c times that of an optimal scheduler for all possible sets of jobs, and for all possible failure scenarios, i.e.,

$$T_{\text{ALG}}(\mathbf{f}, \mathbf{s}) \leq c \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*), \tag{4}$$

where $T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*)$ denotes the optimal makespan with scheduling decision $\mathbf{s}^*$ under failure scenario $\mathbf{f}$. Clearly, this optimal makespan admits the following two lower bounds:

$$T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \geq t_{\max}(\mathbf{f}) \ , \tag{5}$$

$$T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \geq \frac{A(\mathbf{f})}{P} \ , \tag{6}$$

where $t_{\max}(\mathbf{f}) = \max_{j=1\ldots n}(f_j + 1) \cdot t_j$ is the maximum cumulative execution time of any job under $\mathbf{f}$, and $A(\mathbf{f}) = \sum_{j=1}^{n}(f_j + 1) \cdot a_j$ is the total cumulative area.

In Section 4, we establish several approximation results, which are valid for any failure scenario, regardless of its individual probability. This is the strongest result that can be obtained from a theoretical perspective. However, from a practical perspective, given a set of jobs, it is not easy to assess the performance of a scheduling heuristic if the probability $Q(\mathbf{f}) = \prod_{j=1\ldots n} q_j(f_j)$ of each failure scenario $\mathbf{f}$ is not known. Thus, for the experiments in Section 5, we report the expected cost of each heuristic under the standard Exponential probability distribution, as explained below.

3.4 Expected makespan

Suppose the occurrence of silent errors striking the jobs follows an Exponential probability distribution, and that the mean time between error (MTBE) of an individual processor is μ, so the error rate of the processor is given by $\lambda = 1/\mu$. For a job J_j executed on p_j processors, the probability that the job is struck by a silent error during execution is then given by $q_j = 1 - e^{-\lambda p_j \cdot t_j} = 1 - e^{-\lambda a_j}$ [22]. Then, the probability for job J_i to fail f_j times before succeeding on the $(f_j + 1)$-th execution is $q_j(f_j) = q_j^{f_j}(1 - q_j)$.

Given a set $\mathcal{J}$ of parallel jobs, we can now define the *expected makespan* of an algorithm ALG, which is taken over all possible failure scenarios weighted by their probabilities, as:

$$\mathbb{E}(T_{\text{ALG}}) = \sum\nolimits_{\mathbf{f}} Q(\mathbf{f}) \cdot T_{\text{ALG}}(\mathbf{f}, \mathbf{s}) \ . \tag{7}$$

In this case, an algorithm is a c-approximation if we have:

$$\mathbb{E}(T_{\text{ALG}}) \leq c \cdot \mathbb{E}(T_{\text{OPT}}) \ , \tag{8}$$

for all possible sets of jobs, where $\mathbb{E}(T_{\text{OPT}})$ denotes the optimal expected makespan. This is because the inequality is true for each failure scenario, hence for the weighted sum. Obviously, the converse is not true: an algorithm could satisfy Equation (8) (thus being a c-approximation in expectation) but be arbitrarily worse than the optimal on some (low probability) failure scenarios. Still, expected makespans provide a synthetic indicator on the performance of an algorithm under study and enable easy, quantitative comparisons. Thus, we use them for the experimental evaluations in Section 5.

3.5 Static vs. dynamic scheduling

As all the information regarding the set of jobs is available, one approach is to make all scheduling decisions (i.e., starting times $\mathbf{s}$) *statically* at the beginning, and then execute the jobs according to this static schedule. While this approach works for failure-free scenarios, it is problematic when jobs can fail and re-execute. In particular, a static schedule needs to pre-compute a long (possibly infinite) sequence of starting times for all jobs to account for every possible failure scenario, while ensuring the satisfaction of the scheduling constraints. Pre-computing such a static schedule would be very computationally inefficient, especially when there turn out to be only a few failures in a particular run.

In contrast, another more flexible approach is to make scheduling decisions *dynamically* depending on the particular failure scenario that is unveiled from an execution. For example, a scheduling algorithm may decide the starting time for the next execution attempt of a job depending on all job failure scenarios and schedules so far. As a result, even for the same set of jobs, the algorithm may produce different schedules in response to the different failure scenarios that could arise during runtime. In this paper, we adopt this dynamic approach.

4 Resilient scheduling to cope with job failures

In this section, we present a resilient list-based heuristic (R-List) and a resilient shelf-based heuristic (R-Shelf) for scheduling parallel jobs that could fail due to silent errors. We show that the greedy variant of R-List without reservations is a 2-approximation, and a variant with reservations is a 3-approximation. For R-Shelf, even though it provides a 3-approximation in the failure-free case, we show through an example that any resilient shelf-based algorithm may have a makespan ratio of $\Omega(\ln P)$ compared to the optimal in some failure scenario.

4.1 R-List scheduling heuristic

We present a resilient list-based scheduling algorithm, called R-List, that schedules any set of parallel jobs with the capability to handle failures. Algorithm 1 shows the pseudocode of R-List. It extends the classical batch scheduler that combines reservation and backfilling strategies. The algorithm first organizes all jobs in a list (or a queue) based on some priority rule. Then, whenever an existing job J_k completes and hence releases processors (at time 0, a virtual job J_0 can be considered to complete), the algorithm schedules the remaining jobs in the queue. First, it checks if job J_k completes with error. If so, the job will be inserted back into the queue, based on its priority, to be rescheduled later. All jobs in the queue are divided into two groups: the first m jobs with the highest priorities are each given a reservation at the earliest possible time, provided that any reservation made should not delay the starting times of the higher-priority jobs; the subsequent jobs in the queue (if any) are then examined one by one and backfilled to start at the current time, if such backfilling does not affect any reservations for the higher-priority jobs.[2]

The R-List heuristic takes a parameter m, and depending on the value of m chosen, it resembles several different scheduling strategies known in theory and practice:

- $m = |Q|$ (Conservative backfilling [33]): this strategy makes reservations for all pending jobs in the queue;
- $m = 1$ (Aggressive or EASY backfilling [30, 39]): this strategy makes a reservation only for the job at the head of the queue, and uses backfilling to schedule all remaining jobs in the queue;
- $m = 0$ (Greedy scheduler [42, 16, 15]): this strategy does not make any reservation, and uses backfilling to schedule all jobs in the queue.

Note that, in the case of $m > 0$, and when a job J_k with high priority fails, it may be re-inserted back into the first part of the queue (i.e., among the top m jobs). This may require recomputing the existing reservations (made previously) for some jobs in the queue that have lower priority than J_k. From an analysis point of view, we can think of each job completion as a trigger, which deletes all previous reservations and makes a fresh round of reservation and backfilling decisions based on the updated queue.

In the following, we denote by Reservation this variant of R-List with reservations ($m > 0$), and by Greedy the variant with $m = 0$.

4.2 Approximation ratios of R-List

We show that, under any failure scenario, Reservation with a particular priority rule is a $(3 - \frac{4}{P+1})$-approximation, and that Greedy with any prior-

[2] For practical schedulers, this is typically implemented using two separate job queues, one for reservation and one for backfilling.

Algorithm 1: R-LIST

Input: a set $\mathcal{J} = \{J_1, J_2, \cdots, J_n\}$ of rigid jobs, with processor allocation p_j and error-free execution time t_j for each job $J_j \in \mathcal{J}$, a platform with P identical processors, parameter m;

Output: a list schedule with starting times for all jobs in $\mathcal{J}$ till they complete successfully, while satisfying Constraints (2) and (3).

```
begin
    Insert all jobs into a queue Q according to some priority rule;
    whenever an existing job J_k completes do
        if error detected for J_k then
            Q.insert_with_priority(J_k);
        end
        // schedule high-priority jobs using reservation
        for j = 1, 2 ..., min(m, |Q|) do
            J_j ← Q(j);
            Give job J_j an earliest possible reservation without delaying the
              reservation of job J_j', ∀j' = 1, ..., j − 1;
        end
        // schedule low-priority jobs using backfilling
        t ← get_current_time();
        for j = m + 1, ..., |Q| do
            J_j ← Q(j);
            if Job J_j can be scheduled at time t without delaying the reservation
              of job J_j', ∀j' = 1 ... m then
                execute job J_j at time t;
            end
        end
    end
end
```

ity rule is a $(2-\frac{1}{P})$-approximation. According to Equation (8), these results directly imply the same approximation ratios for the respective heuristic variants in terms of the expected makespan.

To assist the analysis, we first define some notations below. Since R-LIST only allocates and de-allocates processors upon job completions (the starting time of a reservation is necessarily at a future job completion time as well), the entire schedule can be divided into a set of consecutive and non-overlapping intervals $\mathcal{I} = \{I_1, I_2, \ldots, I_v\}$, where jobs only start (or complete) at the beginning (or end) of an interval, and v denotes the total number of intervals. Let $p(I_\ell)$ be the processor utilization (i.e., total number of allocated processors) during interval I_ℓ. As R-LIST never idles all processors unless all jobs complete successfully, we have $p(I_\ell) \geq 1$ for all $I_\ell \in \mathcal{I}$. Let $|I_\ell|$ denote the length of interval I_ℓ. The makespan of R-LIST under a particular failure $\mathbf{f}$ scenario can be expressed as $T(\mathbf{f}, \mathbf{s}) = \sum_{1 \leq \ell \leq v} |I_\ell|$.

4.2.1 Result for RESERVATION

We first consider the RESERVATION variant, and analyze its performance while applying a priority rule that favors large jobs and uses any priority for small jobs. We call this rule *Large Job First (LJF)*. Specifically, a job is said to be *large* if its processor allocation is at least $\frac{P+1}{2}$, and *small* otherwise. The LJF rule specifies that: (1) all large jobs have higher priority than all small jobs; (2) the priorities for large jobs are based on decreasing processor allocation; and (3) the priorities for small jobs are defined arbitrarily.

The following proposition shows the performance of RESERVATION in any failure scenario using the above LJF rule. The result matches the 3-approximation ratio [3, 42] known for failure-free jobs.

Proposition 1. *For any set of rigid parallel jobs under any failure scenario* $\mathbf{f}$*, the makespan of* RESERVATION *with the LJF priority rule satisfies:*

$$T_{\mathrm{R}}(\mathbf{f},\mathbf{s}) \leq (3 - \frac{4}{P+1}) \cdot T_{\mathrm{OPT}}(\mathbf{f},\mathbf{s}^*) \ . \tag{9}$$

Proof. Let J_j be a last successfully completed job in the schedule. We divide the set $\mathcal{I} = \{I_1, I_2, \ldots, I_v\}$ of all intervals into two disjoint subsets $\mathcal{I}_1$ and $\mathcal{I}_2$, where $\mathcal{I}_1$ contains the intervals in which job J_j is executing (including all of its execution attempts), and $\mathcal{I}_2 = \mathcal{I} \backslash \mathcal{I}_1$. Let $T_1 = \sum_{I \in \mathcal{I}_1} |I|$ and $T_2 = \sum_{I \in \mathcal{I}_2} |I|$ denote the total lengths of all intervals in $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively. Based on Equation (5), we have $T_1 = (f_j+1) \cdot t_j(p_j) \leq t_{\max}(\mathbf{f}) \leq T_{\mathrm{OPT}}(\mathbf{f},\mathbf{s}^*)$.

We will show that the processor utilization in any interval $I \in \mathcal{I}_2$ satisfies $p(I) \geq \frac{P+1}{2}$. First, we observe that all large jobs are completed sequentially (in decreasing order of processor allocation) at the beginning of the entire schedule, since no two large jobs can be scheduled at the same time, and no small (backfilling) jobs can delay their executions because large jobs have higher priority based on the LJF rule. Thus, if an interval $I \in \mathcal{I}_2$ contains a large job, its processor allocation must satisfy $p(I) \geq \frac{P+1}{2}$.

Now, consider any interval $I \in \mathcal{I}_2$ after all the large jobs have completed, and suppose I lies in between the i-th execution attempt and the $(i+1)$-th execution attempt of J_j, where $0 \leq i \leq f_j$. Hence, if such an interval exists, it means that J_j is a small job (with $p_j \leq \frac{P+1}{2}$), as well as all remaining jobs that are to be executed. Let t be the time at the beginning of this interval I. Recall that we can consider RESERVATION to make a fresh round of reservations and backfillings based on the current job queue Q at time t. Let J_k be the first job in Q that cannot be scheduled (either reserved or backfilled) to run at t. We know that such a job always exists because of the $(i+1)$-th execution attempt of J_j, which is scheduled to run at a later time. Let $\mathcal{J}_t$ be the set of jobs already running at time t or just scheduled to run at time t before job J_k, and let $p(\mathcal{J}_t)$ be the total processor allocation of all jobs in $\mathcal{J}_t$. As J_k cannot be scheduled to run at time t, it must be due to

$p(\mathcal{J}_t) + p_k \geq P + 1$. Since J_k is a small job, i.e., $p_k \leq \frac{P+1}{2}$, it implies that $p(I) \geq p(\mathcal{J}_t) \geq \frac{P+1}{2}$.

Thus, based on Equation (6) and since $p_j \geq 1$, we have $P \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \geq A(\mathbf{f}) \geq \frac{P+1}{2} \cdot T_2 + p_j \cdot T_1 \geq \frac{P+1}{2} \cdot T_2 + T_1$. The overall execution time of RESERVATION with the LJF priority rule therefore satisfies:

$$\begin{aligned} T_{\text{R}}(\mathbf{f}, \mathbf{s}) &= T_1 + T_2 \\ &\leq T_1 + 2 \cdot \frac{P \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) - T_1}{P+1} \\ &= \frac{2P}{P+1} \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) + \left(1 - \frac{2}{P+1}\right) \cdot T_1 \\ &\leq (3 - \frac{4}{P+1}) \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \ . \end{aligned}$$

□

4.2.2 Result for GREEDY

We now consider the GREEDY variant. The following proposition shows the performance of GREEDY in any failure scenario regardless of the priority rule. The result generalizes the same approximation ratio [42, 16, 15] of GREEDY for failure-free jobs.

Proposition 2. *For any set of rigid parallel jobs under any failure scenario* $\mathbf{f}$*, the makespan of* GREEDY *regardless of the priority rule satisfies:*

$$T_{\text{G}}(\mathbf{f}, \mathbf{s}) \leq (2 - \frac{1}{P}) \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \ . \tag{10}$$

Proof. Given the set $\mathcal{I} = \{I_1, I_2, \ldots, I_v\}$ of all intervals in the schedule, let $p_{\min} = \min_{1 \leq \ell \leq v} p(I_\ell)$ denote the minimum processor utilization among them. Since the algorithm never idles all processors unless all jobs complete successfully, we have $p_{\min} \geq 1$. We consider two cases:

Case 1: $p_{\min} \geq \frac{P+1}{2}$. In this case, we have $p(I_\ell) \geq p_{\min} \geq \frac{P+1}{2}$ for all $1 \leq \ell \leq v$. Hence, based on Equation (6), we get $P \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \geq A(\mathbf{f}) = \sum_{\ell=1,\ldots,v} |I_\ell| \cdot p(I_\ell) \geq \frac{P+1}{2} \cdot T_{\text{G}}(\mathbf{f}, \mathbf{s})$. This implies:

$$T_{\text{G}}(\mathbf{f}, \mathbf{s}) \leq \frac{2P}{P+1} \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \leq (2 - \frac{1}{P}) \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \ .$$

Case 2: $p_{\min} < \frac{P+1}{2}$. In this case, let $I_{\min}$ denote the last-executed interval that has processor utilization $p_{\min}$. Consider a job J_j that is running during interval $I_{\min}$. Necessarily, we have $p_j \leq p_{\min}$. We divide the set $\mathcal{I}$ of intervals into two disjoint subsets $\mathcal{I}_1$ and $\mathcal{I}_2$, where $\mathcal{I}_1$ contains the intervals in which job J_j is executing (including all of its execution attempts), and $\mathcal{I}_2 = \mathcal{I} \backslash \mathcal{I}_1$. Let $T_1 = \sum_{I \in \mathcal{I}_1} |I|$ and $T_2 = \sum_{I \in \mathcal{I}_2} |I|$ denote the total lengths of all intervals in $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively. Based on Equation (5), we have $T_1 = (f_j + 1) \cdot t_j(p_j) \leq t_{\max}(\mathbf{f}) \leq T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*)$.

For any interval $I \in \mathcal{I}_2$ that lies between the i-th execution attempt and the $(i+1)$-th execution attempt of J_j in the schedule, where $0 \leq i \leq f_j$, the processor utilization of I must satisfy $p(I) \geq P - p_{\min} + 1$, since otherwise there are at least $p_{\min} \geq p_j$ available processors during interval I and hence the $(i+1)$-th execution attempt of J_j would have been scheduled at the beginning of I.

For any interval $I \in \mathcal{I}_2$ that lies after the (f_j+1)-th (last) execution attempt of J_j, there must be a job J_k running during I and that was not running during $I_{\min}$ (meaning no attempt of executing J_k was made during $I_{\min}$). This is because $p(I) > p_{\min}$, hence the job configuration must differ between I and $I_{\min}$. The processor utilization during interval I must also satisfy $p(I) \geq P - p_{\min} + 1$, since otherwise the processor allocation of J_k will be $p_k \leq p(I) \leq P - p_{\min}$, implying that the first execution attempt of J_k after interval $I_{\min}$ would have been scheduled at the beginning of $I_{\min}$.

Thus, for all $I \in \mathcal{I}_2$, we have $p(I) \geq P - p_{\min} + 1$. Based on Equation (6), we have $P \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \geq A(\mathbf{f}) \geq (P - p_{\min} + 1) \cdot T_2 + p_{\min} \cdot T_1$. Since $p_{\min} \geq 1$, the overall execution time of GREEDY therefore satisfies:

$$\begin{aligned}
T_{\text{G}}(\mathbf{f}, \mathbf{s}) &= T_1 + T_2 \\
&\leq T_1 + \frac{P \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) - p_{\min} \cdot T_1}{P - p_{\min} + 1} \\
&= \frac{P}{P - p_{\min} + 1} \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) + \frac{P - 2p_{\min} + 1}{P - p_{\min} + 1} \cdot T_1 \\
&\leq \frac{2P - 2p_{\min} + 1}{P - p_{\min} + 1} \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \\
&\leq (2 - \frac{1}{P - p_{\min} + 1}) \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \\
&\leq (2 - \frac{1}{P}) \cdot T_{\text{OPT}}(\mathbf{f}, \mathbf{s}^*) \ . \qquad \square
\end{aligned}$$

4.3 R-SHELF scheduling heuristic

We now present a shelf-based scheduling heuristic, called R-SHELF, that schedules any set of parallel jobs onto a series of shelves while handling job failures.

Heuristic description. Algorithm 2 shows the pseudocode of R-SHELF. As in R-LIST, the algorithm starts by organizing all jobs in a queue based on some priority rule. Whenever the jobs in the preceding shelf all complete (at time 0, a virtual shelf S_0 with no job in it can be considered to complete), the algorithm builds a new shelf and adds the remaining jobs to it. First, any job in the preceding shelf that completes with error will be inserted back into the queue based on its priority. Then, the algorithm scans the queue and adds a job to the new shelf if the job can fit in without violating the

Algorithm 2: R-SHELF

```
Input: a set 𝒥 = {J_1, J_2, ··· , J_n} of rigid jobs, with processor allocation p_j and
       error-free execution time t_j for each job J_j ∈ 𝒥, a platform with P
       identical processors, parameter b;
Output: a shelf schedule with starting times for all jobs in 𝒥 till they complete
        successfully.
begin
    Insert all jobs into a queue Q according to some priority rule;
    i ← 0, S_i ← ∅;
    whenever all jobs in S_i complete do
        if error detected for J_k ∈ S_i then
            Q.insert_with_priority(J_k);
        end
        i ← i + 1 and S_i ← ∅; // start a new shelf
        for j = 1, 2 . . . , |Q| do
            J_j ← Q(j);
            if Job J_j can fit in shelf S_i then
                S_i ← S_i ⋃ {J_j};
            else if b = 0 then
                break ; // no backfilling
            end
        end
        execute all jobs in S_i at the current time;
    end
end
```

processor constraint. R-SHELF takes a binary parameter b that determines if backfilling is used in the process:

- $b = 0$ (No backfilling): the heuristic closes the new shelf upon encountering the first job in the queue that does not fit in the shelf. This resembles the Next-Fit (NF) strategy for bin-packing.
- $b = 1$ (Backfilling): the heuristic scans all the jobs in the queue until no more job can be added to the new shelf. This resembles the First-Fit (FF) strategy for bin-packing.

Once the jobs in the new shelf have been selected, they will simultaneously start their executions.

Inapproximability result. For failure-free jobs, the variant of R-SHELF without backfilling and considering jobs in the non-increasing execution time order is equivalent to the Next-Fit Decreasing Height (NFDH) [10] algorithm for strip packing. The algorithm starts with the longest job J_1, which is put on the first shelf, whose height is t_1. Then, the next job J_2 is put on the same shelf if it fits in, meaning that $p_1 + p_2 \leq P$, otherwise a new shelf is started for J_2, whose height is t_2. The algorithm proceeds like this, either putting the next job on the last shelf if it fits in, or creating a new shelf otherwise. Despite its simplicity, the algorithm is shown to be a 3-approximation for

failure-free jobs [10, 42].

Now, when jobs can fail, we show that there exists a job instance $\mathcal{J}$ and a failure scenario $\mathbf{f}$ such that any shelf-based algorithm has a makespan $T_{\mathrm{S}}(\mathbf{f}, \mathbf{s})$ that is arbitrarily higher than the optimal makespan $T_{\mathrm{OPT}}(\mathbf{f}, \mathbf{s}^*)$ regardless of the job priority used. This is in clear contrast with the 3-approximation result for the failure-free case. Note that $T_{\mathrm{OPT}}(\mathbf{f}, \mathbf{s}^*)$ is not necessarily the optimal makespan of a shelf-based schedule.

Proposition 3. *There exists a job instance and a failure scenario such that any shelf-based algorithm has an approximation ratio of* $\Omega(\ln P)$.

Proof. Consider a set $\mathcal{J} = \{J_1, \ldots, J_P\}$ of P uniprocessor jobs, where $t_j = P/j$ and $p_j = 1$ for $1 \leq j \leq P$. For the failure scenario $\mathbf{f}$, we let $f_j = j - 1$ for $1 \leq j \leq P$; hence job J_1 does not fail, job J_2 fails once before success, and job J_P fails $f_P = P - 1$ times before success.

We first consider the R-Shelf algorithm. Because the problem instance above has only P uniprocessor jobs, R-Shelf has no freedom at all: it schedules the first execution of all P jobs in the first shelf of height t_1, then the second execution of jobs J_2 to J_P in the second shelf of height t_2, and so on until the last shelf of height t_P, which includes only the P-th execution of job J_P. Therefore, the makespan of R-Shelf is $T_{\mathrm{S}}(\mathbf{f}, \mathbf{s}) = P + \frac{P}{2} + \cdots + 1 = P \sum_{j=1}^{P} \frac{1}{j}$, while the optimal algorithm schedules the different executions of all jobs right after each other, thus having a makespan of $T_{\mathrm{OPT}}(\mathbf{f}, \mathbf{s}^*) = P$. The ratio $\frac{T_{\mathrm{S}}(\mathbf{f},\mathbf{s})}{T_{\mathrm{OPT}}(\mathbf{f},\mathbf{s}^*)}$ tends to $\ln(P)$ when P tends to infinity, hence it is not bounded.

Furthermore, since the P jobs have decreasing execution time and increasing number of failures, any shelf-based algorithm will have at least one shelf of height t_j, for all $1 \leq j \leq P$, thus having a makespan that is at least $T_{\mathrm{S}}(\mathbf{f}, \mathbf{s})$. Therefore, the same ratio applies to any shelf-based algorithm. □

We conclude this section with an open problem. Instead of a single failure scenario, consider an Exponential probability distribution and the expected makespan as defined in Section 3.4. Will R-Shelf or any shelf-based algorithm admit a constant approximation ratio in expectation? To answer this question seems difficult, because computing the expected makespan seems out of reach analytically. Given $P = 10$ in the above example, we find numerically (using a computer program) that the expected makespan ratio of R-Shelf is 1.00005 for $\lambda = 10^{-7}$ and 1.07 for $\lambda = 10^{-3}$. We have not been able to build an example where this ratio (computed numerically) is greater than 3.

5 Performance Evaluation

We now evaluate and compare the performance of all heuristics presented in Section 4, using different job priority rules and backfilling strategies. The evaluation is performed by simulation using both synthetic jobs and jobs extracted from the log traces of the Mira supercomputer.

5.1 Simulation setup

We compare five different heuristics combined with seven different priority rules. The five heuristics are:

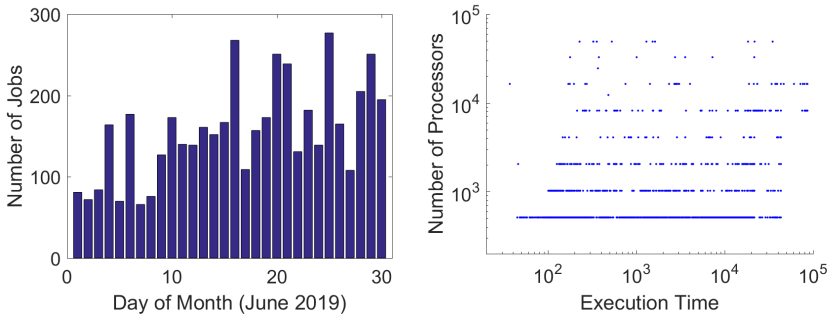
- R-List-0: The list-based algorithm with $m = 0$;
- R-List-1: The list-based algorithm with $m = 1$;
- R-List-Q: The list-based algorithm with $m = |Q|$;
- R-Shelf-B: The shelf-based algorithm with $b = 1$.
- R-Shelf-NB: The shelf-based algorithm with $b = 0$.

For each of these five heuristics, we consider seven different job priority rules:

- Lpt/Spt (Longest/Shortest Processing Time): a job with a longer/shorter processing time will have higher priority;
- Hpa/Lpa (Highest/Lowest Processor Allocation): a job with a higher/lower number of requested processors will have higher priority;
- La/Sa (Largest/Smallest Area): a job with a larger/smaller area will have higher priority;
- Random (Random): the priorities are determined randomly for all jobs.

We simulate two different settings, one using synthetic jobs and the other using real job traces from the Mira logs.

- *Synthetic jobs*: We generate 30 different job sets, each containing 100 jobs. For each job, the processor allocation is generated uniformly at random between 50 and 2000, while the execution time is generated uniformly at random between 100 and 20000 seconds. The total number of processors is set to be $P = 10000$. In the experiments, we also vary P to study its impact.
- *Jobs from Mira logs*: We generate jobs by extracting from the log traces [1] (of June 2019) of the Mira supercomputer, which has $P = 49152$ compute nodes. There were 4699 jobs submitted in June 2019, and we group the ones submitted each day as a set to form 30 sets of jobs. Figure 1(a) plots the number of jobs in each day of the month, varying between 66 and 277. The processor allocations of the jobs vary between 512 and 49152, and the execution times vary between 37 and 86494 seconds. Figure 1(b) plots these two parameters for all jobs in the month (with each point representing a job).



(a) (b)

Figure 1: Data from the trace logs of the Mira supercomputer.

In both settings, silent errors are injected to the jobs based on the Exponential distribution as described in Section 3.4. To study the impact of error rate, we further define the average failure probability for a set of jobs to be $\bar{q} = 1-e^{-\lambda\bar{a}}$, where $\bar{a} = \sum_{j=1}^{n} a_j/n$ is the average area of all jobs in the set. Intuitively, $\bar{q}$ represents the probability that a job with the average area over all jobs would fail due to silent errors. For a given value of $\bar{q}$, we can compute the error rate as $\lambda = -\ln(1-\bar{q})/\bar{a}$, and hence the failure probability of any job J_j with area a_j to be $q_j = 1-e^{-\lambda a_j} = 1-(1-\bar{q})^{a_j/\bar{a}}$. Based on this $\bar{q}$, we then randomly generate 1000 failure scenarios for the set of jobs following the probabilities. For each failure scenario $\mathbf{f}$, we evaluate the makespans of the heuristics, normalized by the lower bound $L(\mathbf{f}) = \max(t_{\max}(\mathbf{f}), A(\mathbf{f})/P)$ as defined in Equations (5) and (6). The normalized makespans are then averaged over the 1000 failure scenarios for comparison.

The simulation code for all experiments is publicly available at `http://www.github.com/vlefevre/job-scheduling`.

5.2 Results for synthetic jobs

We first compare the performance of different heuristics using synthetic jobs. Here, we focus on assessing the impact of two parameters: the average failure probability $\bar{q}$, and the total number of processors P. The results are averaged over the 30 different sets of jobs.

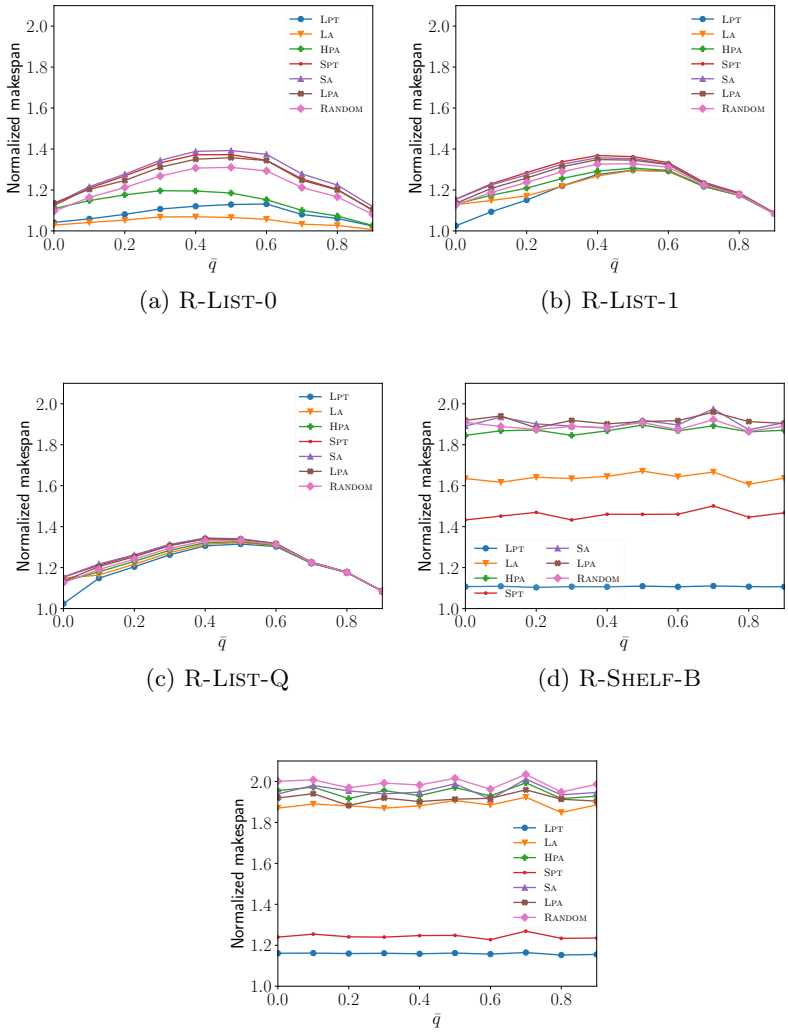
Figure 2 shows the performance of different heuristics when $\bar{q}$ varies from 0 to 0.9. First, we can see that, for all list-based heuristics, the normalized makespans first increase with $\bar{q}$ and then decrease. Indeed, a higher failure probability will result in a larger number of errors, thus increasing the difficulty of scheduling and hence the makespan. However, when the probability is too high, an excessive number of errors will occur, making the optimal scheduler perform equally worse and thus reducing the makespan ratios for the heuristics. For the shelf-based heuristics, the performance appears to be independent of the failure probability. Here, tasks that fail need to wait for the completion of the current shelf to be re-executed, so the num-

ber of shelves is mainly determined by the number of re-executions, which influences both the makespan and an optimal scheduler. The normalized makespan is thus mainly decided by the efficiency of the heuristic to fill one shelf, which does not depend on the failure probabilities. Second, the LPT and LA priorities lead to the best performance for all list-based heuristics, with LPT performing better when $\bar{q}$ is low for R-LIST-1 and R-LIST-Q, and LA performing better for R-LIST-0 under any $\bar{q}$. For the shelf-based heuristics, LPT and SPT are the two best priorities, which is not surprising as the performance of these algorithms is mainly determined by the duration of each shelf.

Figure 4(a) further compares the performance of the five heuristics using some of the best priorities. While most list-based heuristics behave similarly when there is no failure (i.e., $\bar{q} = 0$), R-LIST-0 clearly outperforms the rest when jobs can fail. This corroborates the theoretical result that R-LIST-0 (i.e., GREEDY) has the lowest approximation ratio regardless of the priority rule and failure scenario. Moreover, R-LIST-0 is also the heuristic that is least affected by job failures, with an increase in normalized makespan by less than 10% compared to the case of $\bar{q} = 0$, while the other heuristics experience 20-30% increase in normalized makespan. Finally, R-SHELF-NB appears to be the worst heuristic for small and high probabilities of failure with a makespan that is up to 15% higher than that of R-LIST-0 (when $\bar{q} = 0.9$), while R-LIST-Q is the worst for medium probabilities (e.g., 26% higher than that of R-LIST-0 for $\bar{q} = 0.5$). The results are likely due to: (i) the restriction of R-SHELF-NB for building shelves in a schedule, which leads to poor performance for some failure scenarios (such as the one discussed in Section 4.3), and hence an increase in the expected makespan, and (ii) the fact that R-LIST-Q is more affected by the increasing failure probability.

Figure 3 shows the performance of different heuristics when the number of processors P varies from 5000 to 20000 while the failure probability is fixed at $\bar{q} = 0.3$. Again, we can see that LA and LPT are the two best priority rules for all heuristics, with LA performing better for R-LIST-0 and R-LIST-1, and LPT performing better for other heuristics under all P. Also, the normalized makespans of the heuristics first increase with the number of processors and then tend to decrease. This is because when P is either too small (i.e., total resource is constrained) or too big (i.e., total resource is almost unconstrained), the optimal scheduler tends to have very similar performance as the heuristics.

We further compare the performance of the five heuristics using some of the best priorities in Figure 4(b). As in the previous experiment, the best heuristic is R-LIST-0 with the LA priority, which is the least impacted by the total number of processors (with $< 10\%$ variations in normalized makespan). Also, R-LIST-Q gives the worst performance (with a 23% increase in makespan compared to R-LIST-0 with LA when $P = 15000$) and



(a) R-LIST-0

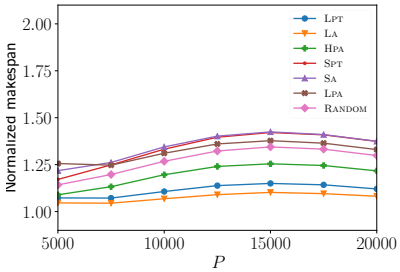
(b) R-LIST-1

(c) R-LIST-Q

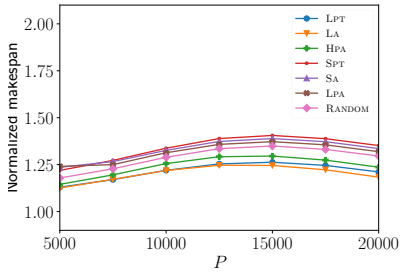
(d) R-SHELF-B

(e) R-SHELF-NB

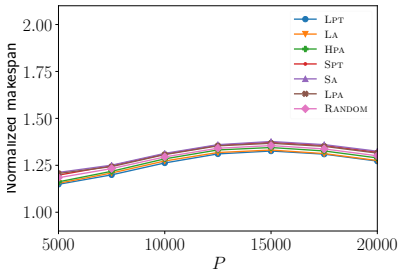
Figure 2: Normalized makespans of different heuristics and priority rules over 30 sets of jobs when $\bar{q}$ varies between 0 and 0.9, and $P = 10000$.



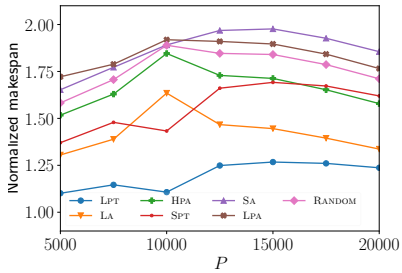
(a) R-LIST-0



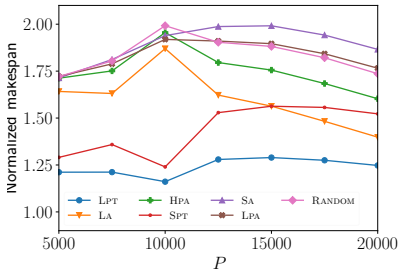
(b) R-LIST-1



(c) R-LIST-Q

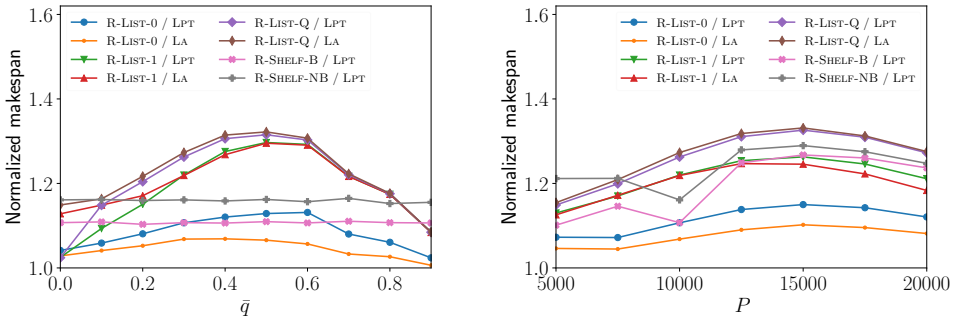


(d) R-SHELF-B



(e) R-SHELF-NB

Figure 3: Normalized makespans of different heuristics and priority rules over 30 sets of jobs when P varies between 5000 and 20000, and $\bar{q} = 0.3$.



(a) (b)

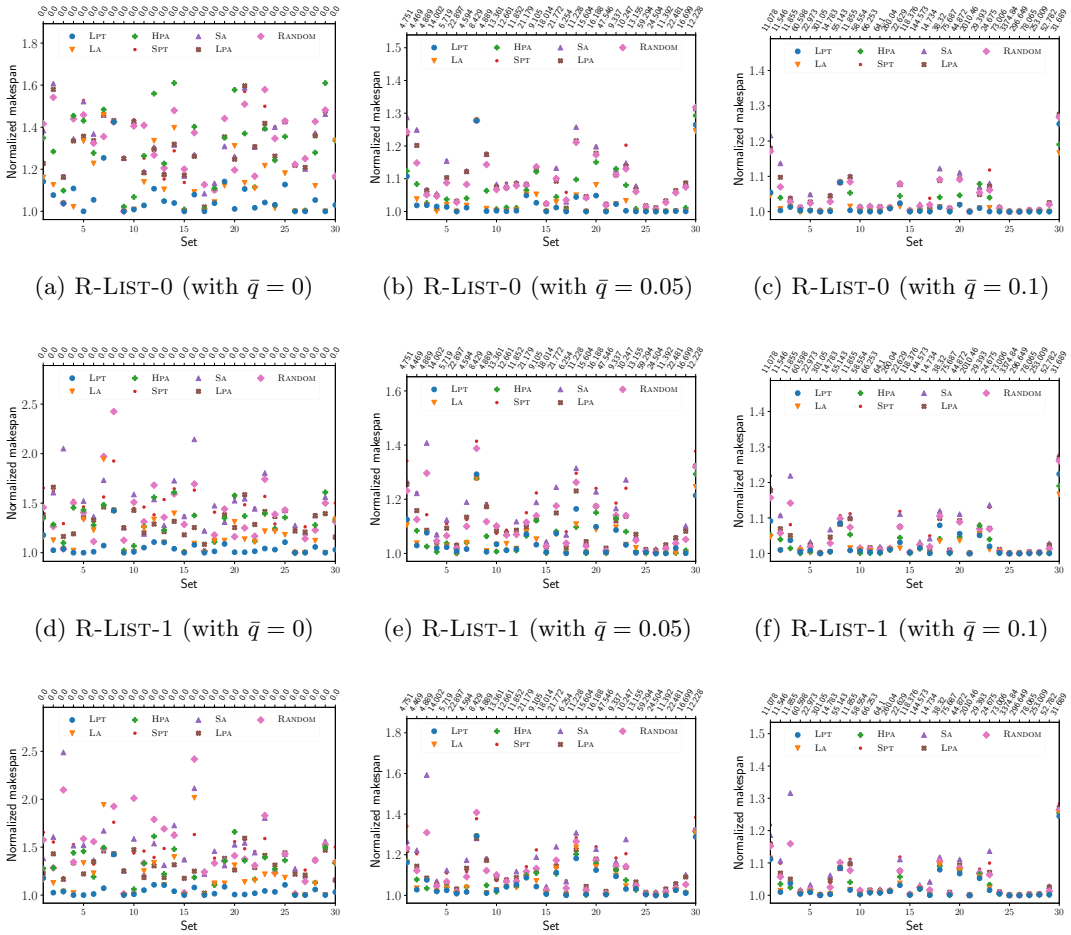
Figure 4: Comparison of different heuristics with the best priority rules Lpt and/or La when: (a) $\bar{q}$ varies between 0 and 0.9, and $P = 10000$; and (b) P varies between 5000 and 20000, and $\bar{q} = 0.3$.

has the largest variation (∼20%) in normalized makespan as the number of processors changes.

From these experiments, we can see that job failures and processor variations do have an impact on the relative performance of different heuristics. Nevertheless, the makespans of all the heuristics (with good priorities) are never more than 40% worse than the theoretical lower bound, which can be much less than the optimal makespan. The results suggest the robustness of these heuristics, and that they should actually perform really well in practice, even with job failures.

5.3 Results for jobs from Mira

We now evaluate the performance of different heuristics using real jobs from the Mira trace logs. Figures 5 and 6 show the normalized makespans of all heuristics and priority rules under all 30 days (sets) of jobs with and without failures. We observe that the Lpt and La priorities again offer the best performance, with Lpt performing better this time for most job sets. This holds for every heuristic on average, especially when there is no failure (i.e., $\bar{q} = 0$). As the failure probability increases, both Lpt and La (and even Hpa) give similar performance. The reason is that the processor allocations and execution times of the jobs in Mira are more skewed than those of the synthetic ones. Here, some jobs use a very large number of processors and have long execution times, which make them fail more often even with small values of $\bar{q}$. As a result, the makespan lower bound is largely determined by the total execution times of these jobs, thus any priority rule that favors these jobs will achieve similar performance. Comparing different heuristics, we can see that R-List-0 again performs the best and R-Shelf-B the worse, especially with higher failure probability ($\bar{q} = 0.1$). This is consistent with the previous findings and corroborates the analysis.

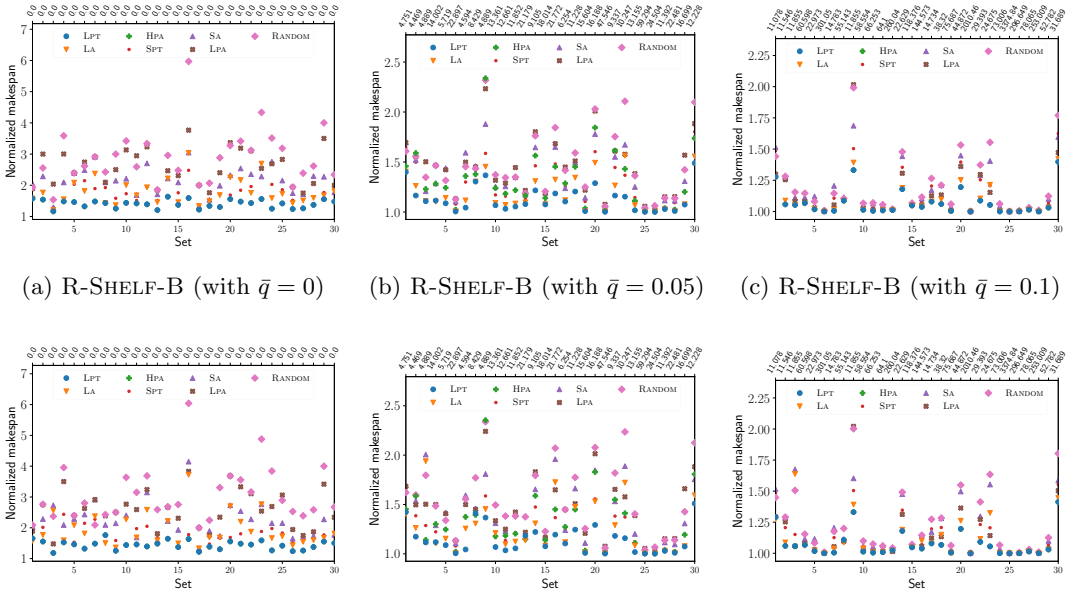


(a) R-LIST-0 (with $\bar{q} = 0$) (b) R-LIST-0 (with $\bar{q} = 0.05$) (c) R-LIST-0 (with $\bar{q} = 0.1$)

(d) R-LIST-1 (with $\bar{q} = 0$) (e) R-LIST-1 (with $\bar{q} = 0.05$) (f) R-LIST-1 (with $\bar{q} = 0.1$)

(g) R-LIST-Q (with $\bar{q} = 0$) (h) R-LIST-Q (with $\bar{q} = 0.05$) (i) R-LIST-Q (with $\bar{q} = 0.1$)

Figure 5: Performance of list-based heuristics for 30 job sets using the Mira trace logs (June 2019) with and without failures. Each row represents a different heuristic (R-LIST-0, R-LIST-1 and R-LIST-Q), and each column represents a different failure probability ($\bar{q} = 0$, $\bar{q} = 0.05$ and $\bar{q} = 0.1$). The average number of failures for each job set is indicated on top of each plot.



(a) R-SHELF-B (with $\bar{q} = 0$) (b) R-SHELF-B (with $\bar{q} = 0.05$) (c) R-SHELF-B (with $\bar{q} = 0.1$)

(d) R-SHELF-NB (with $\bar{q} = 0$) (e) R-SHELF-NB (with $\bar{q} = 0.05$) (f) R-SHELF-NB (with $\bar{q} = 0.1$)

Figure 6: Performance of shelf-based heuristics for 30 job sets using the Mira trace logs (June 2019) with and without failures. Each row represents a different heuristic (R-SHELF-B and R-SHELF-NB), and each column represents a different failure probability ($\bar{q} = 0$, $\bar{q} = 0.05$ and $\bar{q} = 0.1$). The average number of failures for each job set is indicated on top of each plot.

Table 1: Performance of different heuristics using LPT priority for all 30 days (sets) of jobs from June 2019 on the Mira supercomputer.

$\bar{q}$	Average #failures	Average makespan ratio					Standard deviation					Maximum makespan ratio				
		R-LIST			R-SHELF		R-LIST			R-SHELF		R-LIST			R-SHELF	
		0	1	Q	B	NB	0	1	Q	B	NB	0	1	Q	B	NB
0	0	1.067	1.051	1.051	1.407	1.441	8.78×10^{-2}	8.19×10^{-2}	8.23×10^{-2}	1.29×10^{-1}	1.45×10^{-1}	1.425	1.425	1.425	1.633	1.760
0.05	15.2913	1.031	1.049	1.061	1.129	1.141	6.72×10^{-2}	6.87×10^{-2}	7.76×10^{-2}	1.30×10^{-1}	1.40×10^{-1}	1.278	1.292	1.292	1.489	1.510
0.1	254.453	1.016	1.025	1.028	1.071	1.073	4.66×10^{-2}	4.54×10^{-2}	4.97×10^{-2}	1.03×10^{-1}	1.06×10^{-1}	1.249	1.224	1.245	1.398	1.413

Table 1 summarizes the results of the five heuristics using the LPT priority (which is overall the best one) over 30 days (sets) of jobs, which have an average of 157.63 jobs per day (set). As $\bar{q}$ increases to 0.05 and 0.1, the average number of failures rises to around 15 and 254, respectively. All list-based heuristics have good average makespan ratios that are very close to 1 (with low standard deviations), as well as good maximum makespan ratios that are lower than 1.5, while the two shelf-based heuristics have worse performance in comparison, even when failures are not present. The maximum makespans, however, are never more than 80% of the theoretical lower bound. This again corroborates the results in Section 5.2.

Overall, these results confirm the efficacy and robustness of the resilient scheduling heuristics, not only for synthetic jobs, but also for real sets of jobs. In particular, both theory and practice have suggested that R-LIST-0 is the best heuristic when silent errors are present, and LPT and LA are the two best priorities for most cases. In all experiments we have conducted, this heuristic achieves a makespan that is within a few percent of the lower bound on average, and never more than 50% in the worst case.

6 Conclusion

In this paper, we have investigated the problem of scheduling rigid jobs onto a parallel platform subject to silent errors. We have revisited the classical scheduling algorithms in this new framework, where jobs that have been struck by errors must be re-executed (possibly many times) until success. We designed resilient list-based and shelf-based scheduling heuristics, along with different priority rules and backfilling strategies. On the theoretical side, we proved that variants of the list-based heuristic achieve a constant approximation ratio (2 or 3 depending whether reservation is used or not). We also showed that any shelf-based heuristic is no longer a constant-factor approximation, while a failure-free variant was known to be a 3-approximation. Extensive simulations conducted using both synthetic jobs and real traces from the Mira supercomputer demonstrate that these heuristics are quite robust, and achieve makespans close to the optimal. As highlighted by the theoretical analysis, the best strategy remains the unrestricted greedy list-based scheduling with no reservations, and good results are obtained in practice when job priorities are assigned by processing times (favoring jobs

with long execution times) or by areas (favoring jobs with many processors and/or long execution times).

Some problems remain open, in particular for the study of shelf-based algorithms, whose expected makespan under the Exponential probability distribution is not known to be bounded by a constant factor of the optimal or not. A natural extension of this work would be to consider moldable jobs, whose processor allocations can be decided at launch time. However, for jobs with nonlinear speedup curves, changing the number of processors assigned to a job also changes its error probability under the Exponential probability distribution, thereby severely complicating the problem, and thus calling for the design of novel heuristics.

Acknowledgement: The data from Mira logs was generated from resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

References

[1] Argonne Leadership Computing Facility. Mira log traces. https://reports.alcf.anl.gov/data/mira.html.

[2] B. Baker and J. Schwarz. Shelf algorithms for two-dimensional packing problems. *SIAM Journal on Computing*, 12(3):508–525, 1983.

[3] B. S. Baker, E. G. Coffman, and R. L. Rivest. Orthogonal packings in two dimensions. *SIAM Journal on Computing*, 9(4):846–855, 1980.

[4] L. Bautista Gomez and F. Cappello. Detecting silent data corruption through data dynamic monitoring for scientific applications. In *PPoPP*, 2014.

[5] J. Bruno, P. Downey, and G. N. Frederickson. Sequencing tasks with exponential service times to minimize the expected flow time or makespan. *J. ACM*, 28(1):100–113, 1981.

[6] K. M. Chandy and P. F. Reynolds. Scheduling partially ordered tasks with probabilistic execution times. *SIGOPS Oper. Syst. Rev.*, 9(5):169–177, 1975.

[7] B. Chen and A. P. Vestjens. Scheduling on identical machines: How good is LPT in an on-line setting. *Operations Research Letters*, 21(4):165–169, 1997.

[8] C. Chen, G. Eisenhauer, M. Wolf, and S. Pande. LADR: Low-cost application-level detector for reducing silent output corruptions. In *HPDC*, pages 156–167, 2018.

[9] Z. Chen. Online-ABFT: An online algorithm based fault tolerance scheme for soft error detection in iterative methods. *SIGPLAN Not.*, 48(8):167–176, 2013.

[10] E. G. Coffman, M. R. Garey, D. S. Johnson, and R. E. Tarjan. Performance bounds for level-oriented two-dimensional packing algorithms. *SIAM J. Comput.*, 9(4):808–826, 1980.

[11] J. Csirik and G. J. Woeginger. Shelf algorithms for on-line strip packing. *Information Processing Letters*, 63(4):171–175, 1997.

[12] J. Csirik and G. J. Woeginger. On-line packing and covering problems. In A. Fiat and G. J. Woeginger, editors, *Online Algorithms: The State of the Art*, chapter 7, pages 147–177. Springer, 1998.

[13] D. G. Feitelson, L. Rudolph, U. Schwiegelshohn, K. C. Sevcik, and P. Wong. Theory and practice in parallel job scheduling. In *JSSPP*, pages 1–34, 1997.

[14] A. Feldmann, M.-Y. Kao, J. Sgall, and S.-H. Teng. Optimal on-line scheduling of parallel jobs with dependencies. *Journal of Combinatorial Optimization*, 1(4):393–411, 1998.

[15] A. Feldmann, J. Sgall, and S.-H. Teng. Dynamic scheduling on parallel machines. *Theoretical Computer Science*, 130(1):49–72, 1994.

[16] M. R. Garey and R. L. Graham. Bounds for multiprocessor scheduling with resource constraints. *SIAM J. Comput.*, 4(2):187–200, 1975.

[17] M. R. Garey and D. S. Johnson. *Computers and Intractability, a Guide to the Theory of NP-Completeness.* W.H. Freeman and Company, 1979.

[18] E. Gaussier, J. Lelong, V. Reis, and D. Trystram. Online tuning of EASY-backfilling using queue reordering policies. *IEEE Transactions on Parallel and Distributed Systems*, 29(10):2304–2316, 2018.

[19] A. Goel and P. Indyk. Stochastic load balancing and related problems. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*, 1999.

[20] P.-L. Guhur, H. Zhang, T. Peterka, E. Constantinescu, and F. Cappello. Lightweight and accurate silent data corruption detection in ordinary differential equation solvers. In *Euro-Par*, 2016.

[21] X. Han, K. Iwama, D. Ye, and G. Zhang. Strip packing vs. bin packing. In M.-Y. Kao and X.-Y. Li, editors, *Algorithmic Aspects in Information and Management*, pages 358–367. Springer, 2007.

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

[23] K.-H. Huang and J. A. Abraham. Algorithm-based fault tolerance for matrix operations. *IEEE Trans. Comput.*, 33(6):518–528, 1984.

[24] J. L. Hurink and J. J. Paulus. Online algorithm for parallel job scheduling and strip packing. In C. Kaklamanis and M. Skutella, editors, *Approximation and Online Algorithms*, pages 67–74. Springer, 2008.

[25] D. B. Jackson, Q. Snell, and M. J. Clement. Core Algorithms of the Maui Scheduler. In *JSSPP*, pages 87–102, 2001.

[26] K. Jansen. A $(3/2+\epsilon)$ approximation algorithm for scheduling moldable and non-moldable parallel tasks. In *SPAA*, pages 224–235, 2012.

[27] B. Johannes. Scheduling parallel jobs to minimize the makespan. *J. of Scheduling*, 9(5):433–452, 2006.

[28] J. Kleinberg, Y. Rabani, and E. Tardos. Allocating bandwidth for bursty connections. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing (STOC)*, pages 664–673, 1997.

[29] K. Li. Analysis of the list scheduling algorithm for precedence constrained parallel tasks. *Journal of Combinatorial Optimization*, 3(1):73–88, 1999.

[30] D. A. Lifka. The ANL/IBM SP Scheduling System. In *JSSPP*, pages 295–303, 1995.

[31] A. Lodi, S. Martello, and M. Monaci. Two-dimensional packing problems: A survey. *European Journal of Operational Research*, 141(2):241–252, 2002.

[32] Marc Snir et al. Addressing failures in exascale computing. *Int. J. High Perform. Comput. Appl.*, 28(2):129–173, 2014.

[33] A. W. Mu'alem and D. G. Feitelson. Utilization, Predictability, Workloads, and User Runtime Estimates in Scheduling the IBM SP2 with Backfilling. *IEEE Trans. Parallel Distrib. Syst.*, 12(6):529–543, 2001.

[34] E. Naroska and U. Schwiegelshohn. On an on-line scheduling problem for parallel jobs. *Inf. Process. Lett.*, 81(6):297–304, 2002.

[35] J. Niño-Mora. Stochastic scheduling. *Encyclopedia of Optimization*, pages 3818–3824, 2009.

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

[37] M. L. Pinedo. *Scheduling: Theory, Algorithms, and Systems.* Springer-Verlag New York, Inc., Third edition, 2008.

[38] D. B. Shmoys, J. Wein, and D. P. Williamson. Scheduling parallel machines on-line. *SIAM J. Comput.*, 24(6):1313–1331, 1995.

[39] J. Skovira, W. Chan, H. Zhou, and D. A. Lifka. The EASY - LoadLeveler API Project. In *JSSPP*, pages 41–47, 1996.

[40] S. Srinivasan, R. Kettimuthu, V. Subramani, and P. Sadayappan. Characterization of backfilling strategies for parallel job scheduling. In *International Conference on Parallel Processing Workshop*, 2002.

[41] G. Staples. TORQUE resource manager. In *Proceedings of the ACM/IEEE Conference on Supercomputing*, 2006.

[42] J. Turek, J. L. Wolf, and P. S. Yu. Approximate algorithms scheduling parallelizable tasks. In *SPAA*, 1992.

[43] R. R. Weber. Scheduling jobs with stochastic processing requirements on parallel machines to minimize makespan or flowtime. *J Appl Probab*, 19(1):167–182, 1982.

[44] G. Weiss and P. M. Scheduling tasks with exponential service times on non-identical processors to minimize various cost functions. *J Appl Probab*, 17(1):187–202, 1980.

[45] A. K. L. Wong and A. M. Goscinski. Evaluating the EASY-backfill job scheduling of static workloads on clusters. In *CLUSTER*, 2007.

[46] P. Wu, C. Ding, L. Chen, F. Gao, T. Davies, C. Karlsson, and Z. Chen. Fault tolerant matrix-matrix multiplication: Correcting soft errors online. In *ScalA'11*, pages 25–28, 2011.

[47] D. Ye, X. Han, and G. Zhang. A note on online strip packing. *Journal of Combinatorial Optimization*, 17(4):417–423, 2009.

[48] A. B. Yoo, M. A. Jette, and M. Grondona. SLURM: Simple Linux Utility for Resource Management. In *JSSPP*, pages 44–60, 2003.

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

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