Do moldable applications perform better on failure-prone HPC platforms?

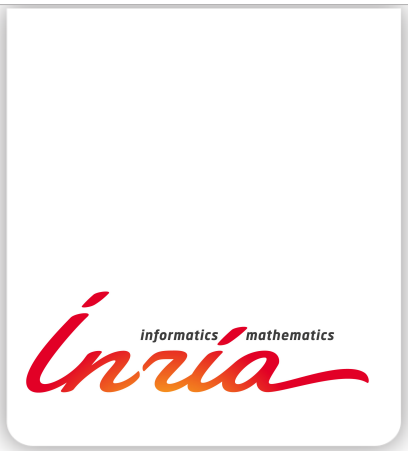
Valentin Le Fèvre, George Bosilca, Aurelien Bouteiller, Thomas Herault, Atsushi Hori, Yves Robert, Jack J. Dongarra

Do moldable applications perform better on failure-prone HPC platforms?

**Valentin Le Fèvre, George Bosilca, Aurelien Bouteiller,
Thomas Herault, Atsushi Hori,
Yves Robert, Jack Dongarra**

RESEARCH REPORT
N° 9174
May 2018
Project-Team ROMA

ISSN 0249-6399 ISRN INRIA/RR--9174--FR+ENG



Do moldable applications perform better on failure-prone HPC platforms?

Valentin Le Fèvre*, George Bosilca†, Aurelien Bouteiller†,
Thomas Herault†, Atsushi Hori‡,
Yves Robert*†, Jack Dongarra†§

Project-Team ROMA

Research Report n° 9174 — May 2018 — 21 pages

Abstract: This paper compares the performance of different approaches to tolerate failures using checkpoint/restart when executed on large-scale failure-prone platforms. We study (i) RIGID applications, which use a constant number of processors throughout execution; (ii) MOLDABLE applications, which can use a different number of processors after each restart following a fail-stop error; and (iii) GRIDSHAPED applications, which are moldable applications restricted to use rectangular processor grids (such as many dense linear algebra kernels). For each application type, we compute the optimal number of failures to tolerate before relinquishing the current allocation and waiting until a new resource can be allocated, and we determine the optimal yield that can be achieved. We instantiate our performance model with realistic applicative scenarios and make it publicly available for further usage.

Key-words: checkpoint, fail-stop error, rigid application, moldable application, grid-shaped application, number of spares, number of tolerated failures, optimal allocation length.

* LIP, École Normale Supérieure de Lyon, CNRS & Inria, France
† University Tennessee Knoxville, USA
‡ RIKEN Center for Computational Science, Japan
§ University of Manchester, UK

RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

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

Performance comparée des applications rigides et élastiques sur des plates-formes de calcul scientifique sujettes aux fautes

Résumé : Ce rapport compare l'efficacité de plusieurs approches pour tolérer un certain nombre d'erreurs fatales et continuer l'exécution en utilisant des techniques de checkpoint et redémarrage sur des plates-formes à grande échelle. Nous étudions trois types d'applications: (i) les applications rigides, qui utilisent toujours le même nombre de processeurs durant toute l'exécution; (ii) les applications élastiques, qui peuvent utiliser un nombre différent de processeurs après chaque redémarrage dû à une erreur fatale; et (iii) les applications devant s'exécuter sur une grille de processeur (telles les noyaux d'algèbre linéaire dense). Pour chaque type, nous calculons le nombre optimal d'erreurs fatales à tolérer avant d'interrompre l'allocation courante et de se place en file d'attente de nouvelles ressources. Nous déterminons aussi le rendement optimal qui peut être atteint. Nous mettons en oeuvre notre modèle de performance avec des scénarios réalistes inspirés des plates-formes actuelles, et nous le mettons à libre disposition pour permettre à chacun d'explorer avec les paramètres de son choix.

Mots-clés : checkpoint, erreur fatale, application rigide, application élastique, application devant s'exécuter sur une grille de processeurs, nombre de processeurs mis en réserve, nombre d'erreurs tolérées, longueur optimale d'une allocation.

1 Introduction

Consider a long-running job that requests N processors from the batch scheduler. Resilience to fail-stop errors[1] is provided by a Checkpoint/Restart (CR) mechanism, which is the de-facto standard approach for High-Performance Computing (HPC) applications. After each failure, the application restarts from the last checkpoint but the number of available processors decreases, assuming the application can continue execution after a failure (e.g., using ULFM [3]). Until which point should the execution proceed before requesting a new allocation with N fresh resources from the batch scheduler?

The answer depends upon the nature of the application. For a RIGID application, the number of processors must remain constant throughout the execution. The question is then to decide the number F of processors (out of the N available initially) that will be used as spares. With F spares, the application can tolerate F failures. The application always executes with $N-F$ processors: after each failure, then it restarts from the last checkpoint and continues executing with $N-F$ processors, the faulty processor having been replaced by a spare. After F failures, the application stops when the $(F+1)$st failure strikes, and relinquishes the current allocation. It then asks for a new allocation with N processors, which takes a *wait time*, D, to start (as other applications are most likely using the platform concurrently). The optimal value of F obviously depends on the value of D, in addition to the application and resilience parameters. The wait time typically ranges from several hours to several days if the platform is over-subscribed (up to 10 days for large applications on the K-computer [21]). The metric to optimize here is the (expected) application yield, which is the fraction of useful work per second, averaged over the N resources, and computed in steady-state mode (expected value for multiple batch allocations of N resources).

For a MOLDABLE application, the problem is different: here we assume that the application can use a different number of processors after each restart. The application starts executing with N processors; after the first failure, the application recovers from the last checkpoint and is able to continue with only $N-1$ processors, albeit with a slowdown factor $\frac{N-1}{N}$. After how many failures F should the application decide to stop[2] and accept to produce no progress during D, in order to request a new allocation? Again, the metric to optimize is the application yield.

Finally, consider an application which must have a given shape (or a set of given shapes) in terms of processor layout. Typically, these shapes are dictated by the algorithm. In this paper, we use the example of a

[1] We use the terms *fail-stop error* and *failure* indifferently.

[2] Another limit is induced by the total application memory Mem_{tot}. There must remain at least ℓ live processors such that $Mem_{tot} \leq \ell \times Mem_{ind}$, where Mem_{ind} is the memory of each processor. We ignore this contraint in the paper but it would be straightforward to take it into account.

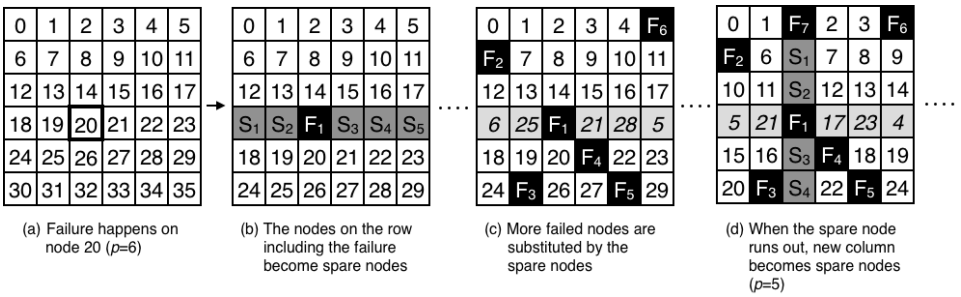


Figure 1: Example of node failures substituted by spare nodes in a 2-D GridShaped application.

GridShaped application, which is required to execute on a rectangular processor grid whose size can dynamically be chosen. Most dense linear algebra kernels (matrix multiplication, LU, Cholesky and QR factorizations) are GridShaped applications, and perform more efficiently on square processor grids than on elongated rectangle ones. The application starts with a square $p \times p$ grid of $N = p^2$ processors. After the first failure, execution continues on a $p \times (p-1)$ rectangular grid, keeping $p-1$ processors as spares for the next $p-1$ failures. After p failures, the grid is shrunk again to a $(p-1) \times (p-1)$ square grid, and so on. We address the same question: after how many failures F should the application stop working on a smaller processor grid and request a new allocation, in order to optimize the application yield?

The major contribution of this paper is to present a detailed performance model and to provide analytical formulas for the expected yield of each application type. We instantiate the model for several applicative scenarios, for which we draw comparisons across application types. Our model is publicly available [18] so that more scenarios can be explored. Notably, the paper qualifies the optimal number of spares for the optimal yield, and the optimal length of a period between two full restarts; it also qualifies how much the yield and total work done within a period are improved by deploying Moldable applications w.r.t. Rigid applications.

The rest of the paper is organized as follows. Section 2 provides an overview of related work. Section 3 is devoted to formally defining the performance model. Section 4 provides formulas for the yield of Rigid, Moldable and GridShaped applications. These formulas are instantiated through the applicative scenarios in Section 5, to compare the different results. Finally, Section 6 provides final remarks and hints for future work.

2 Related work

We first survey related work on checkpoint-restart in Section 2.1. Then we discuss previous contributions on MOLDABLE applications in Section 2.2.

2.1 Checkpoint-restart

Checkpoint/restart (CR) is the most common strategy employed to protect applications from underlying faults and failures on HPC platforms. Generally, CR periodically outputs snapshots (*i.e.*, checkpoints) of the application global, distributed state to some stable storage device. When a failure occurs, the last stored checkpoint is retrieved and used to restart the application.

A widely-used approach for HPC applications is to use a fixed checkpoint period (typically one or a few hours), but it is sub-optimal. Instead, application-specific metrics can (and should) be used to determine the optimal checkpoint period. The well-known Young/Daly formula [22, 7] yields an application optimal checkpoint period, $\sqrt{2\mu C}$ seconds, where C is the time to commit a checkpoint and μ the application Mean Time Between Failures (MTBF) on the platform. We have $\mu = \frac{\mu_{ind}}{N}$, where N is the number of processors enrolled by the application and μ_{ind} is the MTBF of an individual processor [15].

The Young/Daly formula minimizes platform waste, defined as the fraction of job execution time that does not contribute to its progress. The two sources of waste are the time spent taking checkpoints (which motivates longer checkpoint periods) and the time needed to recover and re-execute after each failure (which motivates shorter checkpoint periods). The Young/Daly period achieves the optimal trade-off between these sources to minimize the total waste.

2.2 Moldable and GridShaped applications

RIGID and MOLDABLE applications have been studied for long in the context of scientific applications. A detailed survey on various application types (RIGID, MOLDABLE, malleable) was conducted in [9]. Resizing application to improve performance has been investigated by many authors, including [16, 5, 20, 19] among others. A related recent study is the design of a MPI prototype for enabling tolerance in MOLDABLE MapReduce applications [11].

The TORQUE/Maui scheduler has been extended to support evolving, malleable, and MOLDABLE parallel jobs [17]. In addition, the scheduler may have system-wide spare nodes to replace failed nodes. In contrast, our scheme does not assume a change of behavior from the batch schedulers and resource allocators, but utilizes job-wide spare nodes: a node set including

potential spare nodes is allocated and dedicated to a job at the time of scheduling, that can be used by the application to restart within the same job after a failure.

An experimental validation of the feasibility of shrinking application on the fly is provided in [2]. In this paper, the authors used an iterative solver application to compared two recovery strategies, shrinking and spare node substitution. They use ULFM, the fault-tolerant extension of MPI that offers the possibiliity of dynamically resizing the execution after a failure. Finally, in [10, 13], the authors studied MOLDABLE and GRIDSHAPED applications that continue executing after some failures. They focus on the performance degradation incurred after shrinking or spare node substitution, due to less efficient communications (and in particular collective communications). A major difference with our work is that these studies focus on recovery overhead and do not address overall performance nor yield.

3 Performance model

This section reviews the key parameters of the performance model. Some assumptions are made to simplify the computation of the yield. We discuss possible extensions in Section 6.

3.1 Application/platform framework

We consider perfectly parallel applications that execute on homogeneous parallel platforms. Without loss of generality, we assume that each processor has unit speed: we only need to know that the total amount of work done by p processors within T seconds requires $\frac{p}{q}T$ seconds with q processors.

3.2 Mean Time Between Failures (MTBF)

Each processor is subject to failures which are IID (independent and identically distributed) random variables following an Exponential probability distribution of mean μ_{ind}, the individual processor MTBF. Then the MTBF of a section of the platform comprised of i processors is given by $\mu_i = \frac{\mu_{ind}}{i}$ [15].

3.3 Checkpoints

Processors checkpoint periodically, using the optimal Young/Daly period [22, 7]: for an application using i processors, this period is $\sqrt{2C_i\mu_i}$, where C_i is the time to checkpoint with i processors. We consider two cases to define C_i. In both cases, the overall application memory footprint is considered constant at Mem_{tot}, so the size of individual checkpoints is inversely linear with the number of participating/surviving processors. In the first case, the I/O bandwidth is the bottleneck (which is often the case in HPC platforms

– it takes only a few processors to saturate the I/O bandwidth); then the checkpoint cost is constant and given by $C_i = \frac{Mem_{tot}}{\tau_{io}}$, where τ_{io} is the aggregated I/O bandwidth. In the second case, the processor network card is the bottleneck (which is the case for in-memory checkpointing, or checkpointing to NVRAM), and the checkpoint cost is inversely proportional to number of active processors: $C_i = \frac{Mem_{tot}}{\tau_{xnet} \times i}$, where τ_{xnet} is the available network bandwidth, and $\frac{Mem_{tot}}{i}$ the checkpoint size.

We denote the recovery time with i processors as R_i. For all simulations we use $R_i = C_i$, assuming that the read and write bandwidths are identical.

3.4 Wait Time

Job schedulers allocate nodes to given applications for a given time. They aim at optimizing multiple criteria (depending on the center policy) among which fairness (balancing the job requests between users or accounts), platform utilization (minimizing the number of resources that are idling), job makespan (providing the answer as fast as possible). Combined with a high resource utilization (node idleness is usually in the single digit percentage for a typical HPC platform), a job has to wait a *Wait Time* (D) between its submission and the begining of its execution.

Job schedulers implement the selection based on the list of submitted jobs, each job defining how many processors it needs and for how long. That definition is, in most cases, unchangeable: an application may use less resource than what it requested, but the account will be billed for the requested resource, and it will not be able to re-dimension the allocation during the execution.

Thus, if after some failures, an application has not enough resource left to efficiently complete, it will have to relinquish the allocation, and request a new one. During the wait time D, the application does not execute any computation to progress towards completion: its yield is zero during D seconds.

4 Expected yield

This section is the core of the paper. We compute the expected yield for each application type, Rigid, Moldable and GridShaped.

4.1 Rigid application

We first consider a Rigid application that can be parallelized at compile-time to use any number of processors but cannot change this number until it reaches termination. There are N processors allocated to the application. We use $N-F$ for execution and keep F as spares. The execution is protected from failures by checkpoints of duration C_{N-F}. Each failure striking the

application will incur an in-place restart of duration R_{N-F}, using a spare processor to replace the faulty one. However, when the $(F+1)^{st}$ failure strikes, the job will have to stop and perform a full restart, waiting for a new allocation of N processors to be granted by the job scheduler.

We define $\mathcal{T}_R$ as the expected duration of an execution period until the $(F+1)^{st}$ failure strikes. The first failure is expected to strike after μ_N seconds, the second failure μ_{N-1} seconds after the first one, and so on. Without any overhead, the length of a period would be $\sum_{i=N}^{N-F} \mu_i$. Except for the last failure, each failure incurs some overhead only if it strikes the application. This happens with probability $\frac{N-F}{i}$, where i is the current number of live processors. In that case, the failure requires a restart and some re-execution, namely half the checkpoint period in average. The application always uses $N-F$ processors, hence the checkpoint period remains equal to $\sqrt{2C_{N-F}\mu_{N-F}}$ —as a first-order approximation, we assume that no failure occurs during restart and re-execution, thereby neglecting the probability of two failures within a short time window. On the contrary, if the failure strikes a spare, there is no overhead. The last failure always requires a wait time, and then a restart and re-execution. Therefore, we derive:

$$\mathcal{T}_R = \sum_{i=N}^{N-F} \mu_i + \sum_{i=N}^{N-F+1} \frac{N-F}{i} \left(R_P + \frac{\sqrt{2C_{N-F}\mu_{N-F}}}{2} \right) + D + R_P + \frac{\sqrt{2C_{N-F}\mu_{N-F}}}{2}$$

What is the total amount of work $\mathcal{W}_R$ computed during a period? During the sub-period of length μ_i, there are $\frac{\mu_i}{\sqrt{2C_{N-F}\mu_{N-F}}}$ checkpoints, each of length C_{N-F}, and each processor works during $\frac{\mu_i}{1+\frac{C_{N-F}}{\sqrt{2C_{N-F}\mu_{N-F}}}}$ seconds. There are $N-F$ processors at work, hence

$$\mathcal{W}_R = (N-F) \cdot \sum_{i=N}^{N-F} \frac{\mu_i}{1+\frac{C_{N-F}}{\sqrt{2C_{N-F}\mu_{N-F}}}}$$

During the duration $\mathcal{T}_R$ of the period, in the absence of failures and protection, the application could have used all N processors to compute. Thus the effective yield with protection for the application during $\mathcal{T}_R$ is reduced to $\mathcal{Y}_R$:

$$\mathcal{Y}_R = \frac{\mathcal{W}_R}{N \cdot \mathcal{T}_R}$$

4.2 Moldable Application

We now consider a MOLDABLE application that can use a different number of processors after each restart. The application starts executing with N processors; after the first failure, the application recovers from the last

checkpoint and is able to continue with only $N-1$ processors after paying the restart cost R_{N-1}, albeit with a slowdown factor $\frac{N-1}{N}$ of the parallel work per time unit.

We define $\mathcal{T}_M$ as the expected duration of an execution period until the $(F+1)^{st}$ failure strikes. Without any overhead, the length of a period would be $\sum_{i=N}^{N-F} \mu_i$, the same as for RIGID applications. But there are few differences. First, each failure strikes the application, since it always uses all live processors. Second, the checkpoint period increases after each failure, since the number of live processors decreases. Third, the re-execution after a failure (except the last one) incurs a slowdown factor because we move from i processors to $i-1$ processors. Fourth and finally, the re-execution after the last failure is performed faster, because there are more live processors. Altogether, we derive that

$$\mathcal{T}_M = \sum_{i=N}^{N-F} \mu_i + \sum_{i=N}^{N-F+1} \left(R_{i-1} + \frac{i}{i-1} \cdot \frac{\sqrt{2C_i\mu_i}}{2} \right) + D + R_N + \frac{N-F}{N} \frac{\sqrt{2C_{N-F}\mu_{N-F}}}{2}$$

To compute the total amount of work $\mathcal{W}_M$ during a period, we proceed as before and consider each sub-period. During the sub-period of length μ_i, there are $\frac{\mu_i}{\sqrt{2C_i\mu_i}}$ checkpoints, each of length C_i, and each processor works during $\frac{\mu_i}{1+\frac{C_i}{\sqrt{2C_i\mu_i}}}$ seconds. And there are i processors at work during that sub-period. Altogether:

$$\mathcal{W}_M = \sum_{i=N}^{N-F} i \times \frac{\mu_i}{1+\frac{C_i}{\sqrt{2C_i\mu_i}}}$$

The yield of the MOLDABLE application is then:

$$\mathcal{Y}_M = \frac{\mathcal{W}_M}{N \cdot \mathcal{T}_M}$$

4.3 GridShaped application

Finally, we consider a GRIDSHAPED application, defined as a moldable execution which requires a rectangular processor grid. The application starts with a square $p \times p$ grid of $N = p^2$ processors. After the first failure, execution continues on a $p \times (p-1)$ rectangular grid, keeping $p-1$ processors as spares for the next $p-1$ failures. After p failures, the grid is shrunk again to a $(p-1) \times (p-1)$ square grid, and the execution continues on this reduced-size square grid. After how many failures F should the application stop, in order to maximize the application yield?

The derivation of the expected length of a period and of the total work is more complicated for GRIDSHAPED than for RIGID and MOLDABLE. To

simplify the presentation, we outline the computation of the yield only for values of F of the form $F = 2pf + 1$, hence $p^2 = F + (p - f)^2$, meaning that we stop shrinking and request a new allocation when reaching a square grid of size $(p - f) \times (p - f)$ for some value of $f < p$ to be determined. Obviously, we could stop after any number of faults F, and the publicly available software [18] shows how to compute the optimal value of F without any restriction.

We start by computing an auxiliary variable: the expected time $T_G(p_1, p_2)$ to move from a $p_1 \times p_2$ grid to a $(p_1 - 1) \times p_2$ grid, where $p_1 \geq p_2$. Without restart and re-execution, this time is $\sum_{i=p_1p_2}^{(p_1-1)p_2} \mu_i$. The first failure calls for a restart $R_{p_1p_2}$ and re-execution (at a reduced pace on less resources) of duration $\frac{p_1}{p_1-1}\frac{\sqrt{2C_{p_1p_2}\mu_{p_1p_2}}}{2}$. The j^{th} failure, for $2 \leq j \leq p_2$, will strike the application with probability $\frac{(p_1-1)p_2}{p_1p_2-j+1}$, because it is using $(p_1 - 1)p_2$ processors and keeping $p_1 - j + 1$ spares. The checkpoint period evolves with the number of processors, just as for MOLDABLE applications. We derive:

$$\begin{aligned} T_G(p_1, p_2) = & \sum_{i=p_1p_2}^{(p_1-1)p_2+1} \mu_i \\ & + R_{(p_1-1)p_2} + \frac{p_1}{p_1 - 1}\frac{\sqrt{2C_{p_1p_2}\mu_{p_1p_2}}}{2} \\ & + \sum_{i=p_1p_2-1}^{(p_1-1)p_2+1} \frac{(p_1 - 1)p_2}{i}\left[R_{(p_1-1)p_2} + \frac{\sqrt{2C_{(p_1-1)p_2}\mu_{(p_1-1)p_2}}}{2}\right] \end{aligned}$$

Going from p^2 processors down to $(p-f)^2$ processors thus require a time

$$\begin{aligned} \mathcal{T}_G = & \sum_{g=0}^{f-1} [T_G(p - g, p - g) + T_G(p - g, p - g - 1)] \\ & + \mu_{(p-f)^2} + D + R_{p^2} + \frac{(p - f)^2}{p^2}\frac{\sqrt{2C_{(p-f)^2}\mu_{(p-f)^2}}}{2} \end{aligned}$$

The last restart and sped-up re-execution are the same as for RIGID or MOLDABLE applications.

Similarly, we define the auxiliary variable $W_G(p_1, p_2)$ as the parallel work when moving from a $p_1 \times p_2$ grid to a $(p_1 - 1) \times p_2$ grid, where $p_1 \geq p_2$. There are p_1p_2 processors working during the first sub-period, and $(p_1-1)p_2$ during the following ones. We readily obtain

$$\begin{aligned} W_G(p_1, p_2) = & p_1p_2\mu_{p_1p_2}\left(1 - \frac{C_{p_1p_2}}{\sqrt{2C_{p_1p_2}\mu_{p_1p_2}}}\right) \\ & + \sum_{i=p_1p_2-1}^{(p_1-1)p_2+1} (p_1 - 1)p_2\mu_i\left(1 - \frac{C_{(p_1-1)p_2}}{\sqrt{2C_{(p_1-1)p_2}\mu_{(p_1-1)p_2}}}\right) \end{aligned}$$

Going from p^2 processors down to $(p-f)^2$ processors thus corresponds to a total work

$$\mathcal{W}_G = \sum_{g=0}^{f-1} \left(W_G(p-g, p-g) + W_G(p-g, p-g-1)\right) + (p-f)^2 \times \mu_{(p-f)^2} \left(1 - \frac{C_{(p-f)^2}}{\sqrt{2C_{(p-f)^2}\mu_{(p-f)^2}}}\right)$$

The yield of the GridShaped application is then:

$$\mathcal{Y}_G = \frac{\mathcal{W}_G}{N \cdot \mathcal{T}_G}$$

where $N = p^2$.

5 Applicative scenarios

We consider several applicative scenarios in this section. We start with a platform inspired from existing ones in Section 5.1 and then we study the impact of several key parameters in Section 5.2.

5.1 Main scenario

As a main applicative scenario, we consider a platform with 22,250 nodes (150^2), with a node MTBF of 20 years, and an application that would take 2 minutes to checkpoint (at 22,250 nodes). In other words, we let $N = 22,500$, $\mu_{ind} = 20y$ and $C_i = C = 120s$. These values are inspired from existing platforms: the Titan supercomputer at OLCF [12], for example, holds 18,688 nodes, and experiences a few node failures per day, implying a node MTBF between 18 and 25 years. The filesystem has a bandwidth of 1.4TB/s, and nodes altogether aggregate 100TB of memory, thus a checkpoint that would save 30% of that system should take in the order of 2 minutes to complete. In other words, $C_i = C = 120$ seconds for all $i \leq 18,688$.

Figure 2 shows the yield that can be expected if doing a full restart after an optimal number of failures, as a function of the wait time, for the three kind of applications considered (Rigid, Moldable and GridShaped). We also plot the expected yield when the application experiences a full restart after each failure (NoSpare). First, one sees that the three approaches that avoid paying the cost of a wait time after every failure experience a comparable yield, while the performance of the NoSpare approach quickly degrades to a small efficiency (30% when the wait time is around 14h).

The zoom box to differentiate the Rigid, Moldable and GridShaped yield shows that the Moldable approach has a slightly higher yield than the other ones, but only for a minimal fraction of the yield. This is expected, as

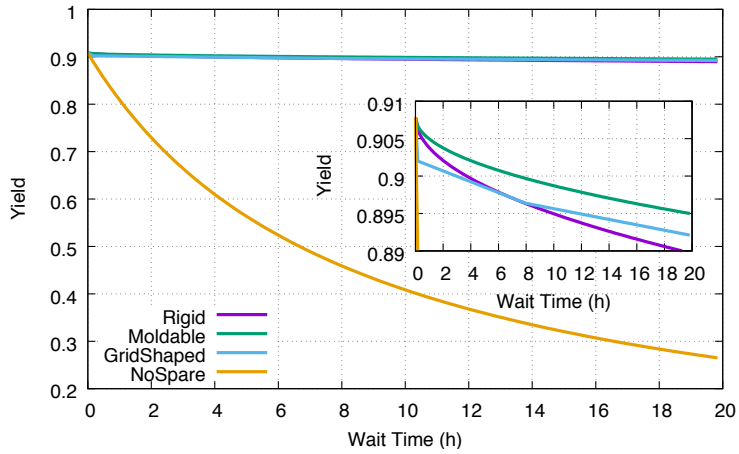


Figure 2: Optimal yield as function of the wait time, for the different types of applications.

the Moldable approach takes advantage of all living processors, while the GridShaped and Rigid approaches sacrifice the computing power of the spare nodes waiting for the next failure. However, the size of the gain is small to the point of being negligible. The GridShaped approach experiences a yield that changes in steps, oscillating around the Moldable yield. Both these phenomenons are explained by the next figure.

Figure 3 shows the number of failures after which the application should do a full restart, to obtain an optimal yield, as a function of the wait time, for the three kind of applications considered. We observe that this optimal is quickly reached: even with long wait times (e.g. 10h), 200 to 250 failures (depending on the method) should be tolerated within the allocation before relinquishing it. This is small compared to the number of nodes: less than 1% of the resource should be dedicated as spares for the Rigid approach, and after losing 1% of the resource, the Moldable approach should request a new allocation.

This is remarkable, taking into account the poor yield obtained by the approach that does not tolerate failures within the allocation. Even with a small wait time (assuming the platform would be capable of re-scheduling applications that experience failures in less than 2h), Figure 2 shows that the yield of the NoSpare approach would decrease to 70%. This represents a waste of 30%, which is much higher than the recommended waste of 10% for resilience in the current HPC platforms recommendations [6, 4]. Comparatively, provisioning, within the allocations, only 1% of additional resource would allow to maintain a yield at 90%, for every approach considered.

The GridShaped approach experiences steps that correspond to using all the spares created when redeploying the application over a smaller grid

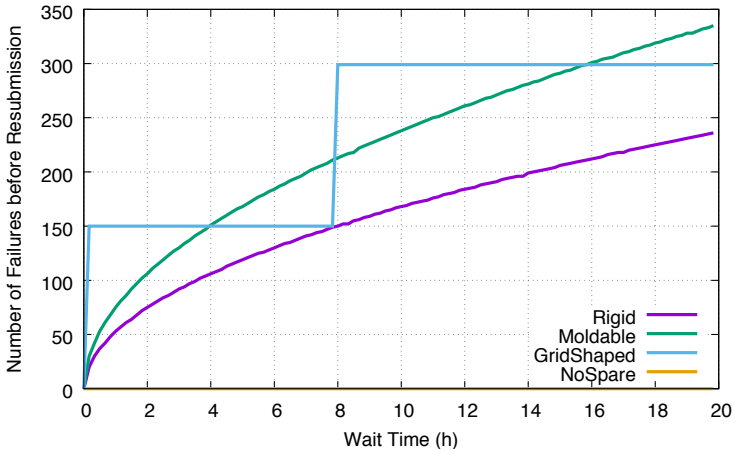


Figure 3: Optimal number of failures tolerated between two full restarts, as function of the wait time, for the different types of applications.

before relinquishing the allocation. As illustrated in Figure 2, the yield evolves in steps, changing the slope of a linear approximation radically when redeploying over a smaller grid. This has for consequence that the maximal yield is always at a slope change point, thus at the frontier of a new grid size. It is still remarkable that even with very small wait times, it is more beneficial to use spares (and thus to lose a full row of processors) than to redeploy immediately.

Figure 4 shows the maximal length of an allocation: after such duration, the job will have to fully restart in order to maintain the optimal yield. This figure illustrates the real difference between the RIGID and MOLDABLE approaches: although both approaches are capable of extracting the same yield, the MOLDABLE approach can do so with significantly longer periods between full restarts. This is important when considering real life applications, because this means that the applications using a MOLDABLE approach have a higher chance to complete before the first full restart, and overall will always complete in a lower number of allocations than the RIGID approach.

Finally, Figure 5 shows an upper limit of the duration of the wait time in order to guarantee a given yield for the three applications. In particular, we see that to reach a yield of 90%, an application which would restart its job at each fault would need that restart to be done in less than 6 minutes whereas the RIGID and GRIDSHAPED approaches need a full restart in less than 3 hours approximately. This bound goes up to 7 hours for the MOLDABLE approach. In comparison, with a wait time of 1 hour, the yield obtained using NOSPARE is only 80%. This shows that, using these parameters, it seems impossible to guarantee the recommended waste of 10% without tolerating (a small) number of failures before rescheduling the job.

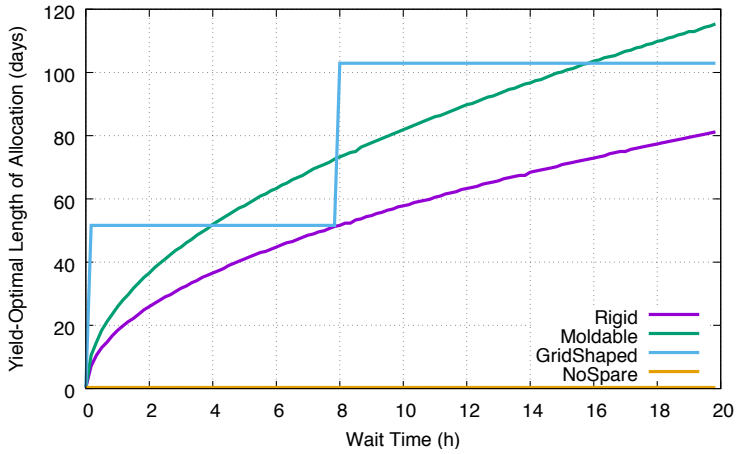


Figure 4: Optimal length of allocations, for the different types of applications.

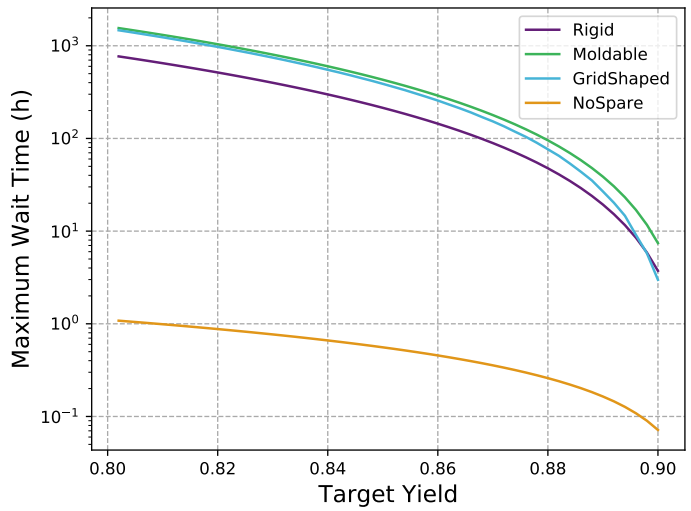
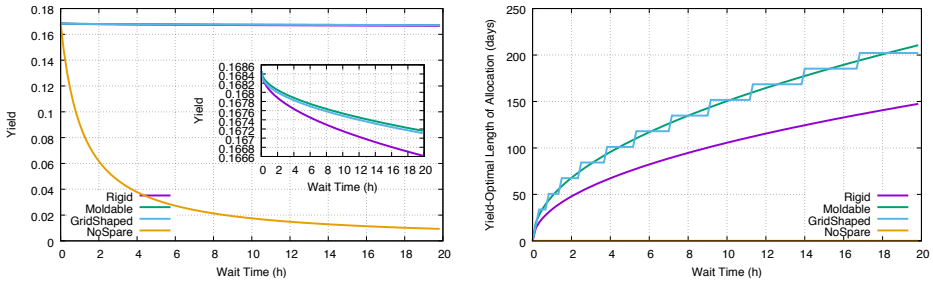
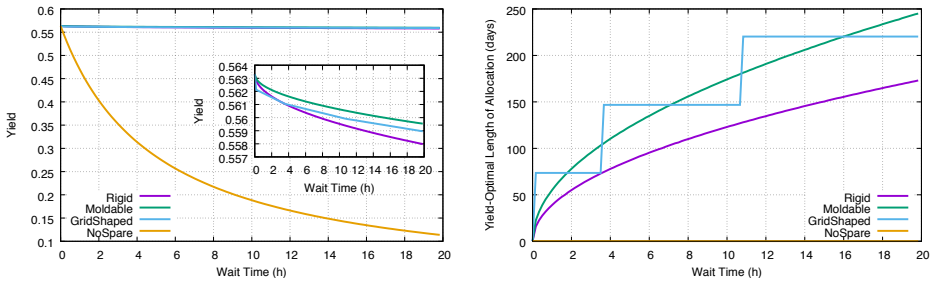


Figure 5: Maximum wait time allowed to reach a target yield.



(a) $\mu_{ind} = 5$ years



(b) $\mu_{ind} = 50$ years

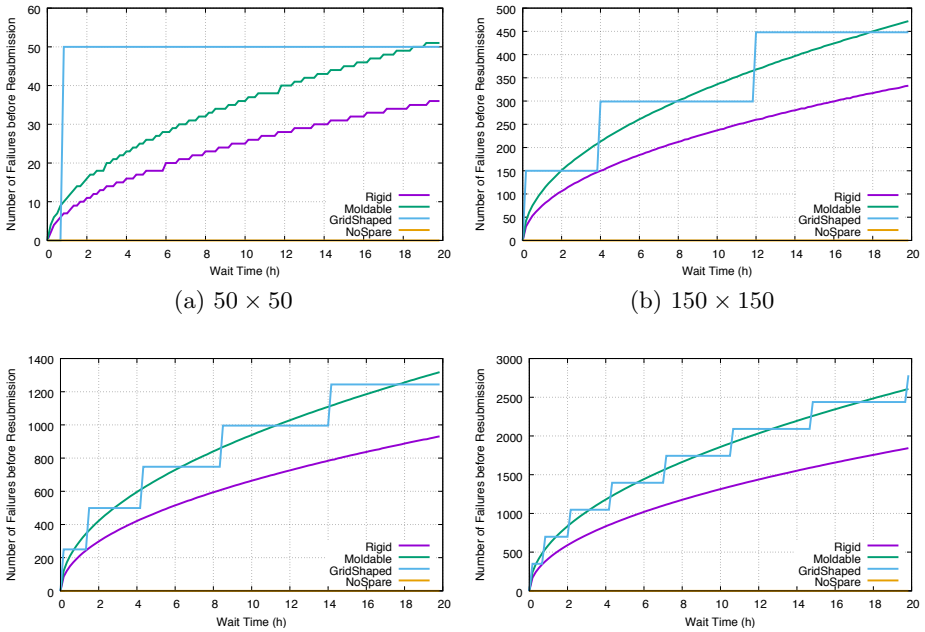
Figure 6: Yield and optimal allocation length of as a function of the wait time with $N = 350 \times 350$, and $C = 30$ minutes.

5.2 Varying key parameters

We performed extensive simulations to assess the impact of key parameters. We tried all combinations of MTBF (5 years, 10 years, 20 years, 50 years), checkpointing cost (2 minutes, 10 minutes, 30 minutes, 60 minutes) and application size ($50 \times 50 = 2500$, $150 \times 150 = 22500$, $250 \times 250 = 62500$, $350 \times 350 = 122500$). Not all results are presented for conciseness, but they all give very similar results compared to the main scenario of Section 5.1.

Figure 6 shows the yield and the corresponding allocation length for two (extreme) values of the MTBF, when using the largest application size $N = 350 \times 350$. The top subfigure is for $\mu_{ind} = 5$ years while the bottom subfigure is for $\mu_{ind} = 50$ years. As expected, the yield increases when the MTBF decreases. However, the variation of μ_{ind} only slightly impacts the allocation length (which stays around 100 days for the MOLDABLE approach with a wait time of 4 hours). The only major difference is for the GRIDSHAPED, whose allocation length closely follows that of MOLDABLE: when the MTBF is low, it is better to tolerate a larger number of failures before resubmission. Since the size of the steps are defined by the application size, the impact on the allocation length becomes smaller when there are more errors.

Figure 7 shows the optimal number of faults to tolerate for the four



(a) 50×50 (b) 150×150

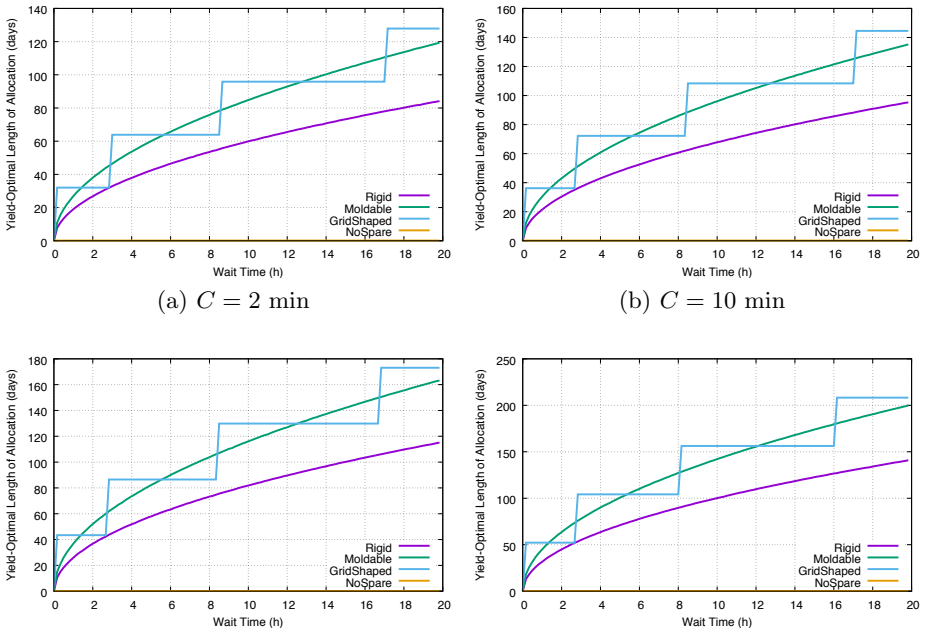
(c) 250×250 (d) 350×350

Figure 7: Optimal number of faults before rescheduling the application for different application sizes.

different application sizes (with $\mu_{ind} = 10$ years and $C_i = C = 10$ minutes). We can see from this experiment that the number of tolerated failures stays within a small pourcentage of the total number of processors. In particular, the last step of the GridShaped application corresponds to $\approx 2\%$ of the total application size in all the four cases.

Figure 8 aims at showing the impact of the checkpointing cost on the allocation length. The trend is that a higher checkpointing cost induces a longer allocation length. This can be explained by the fact that the allocation length takes into account the checkpoint/restart strategy into its computation. Thus longer checkpoints induces more time spent for an equivalent work done. Overall, the impact of the checkpointing cost stays minimal compared to the impact of the wait time or the MTBF.

Finally, Figure 9 describes the yield obtained when using different models for the checkpointing cost: either the checkpoint is constant (independent of the number of processors: left figure) or it is inversely proportional to the number of processors (right figure). As these plots show, the difference between the two models does not have a noticeable impact on the yield of the applications. This can be explained as follows: as Figure 7 showed, only a small number of faults is allowed before resubmission, in comparison to



(a) $C = 2$ min (b) $C = 10$ min

(c) $C = 30$ min (d) $C = 60$ min

Figure 8: Optimal number of faults before rescheduling the application for different checkpointing costs.

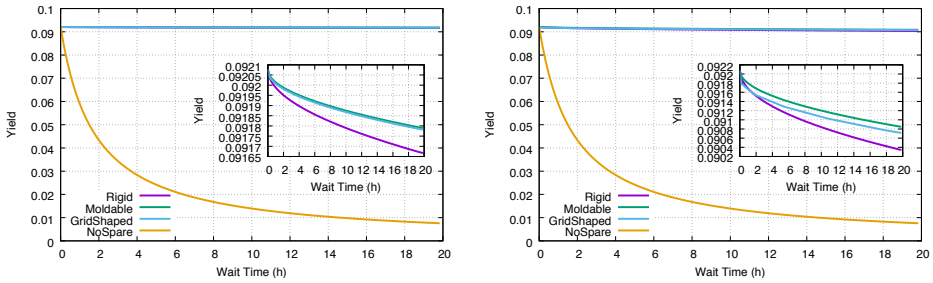


Figure 9: Constant checkpoint cost ($C_i = 60$ min) on the left, and increasing checkpoint cost ($C_i = \frac{N}{i} \times 60$ min) on the right, with $\mu_{ind} = 5$ years and $N = 350 \times 350$.

the application size. Changing the number of active processors by a few percentage does not really make a difference for the checkpoitnt cost, which remains almost the same in both models.

6 Conclusion

In this paper, we have compared the performance of Rigid, Moldable and GridShaped applications when executed on large-scale failure-prone platforms. For each application type, we have computed the optimal number of faults that should be tolerated before requesting a new allocation, as a function of the wait time. Through realistic applicative scenarios inspired by state-of-the-art platforms, we have shown that the three application types experience an optimal yield when requesting a new allocation after experiencing a number of failures that represents a small percentage of the initial number of resources (hence a small percentage of spares for Rigid applications), and this even for large values of the wait time. On the contrary, the NoSpare strategy, where a new allocation is requested after each failure, sees its yield dramatically decrease when the wait time increases. We also observed that Moldable applications enjoy much longer execution periods in between two re-allocations, thereby decreasing the total execution time as compared to Rigid applications (and GridShaped applications lying in between).

Future work will be devoted to exploring more applicative scenarios. We also intend to extend the model in several directions. On the application side, we aim at dealing with non-perfectly parallel applications but instead with applications whose speedup profile obeys Amdahl's law [1]. We will also introduce a more refined speedup profile for GridShaped applications, with an execution speed that depends on the grid shape (a square being usually faster than an elongated rectangle). On the resilience side, we will

address forward-recovery schemes, such as ABFT [14, 8], in replacement of, or in combination with, checkpoint-restart techniques.

References

[1] G. Amdahl. The validity of the single processor approach to achieving large scale computing capabilities. In *AFIPS Conference Proceedings*, volume 30, pages 483–485. AFIPS Press, 1967.

[2] Rizwan A. Ashraf, Saurabh Hukerikar, and Christian Engelmann. Shrink or substitute: Handling process failures in HPC systems using in-situ recovery. *CoRR*, abs/1801.04523, 2018.

[3] Wesley Bland, Aurelien Bouteiller, Thomas Herault, George Bosilca, and Jack Dongarra. Post-failure recovery of MPI communication capability: Design and rationale. *International Journal of High Performance Computing Applications*, 27(3):244–254, 2013.

[4] Franck Cappello, Al Geist, William Gropp, Sanjay Kale, Bill Kramer, and Marc Snir. Toward exascale resilience: 2014 update. *Supercomputing frontiers and innovations*, 1(1), 2014.

[5] Walfredo Cirne and Francine Berman. Using moldability to improve the performance of supercomputer jobs. *J. Parallel Distrib. Comput.*, 62(10):1571–1601, 2002.

[6] CORAL: Collaboration of Oak Ridge, Argonne and Livermore National Laboratorie. Draft CORAL-2 build statement of work. Technical Report LLNL-TM-7390608, Lawrence Livermore National Laboratory, March, 30 2018.

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

[8] Peng Du, Aurelien Bouteiller, et al. Algorithm-based fault tolerance for dense matrix factorizations. In *PPoPP*, pages 225–234. ACM, 2012.

[9] Pierre-François Dutot, Grégory Mounié, and Denis Trystram. Scheduling parallel tasks approximation algorithms. In Joseph Y.-T. Leung, editor, *Handbook of Scheduling - Algorithms, Models, and Performance Analysis*. CRC Press, 2004.

[10] Aiman Fang, Hajime Fujita, and Andrew A. Chien. Towards understanding post-recovery efficiency for shrinking and non-shrinking recovery. In *Euro-Par 2015: Parallel Processing Workshops*, pages 656–668. Springer, 2015.

[11] Yanfei Guo, Wesley Bland, Pavan Balaji, and Xiaobo Zhou. Fault tolerant MapReduce-MPI for HPC clusters. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2015, Austin, TX, USA, November 15-20, 2015*, pages 34:1–34:12, 2015.

[12] Saurabh Gupta, Tirthak Patel, Christian Engelmann, and Devesh Tiwari. Failures in large scale systems: Long-term measurement, analysis, and implications. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '17, pages 44:1–44:12, New York, NY, USA, 2017.

[13] Atsushi Hori, Kazumi Yoshinaga, Thomas Herault, Aurélien Bouteiller, George Bosilca, and Yutaka Ishikawa. Sliding substitution of failed nodes. In *Proceedings of the 22Nd European MPI Users' Group Meeting*, EuroMPI '15, pages 14:1–14:10, New York, NY, USA, 2015. ACM.

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

[15] Thomas Hérault and Yves Robert, editors. *Fault-Tolerance Techniques for High-Performance Computing*, Computer Communications and Networks. Springer Verlag, 2015.

[16] J. E. Moreira and V. K. Naik. Dynamic resource management on distributed systems using reconfigurable applications. *IBM Journal of Research and Development*, 41(3):303–330, 1997.

[17] Suraj Prabhakaranw. *Dynamic Resource Management and Job Scheduling for High Performance Computing*. PhD thesis, Technische Universität Darmstadt, 2016.

[18] Simulation Software. Computing the yield. `https://github.com/vlefevre/continuability`, 2018.

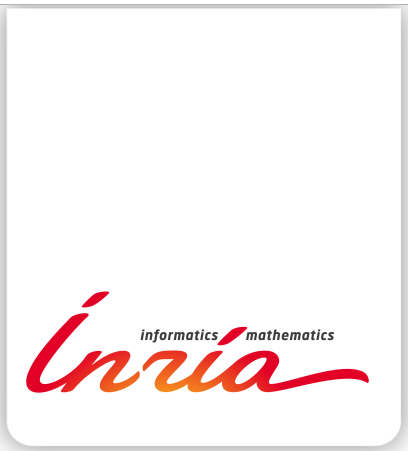
[19] Rajesh Sudarsan and Calvin J. Ribbens. Design and performance of a scheduling framework for resizable parallel applications. *Parallel Computing*, 36(1):48–64, 2010.

[20] Rajesh Sudarsan, Calvin J. Ribbens, and Diana Farkas. Dynamic resizing of parallel scientific simulations: A case study using LAMMPS. In *Int . Conf . Computational Science ICCS*, pages 175–184. Procedia, 2009.

[21] Keiji Yamamoto, Atsuya Uno, Hitoshi Murai, Toshiyuki Tsukamoto, Fumiyoshi Shoji, Shuji Matsui, Ryuichi Sekizawa, Fumichika Sueyasu, Hiroshi Uchiyama, Mitsuo Okamoto, Nobuo Ohgushi, Katsutoshi

Takashina, Daisuke Wakabayashi, Yuki Taguchi, and Mitsuo Yokokawa. The K computer Operations: Experiences and Statistics. *Procedia Computer Science*, 29:576–585, 2014. Int. Conf. Computational Science (ICCS).

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



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