

Improving Performance of Iterative Methods by Lossy Checkpointing

Dingwen Tao
University of California, Riverside
Riverside, CA, USA
dtao001@cs.ucr.edu

Sheng Di
Argonne National Laboratory
Lemont, IL, USA
sdi1@anl.gov

Xin Liang
University of California, Riverside
Riverside, CA, USA
xliang007@ucr.edu

Zizhong Chen
University of California, Riverside
Riverside, CA, USA
chen@cs.ucr.edu

Franck Cappello
Argonne National Laboratory
Lemont, IL, USA
cappello@mcs.anl.gov

ABSTRACT

Iterative methods are commonly used approaches to solve large, sparse linear systems, which are fundamental operations for many modern scientific simulations. When the large-scale iterative methods are running with a large number of ranks in parallel, they have to checkpoint the dynamic variables periodically in case of unavoidable fail-stop errors, requiring fast I/O systems and large storage space. To this end, significantly reducing the checkpointing overhead is critical to improving the overall performance of iterative methods. Our contribution is fourfold. (1) We propose a novel lossy checkpointing scheme that can significantly improve the checkpointing performance of iterative methods by leveraging lossy compressors. (2) We formulate a lossy checkpointing performance model and derive theoretically an upper bound for the extra number of iterations caused by the distortion of data in lossy checkpoints, in order to guarantee the performance improvement under the lossy checkpointing scheme. (3) We analyze the impact of lossy checkpointing (i.e., extra number of iterations caused by lossy checkpointing files) for multiple types of iterative methods. (4) We evaluate the lossy checkpointing scheme with optimal checkpointing intervals on a high-performance computing environment with 2,048 cores, using a well-known scientific computation package PETSc and a state-of-the-art checkpoint/restart toolkit. Experiments show that our optimized lossy checkpointing scheme can significantly reduce the fault tolerance overhead for iterative methods by 23%~70% compared with traditional checkpointing and 20%~58% compared with lossless-compressed checkpointing, in the presence of system failures.

KEYWORDS

Iterative Methods; Numerical Linear Algebra; Resilience; Checkpoint/Restart; Lossy Compression; Performance Optimization

Corresponding author: Sheng Di, Mathematics and Computer Science Division, Argonne National Laboratory, 9700 Cass Avenue, Lemont, IL 60439, USA.

ACM acknowledges that this contribution was authored or co-authored by an employee, contractor, or affiliate of the United States government. As such, the United States government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for government purposes only.

HPDC '18, June 11–15, 2018, Tempe, AZ, USA

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-5785-2/18/06...\$15.00

<https://doi.org/10.1145/3208040.3208050>

ACM Reference Format:

Dingwen Tao, Sheng Di, Xin Liang, Zizhong Chen, and Franck Cappello. 2018. Improving Performance of Iterative Methods by Lossy Checkpointing. In *HPDC '18: International Symposium on High-Performance Parallel and Distributed Computing, June 11–15, 2018, Tempe, AZ, USA*. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3208040.3208050>

1 INTRODUCTION

Scientific simulations involving partial differential equations (PDEs) require solving sparse linear system within each timestep. At large scale, sparse linear systems are solved by using iterative methods, such as the conjugate gradient (CG) method. Thus, iterative methods are one of the most crucial components determining the scalability and efficiency of HPC applications. For example, Becciani et al. [11] presented a study of solving a 5-parameter astrometric catalogue at the micro-arcsecond level for about 1 billion stars of our Galaxy under a cornerstone mission (called Gaia) launched by European Space Agency. Their experimental results show that solving the resulting sparse linear system of 7.2×10^{10} equations for the last period of the Gaia mission can take 1,000 to 4,000 iterations for convergence, totaling up to 1.96×10^5 seconds (i.e., more than 54 hours) on 2,048 BlueGeneQ nodes. When running on high-performance computing (HPC) environments using potentially tens of thousands of nodes and millions of cores for hours or days towards exascale computing [9], fail-stop errors are inevitable. Accordingly, how to effectively protect the iterative methods against such failures is an important research issue, determining the overall performance of iterative methods in HPC environments.

Many algorithm-based fault tolerance approaches have been proposed to tolerate silent data corruptions with iterative methods, and they work efficiently because of little storage overhead. Tolerating fail-stop errors, however, is much more challenging because it requires checkpointing or saving multiple large vector data sets at runtime, leading to large checkpointing overhead.

For many PDE-based scientific simulations, the sparse linear system includes most of the variables that are involved in the application, so checkpointing for iterative methods determines overall checkpointing performance [28]. For example, SIMPLE (Semi-Implicit Method for Pressure-Linked Equations) [39] algorithm is a widely used numerical method to solve the Navier-Stokes equations [19] for Computational Fluid Dynamics (CFD) problems. For 3D CFD problems, there are totally nine fluid-flow scalar variables,

five of which need to be checkpointed during iterative methods. As a result, significantly improving the checkpointing performance of the iterative methods that it uses can significantly improve the application performance, since most of application state used by iterative methods. We refer readers to [38] for more details of 3D CFD problems and SIMPLE-like algorithms in parallel.

In this work, we propose an efficient execution scheme, specifically a lossy checkpointing scheme, in order to improve the overall performance for iterative methods running in the presence of failures. Unlike the traditional checkpointing approach, our lossy checkpointing scheme integrates a lossy compression technique into the checkpoint/restart model. That is, the checkpointing data is compressed by a lossy compressor before being moved to the parallel file system (PFS), which is an approach that can significantly reduce the run-time checkpointing overhead. Upon a failure, the latest checkpointing file is loaded and goes through a decompression step to reconstruct the checkpointing data for the recovery of the iterative execution.

Checkpoint/restart research has been conducted for decades in order to optimize the performance of various large-scale scientific executions, but lossy-compressed checkpointing is rarely studied. Lossy compressed checkpointing raises two challenging issues. (1) What is the impact of lossy checkpointing data on the execution performance? Specifically, can the iterative methods still converge, or how many extra iterations will be introduced after restarting from a lossy checkpoint? (2) Is adopting lossy compression in the checkpointing model a worthwhile method for improving the overall performance? Specifically, how much performance gain can be achieved based on the checkpoints with reduced size?

To address such two key issues, we make following contributions.

- We propose a novel lossy checkpointing scheme that significantly improves the performance for iterative methods. In particular, we exploit a lossy checkpointing scheme under which both the lossy compression and checkpointing can be performed efficiently for the iterative methods.
- We design a performance model that can formulate the overall performance of the execution with lossy checkpointing in the presence of failures. In particular, we derive an upper bound for the extra number of iterations caused by the lossy checkpoints against the reduced checkpointing overheads, which is a sufficient condition to determine whether the lossy checkpointing can get a performance gain for an iterative method in numerical linear algebra.
- We explore the impact of the lossy checkpointing on the extra number of iterations for multiple iterative methods, including stationary iterative methods, GMRES, and CG.
- We evaluate our lossy checkpointing scheme with optimized checkpointing intervals based on multiple iterative methods provided by PETSc, using both lossless and lossy compressors, on a parallel environment with up to 2,048 cores. Experiments show that our solution reduces the fault tolerance overhead by 23%~70% compared with traditional checkpointing and 20%~58% compared with lossless checkpointing.

The rest of the paper is organized as follows. In Section 2, we discuss related work. In Section 3, we describe the traditional checkpointing method without lossy compressors. In Section 4, we propose our lossy checkpointing scheme with state-of-the-art lossy compression techniques included, and we provide an in-depth analysis of checkpoint/restart overhead and the impact of the lossy checkpointing on convergence. In Section 5, we present our experimental evaluation results. In Section 6, we conclude with a brief discussion of future work.

2 RELATED WORK

Recently, a study of the Blue Waters system [12] showed that an event that required remedial repair action occurred every 4.2 hours on average and that systemwide events occurred approximately every 160 hours. To avoid remedial actions such as redoing computations, researchers have designed many fault tolerance techniques for HPC applications [15, 24, 32, 33, 52–54].

Checkpoint/Restart Techniques. One of the most widely used techniques is the checkpoint/restart model, and the corresponding optimization strategies have been studied for many years. Plank et al. [40] proposed a diskless checkpointing approach that reduces the checkpoint overhead by storing the checkpoints locally in processor memories. However, diskless checkpointing can survive only partial failures: it is unable to deal with the failure of the whole system. A multilevel checkpoint/restart model [10, 37] was proposed to provide tolerance for different types of failures. Fault Tolerance Interface (FTI) [10], for example, supports four levels of checkpointing: local storage device, partner-copy, Reed-Solomon encoding, and PFS. Di et al. [22, 23] proposed a multilevel checkpoint/restart model based on FTI to optimize the checkpoint intervals for different levels. In addition to the traditional checkpointing model, a few studies have demonstrated the feasibility of using compression techniques to improve the checkpointing performance. Islam et al. [29] adopted data-aware aggregation and lossless data compression to improve the checkpoint/restart performance. Sasaki et al. [45] proposed a lossy compression technique based on wavelet transformation for checkpointing and explored its impact in a production climate application. Calhoun et al. [14] verified the feasibility of using lossy compression in checkpointing two specific PDE simulations experimentally. Their results show that the compression errors in the checkpointing files can be masked by the numerical errors in the discretization, leading to improved performance without degraded overall accuracy in the simulation. To the best of our knowledge, our work is the first attempt to build a generic, theoretical performance model considering the impact of lossy compression techniques on the HPC checkpointing model and significantly improve the overall performance for multiple iterative methods, such as stationary iterative methods, GMRES, and CG.

Fault Tolerance Techniques for Iterative Methods. Iterative methods are widely used for solving systems of equations or computing eigenvalues of large sparse matrices. Although some fault-tolerant iterative methods have been designed, most are from an algorithmic level, and the performance is highly dependent on the specific characteristics of algorithms. For example, Tao et al. [50] proposed

an online algorithm-based fault tolerance (ABFT) approach to detect and recover soft errors for general iterative methods. For some specific iterative algorithms, Chen [17] developed an online ABFT approach for a subset of the Krylov methods by leveraging the orthogonality relationship of two vectors. Bridges et al. [13] and Elliot et al. [26] targeted GMRES based on its special characteristics and proposed a fault-tolerant version via selective reliability. Similar to that work, Sao and Vuduc [44] studied self-stabilizing corrections after error detection for CG algorithm. For fail-stop failures, Langou et al. [31] designed an algorithm-based recovery scheme for iterative methods, called lossy approach, that recovers an approximation of the lost data, but it is limited to the block Jacobi algorithm. Chen [16] proposed an algorithm-based recovery method that utilizes inherent redundant information for accurately recovering the lost data, but it is limited to the memory failure situation. Agullo et al. [4] proposed a technique that can recover from process failures followed by restarting strategies in Krylov subspace solvers where lost entries of the iterate are interpolated to define a new initial guess before restarting the Krylov method. Asynchronous iterations [6] proposed by Bahi et al. are linear solvers designed to tolerate message delays when applying the matrix in parallel.

Scientific Data Compression. Scientific data compression has been studied for years. The data compressors can be split into two categories: lossless and lossy. Lossless compressors make sure that the reconstructed data set after the decompression is exactly the same as the original data set. Such a constraint may significantly limit the compression ratio (up to 2 in general [42]) on the compression of scientific data. The reason is that scientific data are composed mainly of floating-point values and their tailing mantissa bits could be too random to compress effectively. State-of-the-art lossy compressors include SZ [25, 48], ZFP [34], ISABELA [30], FPZIP [35], SSEM [45], and NUMARCK [18]. Basically, they can be categorized into two models: prediction based and transform based. A prediction-based compressor predicts data values for each data point and encodes the difference between every predicted value and its corresponding real value based on a quantization method. Typical examples are SZ [25, 48], ISABELA [30], and FPZIP [35]. The block-transform-based compressor transforms the original data to another space where most of the generated data is very small (close-to-zero), such that the data can be stored with a certain loss in terms of user-required error bounds. For instance, SSEM [45] and ZFP [34] adopt a discrete Wavelet transform and a customized orthogonal transform, respectively. Lossy compression techniques, however, are used mainly for saving storage space and reducing the I/O cost of dumping the analysis data. How to make use of the lossy compressors to improve the checkpointing performance with iterative methods is still an open question.

3 TRADITIONAL CHECKPOINTING TECHNIQUE FOR ITERATIVE METHODS

Before presenting our lossy checkpointing scheme, we investigate the traditional checkpointing techniques for iterative methods.

According to a study of recovery patterns for iterative methods by Langou et al. [31], we need to classify the variables of the algorithms in order to form a fault-tolerant iterative method with the

checkpoint/recovery model. All the variables can be categorized into three types:

- Static variables: need to be stored once, for example, the system matrix A , the preconditioner matrix M , and the right-hand side vector b ;
- Dynamic variables: change along the iterations, for example, the approximate solution vector $x^{(i)}$;
- Recomputed variables: are worth being recomputed after a failure rather than being checkpointed; for example, the residual vector r can be recomputed by $r^{(i)} = b - Ax^{(i)}$. The term “worth” here means that recomputing some variables could be faster than obtaining them through a checkpoint.

Although the recomputed variables also need to be recovered during restarting after failures/errors, we still classify them as a separate category because they are recovered by a different strategy. How to recover a variable depends on the recovery overheads of the particular strategy. A scalar computed through global vector dot product, for example, is too expensive to compute, so it will be treated as a dynamic variable during the checkpointing.

After the classification is finished, we can form the fault-tolerant iterative methods with the checkpoint/recovery model as follows.

- Checkpoint
 - (1) Checkpoint static variables only at the beginning before going into the execution with iterations,
 - (2) Checkpoint dynamic variables every several iterations.
- Recovery
 - (1) Recover a correct computational environment,
 - (2) Recover static variables,
 - (3) Recover dynamic variables,
 - (4) Recover recomputed variables based on the reconstructed static and dynamic variables.

Based on this scheme, we can construct fault-tolerant iterative methods based on the checkpoint/recovery technique. We use the preconditioned CG algorithm as an example, as shown in Algorithm 1. This algorithm is one of the most commonly used iterative methods to solve sparse, symmetric, and positive-definite (SPD) linear systems. It computes successive approximations to the solution (vector $x^{(i)}$), residuals corresponding to the approximate solutions (vector $r^{(i)}$), and search directions (vector $p^{(i)}$) used to update both the approximate solutions and the residuals. Each iteration involves one sparse matrix-vector multiplication (line 10), three vector updates (lines 12, 13, and 17), and two vector inner products (lines 11 and 15). We refer readers to [8] for more details about CG method.

For the CG algorithm, the matrix A , preconditioner M , and right-hand side vector b are static variables. The number of iterations i , the scalar ρ , the direction vector $p^{(i)}$, and the approximate solution vector $x^{(i)}$ are dynamic variables. The residual vector $r^{(i)}$ is the recomputed variable, since we want to reduce checkpoint time and storage consumption. Based on the checkpoint/recovery model for iterative methods discussed above, we perform checkpointing for i , ρ , $p^{(i)}$, and $x^{(i)}$ every $ckpt_intvl$ iterations; and we perform recovering for A , M , i , ρ , $p^{(i)}$, and $x^{(i)}$ after a failure.

So far, we have constructed a fault-tolerant PCG solver with the checkpoint/recovery technique that has a strong resilience to failure-stop errors. Based on this scheme, we now can construct the

Algorithm 1 Fault-tolerant preconditioned conjugate gradient (PCG) algorithm with traditional checkpointing.

Input: linear system matrix A , preconditioner M , and right-hand side vector b

Output: approximate solution x

```

1: Compute  $r^{(0)} = b - Ax^{(0)}$ ,  $z^{(0)} = M^{-1}r^{(0)}$ ,  $p^{(0)} = z^{(0)}$ ,  $\rho_0 = r^{(0)T}z^{(0)}$  for some initial guess  $x^{(0)}$ 
2: for  $i = 0, 1, \dots$  do
3:   if  $((i > 0)$  and  $(i\%ckpt\_intvl = 0))$  then
4:     Checkpoint:  $i, \rho_i$  and  $p^{(i)}, x^{(i)}$ 
5:   end if
6:   if  $((i > 0)$  and  $(\text{recover}))$  then
7:     Recover:  $A, M, i, \rho_i, p^{(i)}, x^{(i)}$ 
8:     Compute  $r^{(i)} = b - Ax^{(i)}$ 
9:   end if
10:   $q^{(i)} = Ap^{(i)}$ 
11:   $\alpha_i = \rho_i / p^{(i)T}q^{(i)}$ 
12:   $x^{(i+1)} = x^{(i)} + \alpha_i p^{(i)}$ 
13:   $r^{(i+1)} = r^{(i)} - \alpha_i q^{(i)}$ 
14:  solve  $Mz^{(i+1)} = r^{(i+1)}$ 
15:   $\rho_{i+1} = r^{(i+1)T}z^{(i+1)}$ 
16:   $\beta_i = \rho_{i+1} / \rho_i$ 
17:   $p^{(i+1)} = z^{(i+1)} + \beta_i p^{(i)}$ 
18:  check convergence; continue if necessary
19: end for

```

fault-tolerant algorithm for any iterative method as follows. During the recovery, the first step is to recover a correct computational environment, such as an MPI environment. It is usually achieved by performing a global restart of the execution. Without loss of generality, we assume that the correct environment has been already recovered and that the recovered number of processors and tasks is the same as the previous failed one.

During the execution of iterative methods with checkpointing techniques, the overall checkpointing/restart cost is dominated by the dynamic variables instead of static variables. The reason is twofold. On the one hand, static variables are not involved in the checkpointing period but only the recovery step, while the optimal checkpointing frequency is generally considerably higher than the recovery frequency (i.e., failure rate). Suppose the mean time to interruption (MTTI) is 4 hours (i.e., 1 failure per 4 hours affecting the execution) and setting one checkpoint takes 18 seconds. Then the optimal checkpointing frequency is 5 checkpoints per hour according to Young's formula [55], which is 30 times as large as the failure rate. On the other hand, the static variables generally have comparable sizes with dynamic variables. Specifically, the static variables in the iterative methods are composed of the linear system matrix A , the preconditioner M , and right-hand side vector b . According to SuiteSparse Matrix Collection [20] (formerly known as University of Florida Sparse Matrix Collection), the number of nonzeros (i.e., the data that needs to be stored) in matrix A is usually of similar order to or a constant times (e.g., $1x \sim 10x$) large than the dimension of dynamic vectors. For preconditioner M , it can be much more sparse than A . For example, the most commonly used

preconditioning methods—*block Jacobi* and *incomplete LU factorization* (ILU)—need to store only the block diagonal matrix of A and the matrix L, U (where $A \approx LU$), respectively. Therefore, the data size of static variables is usually the same order as or a constant times large than that of dynamic variables. Taking these factors into account, we see that the overall checkpoint/restart overhead depends mainly on dynamic variables. Thus we focus mainly on reducing the checkpoint/recovery overhead of the dynamic variables in iterative methods by lossy compressors. Note that when we build the lossy checkpointing performance model (Section 4.3) and perform the evaluation (Section 5), we take into account all the three types of variables instead of only dynamic variables.

4 LOSSY CHECKPOINTING SCHEME FOR ITERATIVE METHODS

In this section, we first analyze the expected overhead of checkpointing techniques for iterative methods. We prove that reducing the checkpointing overhead (e.g., by leveraging compression techniques) can significantly improve the overall performance, especially for future HPC systems. This analysis motivates us to design an approach to reduce the checkpointing overhead. Then, we propose our lossy checkpointing scheme that can be easily applied to iterative methods in numerical linear algebra. We also present a new performance model for our lossy checkpointing scheme. Based on the model, we derive an upper bound for the number of extra iterations caused by lossy checkpoints against the reduced checkpointing overhead, to guarantee the performance improvement of the lossy checkpointing scheme. We theoretically and empirically analyze the impact of lossy checkpointing on the convergence of iterative methods considering multiple types of iterative methods.

4.1 Theoretical Analysis of Checkpointing Overhead for Iterative Methods

If a failure happens, we restart the computation from the last checkpointed variables, as shown in Algorithm 2. This process is normally called *rollback*. Rollback means that some previous computations need to be performed again. Thus, the checkpointing frequency or time interval needs to be determined carefully. Here the checkpointing interval means the mean time between two checkpoints. On the one hand, a larger checkpointing interval means a longer rollback in case of failure, indicating more workload to be recomputed after the recovery; on the other hand, a smaller checkpointing interval means more frequent checkpointing, leading to higher checkpointing overhead. How to calculate the optimal checkpointing intervals has been studied for many years [3, 55]. Our following analysis is based on the recovery pattern of iterative methods constructed by Langou et al. [31].

We use the notation in Table 1 to analyze the expected fault tolerance overhead. The overall execution time T_t can be expressed as

$$T_t = NT_{it} + T_{ckp} \frac{N}{k} + \frac{T_t}{T_f} (T_{rc} + T_{rb}).$$

Without loss of generality, based on Young's formula [55], the optimal checkpointing interval should be chosen as

$$k \cdot T_{it} = \sqrt{2T_f \cdot T_{ckp}}, \quad (1)$$

Table 1: Notation for traditional checkpointing model

T_{it}	Mean time of an iteration
T_{ckp}	Mean time to perform a checkpoint
T_{rc}	Mean time to recover the application with the correct environment and data from the last checkpoint
T_{rb}	Mean time to perform a rollback of some redundant computations
T_f	Mean time to interruption
$T_{overhead}^{CR}$	Mean time overhead of checkpoint/recovery
λ	Failure rate, i.e., $1/T_f$
k	Checkpoint frequency - a checkpoint is performed every k iterations
N	Number of iterations to converge without failures

and the expected mean time to perform a roll back, namely, T_{rb} , is $kT_{it}/2$. Thus,

$$\begin{aligned} T_t &= NT_{it} + T_{ckp} \frac{T_t}{\sqrt{2T_f \cdot T_{ckp}}} + \frac{T_t}{T_f} (T_{rc} + \frac{\sqrt{2T_f \cdot T_{ckp}}}{2}) \\ &= NT_{it} + T_t \left(\sqrt{\frac{2T_{ckp}}{T_f}} + \frac{T_{rc}}{T_f} \right) = NT_{it} + T_t (\sqrt{2\lambda T_{ckp}} + \lambda T_{rc}). \end{aligned}$$

Similar to [31], we therefore can get the expected overall execution time as

$$T_t = \frac{NT_{it}}{1 - \sqrt{2\lambda T_{ckp}} - \lambda T_{rc}}, \quad (2)$$

and the fault tolerance overhead is

$$T_{overhead}^{CR} = T_t - NT_{it} = NT_{it} \cdot \frac{\sqrt{2\lambda T_{ckp}} + \lambda T_{rc}}{1 - \sqrt{2\lambda T_{ckp}} - \lambda T_{rc}}, \quad (3)$$

where NT_{it} is the basic productive execution time with N iterations to converge. Note that in the paper, we use *fault tolerance overhead* to refer to the performance overhead caused by checkpoints/recoveries and failure events, which is equal to the total running time taking away the basic productive execution time (i.e., $T_t - NT_{it}$).

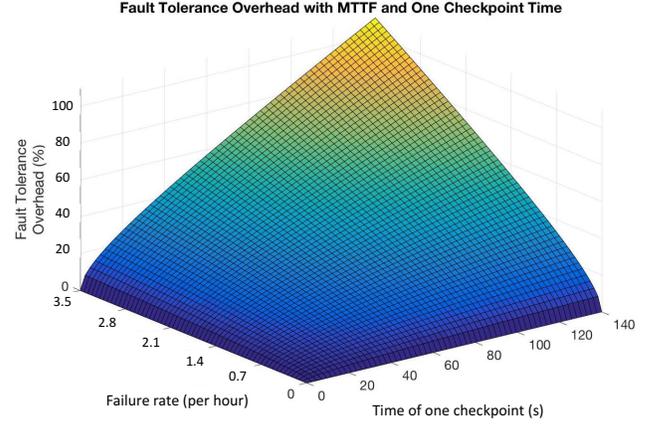
We assume $T_{rc} \approx T_{ckp}$ without loss of generality. Then we can simplify the expected fault tolerance overhead as follows.

$$T_{overhead}^{CR} \approx NT_{it} \cdot \frac{\sqrt{2\lambda T_{ckp}} + \lambda T_{ckp}}{1 - \sqrt{2\lambda T_{ckp}} - \lambda T_{ckp}} \quad (4)$$

Moreover, we can calculate the ratio of the expected fault tolerance overhead to the basic productive execution time as Equation (5).

$$\frac{T_{overhead}^{CR}}{NT_{it}} = \frac{\sqrt{2\lambda T_{ckp}} + \lambda T_{ckp}}{1 - \sqrt{2\lambda T_{ckp}} - \lambda T_{ckp}} \quad (5)$$

Now the expected fault tolerance overhead depends only on the failure rate λ and time of one checkpoint T_{ckp} . Based on this formula, we can plot the expected overhead of checkpoint/recovery based on different λ and T_{ckp} , as shown in Figure 1. We choose λ from 0 to 3.5 failures per hour (i.e., MTTF from about 20 minutes to infinity) and T_{ckp} from 0 to 140 seconds. Note that the MTTF represents the expected period at which the application execution is

**Figure 1: Expected fault tolerance overhead with different failure rates and checkpoint time.**

interrupted. Based on our experimental evaluation, checkpointing one dynamic vector x once without compression takes about 120 seconds with 2,048 processes/cores on the Bebop cluster [1] at Argonne National Laboratory. In our experiment, the number of elements in the vector is set to 10^{10} (with 78.8 GB double-precision floating-point data), which is the largest problem size that the three iterative methods (Jacobi, GMRES, and CG) can be run on the Bebop using 2,048 cores. We adopt the FTI library [10] with MPI-IO for checkpointing because of its high I/O efficiency confirmed in recent studies [51]. More details are presented in the experimental evaluation section.

Figure 1 illustrates that the expected fault tolerance overhead can be as high as 40% with $T_{ckp} = 120s$ if the MTTF is about hourly. On future extreme-scale systems with millions of components, the failure rate may be higher, and the fault tolerance overhead issue could be more severe. From Figure 1, we see that reducing the checkpointing time can significantly improve the overall performance of checkpoint/restart, especially under a higher error rate scenario.

4.2 Lossy Checkpointing Scheme for Iterative Methods

Our lossy checkpointing scheme based on an iterative method has two key steps.

- Compress dynamic variables with lossy compressor before each checkpointing.
- Decompress compressed dynamic variables after each recovering.

We still use the CG algorithm as an example, as shown in Algorithm 2, and the lossy checkpointing scheme can be applied to other iterative methods similarly. Because of space limitations, we present only the lossy checkpointing part without the original computations in Algorithm 2. The lossy compression and decompression procedures are marked in bold. We note that the CG algorithm maintains a series of orthogonality relations between the residual vectors r and the direction vectors p . Specifically, (1) $p^{(k)}$ and $Aq^{(j)}$, (2) $r^{(k)}$ and $p^{(j)}$, and (3) $r^{(k)}$ and $r^{(j)}$ are orthogonal to each other, where $j < k$. However, these orthogonality relations may be broken

Algorithm 2 Fault-tolerant preconditioned conjugate gradient algorithm with lossy checkpointing technique

Input: linear system matrix A , preconditioner M , and right-hand side vector b

Output: approximate solution x

```

1: Initialization: same as line 1 in Algorithm 1
2: for  $i = 0, 1, \dots$  do
3:   if  $((i > 0) \text{ and } (i\%ckpt\_intvl = 0))$  then
4:     Compress:  $x^{(i)}$  with lossy compressor
5:     Checkpoint:  $i$  and compressed  $x^{(i)}$ 
6:   end if
7:   if  $((i > 0) \text{ and } (\text{recover}))$  then
8:     Recover:  $A, M, i$  and compressed  $x^{(i)}$ 
9:     Decompress:  $x^{(i)}$  with lossy compressor
10:    Compute  $r^{(i)} = b - Ax^{(i)}$ 
11:    Solve  $Mz^{(i)} = r^{(i)}$ 
12:     $p^{(i)} = z^{(i)}$ 
13:     $\rho_i = r^{(i)T} z^{(i)}$ 
14:   end if
15:   Computation: same as lines 10–17 in Algorithm 1
16: end for

```

after a recovery because of the errors introduced by lossy compression. Unfortunately, the convergence rate of the CG algorithm (which is superlinear) is highly dependent on these orthogonality relations. Hence, after a recovery from lossy checkpointing, the CG algorithm may lose the superlinear convergence rate, leading to a slow convergence [44]. To avoid this situation, we adopt a restarted scheme for the CG algorithm (*restarted CG*) [41], in which the computed approximate solution x_i is periodically treated as a new guess. In this case, we need to checkpoint only the vector x_i during the execution. The decompressed x_i is used as a new initial guess, and a new series of orthogonal vectors is reconstructed for the execution, such that a superlinear convergence rate can be rebuilt after restart.

Some studies of iterative methods have proved that such a restarted version of iterative methods (i.e., restarting by treating the current approximate solution vector as a new initial guess) may bring important advantages [4, 41, 43]. On the one hand, it suffers from less time and space complexity compared to their classic counterparts. For example, in practice, GMRES is often used to be restarted every a number of iterations (denoted by k) with the vector x_k as a new initial guess; and it is often denoted by GMRES(k). Without the periodically restarting feature, the total time and space complexity of GMRES will both grow with an increasing rate of N^2 over the time step N . By contrast, the time and space complexity of GMRES(k) will be limited under a constant cost over the execution. On the other hand, some studies [4, 43] have proved that the *restarted scheme* may not delay the convergence but even accelerate it, in that the periodically refreshed settings may enable the convergence to jump out of local search of the solution. In Section 4.4, we present more details regarding CG and GMRES with lossy checkpointing. In the following discussion, we always use CG and GMRES to denote the restarted CG and GMRES, respectively, in the context of lossy checkpointing. For these restarted iterative

methods, the only dynamic variable we need to checkpoint is the approximate solution vector x . As shown in Section 3, however, even checkpointing one dynamic vector will still lead to a severe performance issue for current or future HPC systems.

Users can follow the below workflow to easily integrate our lossy checkpointing for iterative methods with existing HPC applications: (1) initialize application; (2) *register variables external to the solver to checkpoint*; (3) start application's computations/iterations; (4) enter the solver's library; (5) *register the solver's variables to checkpoint in the library*; (6) iterate the solver; (7) *save or restore the application and solver's variables*; (8) continue to iterate the solver; (9) exit the solver's library; (10) continue application's computations/iterations (if needed); (11) exit application. Specifically, users can use the APIs, Protect() and Snapshot(), provided by our lossy checkpointing library to register and save/restore variables, as discussed in (2), (5), and (7).

4.3 Performance Model of Lossy Checkpointing

In this subsection, we build a performance model for the lossy checkpointing scheme, which is fundamental for analyzing the lossy checkpointing performance theoretically. Based on this performance model, we further derive a sufficient condition, an upper bound of the extra number of iterations caused by lossy data (i.e., Equation (9)), for guaranteeing the performance improvement of the lossy checkpointing scheme. Building the performance model requires a few more parameters, as listed in Table 2.

Table 2: Notations used in the lossy checkpointing performance model

T_{comp}	Mean time of performing lossy compression
T_{decomp}	Mean time of performing lossy decompression
T_{ckp}^{trad}	Mean time of performing one traditional checkpoint
T_{ckp}^{lossy}	Mean time of performing a lossy checkpointing
$T_{overhead}^{lossyCR}$	Time overhead of performing lossy checkpoint/recovery
N'	Mean number of extra iterations caused by per lossy recovery

Since lossy compression introduces errors in the reconstructed dynamic variable(s), the solver may suffer from a delay to converge. Suppose one recovery will cause extra N' iterations to the convergence on average, then the total execution time can be rewritten as

$$T_t = NT_{it} + T_{ckp}^{lossy} \frac{N}{k} + \frac{T_t}{T_f} (N'T_{it} + T_{rc}^{lossy} + T_{rb}), \quad (6)$$

because lossy checkpointing needs to perform one decompression during each recovery, lossy checkpointing needs to perform one compression during each checkpoint, and each recovery will delay N' iterations on average. Note that T_{ckp}^{lossy} and T_{rc}^{lossy} include the compression time T_{comp} and decompression time T_{decomp} , respectively. According to [25, 34, 48], T_{comp} and T_{decomp} are usually stable for a fixed compression accuracy.

Although the checkpointing/restarting time may differ across various iterations because of different data sizes due to various compression ratios, most well-known iterative methods can converge quickly such that the value of each element in the approximate

solution changes little in the following execution. Hence, the checkpointing data and its size will not change dramatically after several initial iterations, and T_{ckp}^{lossy} and T_{rc}^{lossy} can be assumed to be independent of iterations without loss of generality.

Similar to Section 3, we can derive the expected execution time with lossy checkpointing as

$$T_t = \frac{NT_{it}}{1 - \sqrt{2\lambda T_{ckp}^{lossy}} - \lambda T_{rc}^{lossy} - \lambda N'T_{it}}$$

and the expected performance overhead of lossy checkpointing as

$$T_{overhead}^{lossyCR} = NT_{it} \cdot \frac{\sqrt{2\lambda T_{ckp}^{lossy}} + \lambda T_{rc}^{lossy} + \lambda N'T_{it}}{1 - \sqrt{2\lambda T_{ckp}^{lossy}} - \lambda T_{rc}^{lossy} - \lambda N'T_{it}}. \quad (7)$$

Similarly, we can use T_{ckp}^{lossy} to approximate T_{rc}^{lossy} and simplify the performance overhead formula to

$$T_{overhead}^{lossyCR} \approx NT_{it} \cdot \frac{\sqrt{2\lambda T_{ckp}^{lossy}} + \lambda T_{ckp}^{lossy} + \lambda N'T_{it}}{1 - \sqrt{2\lambda T_{ckp}^{lossy}} - \lambda T_{ckp}^{lossy} - \lambda N'T_{it}}. \quad (8)$$

Now, we can derive a sufficient condition for iterative methods such that the lossy checkpointing scheme with a lossy compressor is able to obtain a performance gain over the traditional checkpointing scheme without lossy compression techniques.

THEOREM 1. *Denote λ and T_{it} by the expected failure rate and expected execution time of an iteration, respectively. The lossy checkpointing scheme will improve the execution performance for an iterative method as long as the following inequality holds.*

$$N' \leq (f(T_{ckp}^{trad}, \lambda) - f(T_{ckp}^{lossy}, \lambda)) / (\lambda T_{it}), \quad (9)$$

where $f(t, \lambda) = \sqrt{2\lambda t} + \lambda t$

PROOF. To have the lossy checkpointing overhead be lower than that of traditional checkpointing, we make Equation (8) smaller than Equation (4):

$$\frac{\sqrt{2\lambda T_{ckp}^{lossy}} + \lambda T_{ckp}^{lossy} + \lambda N'T_{it}}{1 - \sqrt{2\lambda T_{ckp}^{lossy}} - \lambda T_{ckp}^{lossy} - \lambda N'T_{it}} \leq \frac{\sqrt{2\lambda T_{ckp}^{trad}} + \lambda T_{ckp}^{trad}}{1 - \sqrt{2\lambda T_{ckp}^{trad}} - \lambda T_{ckp}^{trad}}.$$

Further simplifying this inequality, we can get the following formula with respect to the upper bound of N' .

$$N' \leq \frac{(\sqrt{2\lambda T_{ckp}^{trad}} + \lambda T_{ckp}^{trad}) - (\sqrt{2\lambda T_{ckp}^{lossy}} + \lambda T_{ckp}^{lossy})}{\lambda T_{it}}$$

Rewriting this inequality with $f(t, \lambda) = \sqrt{2\lambda t} + \lambda t$ will lead to Equation (9). \square

We give an example to explain how to use Theorem 1 in practice. Based on our experiments running GMRES on the Bebop cluster with 2,048 cores, we noted that the lossy compression technique can reduce the checkpointing time T_{ckp} from 120 seconds to 25 seconds for GMRES with a checkpoint of about 80 GB data (details are presented later in Figure 5). Suppose the MTTI of a system is one hour (i.e., $\lambda = 1/3600$) and that the GMRES algorithm requires 5,875 iterations with a total of 7,160 seconds to converge. Then the

mean time of one iteration, namely, T_{it} , is about 1.2 seconds. We can derive the maximum acceptable number of extra iterations to be 500 based on Equation (9). Hence, using a lossy checkpointing scheme is worthwhile if one recovery (with compression error introduced by lossy checkpointing) causes extra 500 or fewer iterations (about 9% of total iterations) to converge,

4.4 Impact Analysis of Lossy Checkpointing on Iterative Methods

In this subsection, we analyze the impact of lossy checkpointing on iterative methods, including stationary iterative methods, GMRES, and CG. Based on our analysis, we conclude that our lossy checkpointing technique can be applied to most of the iterative methods in numerical linear algebra for reducing the fault tolerance overhead.

4.4.1 Stationary Iterative Methods. We analyze the impact of lossy checkpointing on the convergence of four representative iterative methods: Jacobi, Gauss-Seidel, successive overrelaxation, and symmetric successive overrelaxation. The stationary iterative methods can be expressed in the following simple form,

$$x^{(i)} = Gx^{(i-1)} + c,$$

where G and c are a constant matrix and a vector, respectively.

Let R denote the spectral radius of matrix G , which is the largest eigenvalue of the matrix G . The convergence rate of a stationary iterative method is determined by its value. Specifically, let x^* denote the exact solution of the linear system:

$$\|x^{(i)} - x^*\| \approx R^i \cdot \|x^{(0)} - x^*\|.$$

Since the initial guess $x^{(0)}$ could be any vector and it is set to 0 in general, we have

$$\|x^{(i)} - x^*\| \approx R^i \cdot \|x^*\|. \quad (10)$$

Suppose the stationary methods encounter a failure and restart at the t^{th} iteration, and we denote that the lossy compression introduces an error vector e to $x^{(t)}$ by following relative error bound without loss of generality. Here the relative error bound means $|x_i^{(t)} - x_i^{(t)}| \leq eb \cdot |x_i^{(t)}|$ for $1 \leq i \leq n$, where eb is the relative error bound, $x_i^{(t)}$ is the i th element of vector $x^{(t)}$, and n is the vector length. The computation will start from $x^{(t)} + e$ (denoted by $x'^{(t)}$) instead of $x^{(t)}$. We derive the following theorem to obtain the range of the **expected upper bound of the number of extra iterations** for the stationary iterative methods.

THEOREM 2. *Based on the convergence rate (Equation (10)), the expected upper bound of the number of extra iterations for the stationary iterative methods falls into the interval $[\frac{N+1}{2} - \log_R(R^{\frac{N+1}{2}} + eb), N - \log_R(R^N + eb)]$, where eb is a constant relative error bound and R and N remain the same definitions as in the earlier discussion.*

PROOF. Based on the definition of the relative error bound eb , we have $\|e\| \leq eb \cdot \|x^{(t)}\|$. Then, we can get

$$\begin{aligned} \|x'^{(t)} - x^*\| &= \|x^{(t)} + e - x^*\| \leq \|x^{(t)} - x^*\| + \|e\| \\ &\leq R^t \cdot \|x^*\| + eb \cdot \|x^{(t)}\|. \end{aligned} \quad (11)$$

After another m iterations from erroneous vector $x^{(t)}$, we have

$$\|x^{(t+m)} - x^*\| \approx R^m \cdot \|x^{(t)} - x^*\|.$$

Then, based on Equation (11), we derive the following inequality.

$$\|x^{(t+m)} - x^*\| \leq R^m (R^t \cdot \|x^*\| + eb \cdot \|x^{(t)}\|) \quad (12)$$

Let us consider how to choose an m to ensure $\|x^{(t+m)} - x^*\| \leq \|x^{(t)} - x^*\|$, so that the residual norm between the approximate solution and exact solution will return to the previous value after m steps. Based on Equation (12), if we assure

$$R^m (R^t \cdot \|x^*\| + eb \cdot \|x^{(t)}\|) \leq \|x^{(t)} - x^*\|, \quad (13)$$

then $\|x^{(t+m)} - x^*\| \leq \|x^{(t)} - x^*\|$ will hold. Also, based on Equation (10), $\|x^{(t)} - x^*\| \approx R^t \|x^*\|$, Equation (13) is equivalent to

$$R^m (R^t \cdot \|x^*\| + eb \cdot \|x^{(t)}\|) \leq R^t \|x^*\|.$$

Therefore, $m \geq \log_R \frac{R^t \cdot \|x^*\|}{(R^t \cdot \|x^*\| + eb \cdot \|x^{(t)}\|)}$.

Without loss of generality, $\|x^{(t)}\|$ is likely close to $\|x^*\|$ after running with a few initial iterations, so we have the following approximation:

$$\log_R \frac{R^t \cdot \|x^*\|}{(R^t \cdot \|x^*\| + eb \cdot \|x^{(t)}\|)} \approx \log_R \frac{R^t}{(R^t + eb)}.$$

As a result of these inequalities, as long as $m \geq \log_R \frac{R^t}{(R^t + eb)}$, we will assure $\|x^{(t+m)} - x^*\| \leq \|x^{(t)} - x^*\|$. In other words, the stationary iterative methods need to take extra $\log_R \frac{R^t}{(R^t + eb)}$ iterations *at most* for convergence to the same accuracy. To conclude, if the stationary methods restart at the t th iteration with relative error bound eb , the upper bound of extra iterations N' is $t - \log_R(R^t + eb)$.

We now can calculate the range of the expected upper bound of extra iterations for the stationary iterative methods based on the monotonicity and convexity of $t - \log_R(R^t + eb)$ and Jensen inequality. Because of space limitations, we omit the details here. \square

4.4.2 GMRES. The generalized minimum residual method proposed by Saad and Schultz [43] is a Krylov subspace method for solving a large, sparse linear system with no constraint on the coefficient system matrix, especially for solving nonsymmetric systems. The method minimizes the norm of residual vector over a Krylov subspace at every iteration. Considering the cost growth of GMRES, it is often executed with the restarting scheme. In the following discussion, we use GMRES and restarted GMRES interchangeably. Although GMRES has a good ability to resist silent data corruption [26], protecting GMRES against fail-stop errors still has to rely on a checkpointing technique.

Unlike the stationary iterative methods, analyzing the extra convergence steps for nonstationary methods is difficult in theory. However, we propose an adaptive scheme to determine the error bound for GMRES with lossy checkpointing as follows.

THEOREM 3. *For the GMRES method, after a restart with lossy checkpointing, the new residual norm is controlled close to or at least on the same order as the previous residual if the relative error bound eb is set to $O(\|r^{(t)}\|/\|b\|)$.*

PROOF. Similar to Equation (11), we have the following.

$$\begin{aligned} \|r^{(t)}\| &= \|b - Ax^{(t)}\| = \|b - Ax^{(t)} + A(x^{(t)} - x^{(t)})\| \\ &\leq \|r^{(t)}\| + \|Ae\| \leq \|r^{(t)}\| + eb \cdot \|Ax^{(t)}\| \\ &= \|r^{(t)}\| + eb \cdot \|b - r^{(t)}\| \leq (1 + eb)\|r^{(t)}\| + eb \cdot \|b\| \\ &\approx \|r^{(t)}\| + eb \cdot \|b\| \end{aligned} \quad (14)$$

If eb is set to $O(\|r^{(t)}\|/\|b\|)$, then $eb \cdot \|b\|$ is $O(\|r^{(t)}\|)$; hence, $\|r^{(t)}\| + eb \cdot \|b\|$ is $O(\|r^{(t)}\|)$, which means that the new residual norm $\|r^{(t)}\|$ will be of the same order as the previous residual norm $\|r^{(t)}\|$ based on Equation (14). \square

Thanks to error-bounded compressors such as SZ and ZFP, one can easily control the distortion of data within $eb \cdot \|x^{(t)}\|$. Theorem 3 indicates that the convergence rate of GMRES will not degrade if the distortion of lossy checkpointing data follows a relative error bound $\|r^{(t)}\|/\|b\|$, where t is the current iteration number.

Now we can get a reasonable expected number of extra iterations for GMRES. As presented in Langou et al.'s study [31], if it is the same order of residual norm with which the restarted GMRES forms a new approximate solution, GMRES will converge to the same accuracy with no delay or even exhibit an accelerated convergence sometimes. The key reason is that the GMRES is easy to stagnate in its practical execution. If a failure occurs during the stagnation, the alternated recovered data can form a new approximate solution with different spectral properties, which may help GMRES jump out of the stagnation. This phenomenon has been theoretically and empirically observed and proved by Langou et al. [31]. Considering such a feature, the restarted GMRES with our proposed lossy checkpointing can converge without any delay based on the compression error bound suggested by Theorem 3 with an ensured, controlled residual norm. Thus, we can set the expected N' of GMRES with lossy checkpointing to 0. As a result, our lossy checkpointing is highly suitable for the restarted GMRES.

4.4.3 Conjugate Gradient. The conjugate gradient method is usually used in non-restarted style and has a superlinear convergence rate. As discussed in Section 4.2, however, we adopt restarted CG with lossy checkpointing. After a restart, it has to re-establish a new Krylov subspace based on the new initial guess. In our case, the new initial guess is the recovered vector $x^{(t)}$ (decompressed by lossy compressor). This process can lead to a delay of convergence to some extent. Unlike the GMRES method, even if we can ensure that $eb \cdot \|x^{(t)}\|$ is the same order as $\|x^{(t)} - x^*\|$, shown in Equation (14), the extra convergence steps for CG exhibit a property of randomness. Thus, for the CG method, we turn from theoretical analysis to an empirical evaluation for N' .

For each experiment, we randomly select an iteration to compress the approximate solution vector, decompress it to continue the computations, and then count the number of extra iterations. We evaluate the average extra iterations with different relative error bounds, as shown in Figure 2. The figure shows that the errors introduced by lossy checkpointing may delay the convergence of CG method to a certain extent. Based on our evaluation, the average extra iterations varies from 10% to 25% with different error bounds.

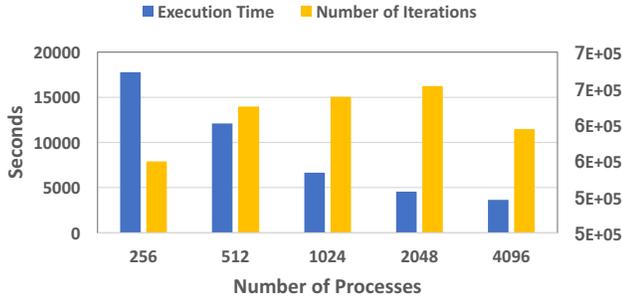


Figure 3: Average productive execution times and numbers of convergence iterations for solving matrix KKT240 once using GMRES and Jacobi preconditioner with different number of processes on Bebop.

a nonlinear programming problems for a 3D PDE-constrained optimization problem [47]. We refer readers to [46] for more details of the matrices. Note that we use GMRES for demonstration purpose, since it is much faster than Jacobi and CG cannot handle indefinite matrix. We test all the preconditioners listed in the PETSc’s website [2] and choose the best one (i.e., Jacobi preconditioner). We use the relative convergence tolerance of $1e-6$. Figure 3 shows that the average productive execution time for solving KKT240 once with GMRES needs to take more than one hour with 4,096 processes. Moreover, we note that the dimensions of the matrices collected by SuiteSparse grow exponentially with years [20]. Therefore, it will be more common to spend hours to days running iterative methods with a large number of ranks in parallel for very large-scale sparse linear systems; on the other hand, the mean time between failures for petascale supercomputers could be hourly or even less than one hour, as demonstrated by a recent study [36] based on a three-year statistic of Sunway TaihuLight supercomputer [27]. These results demonstrate that checkpointing during iterative methods is very important for the future HPC applications and exascale systems.

5.3 Theoretical Performance Investigation

We next perform the experiments with three checkpointing solutions under a fixed checkpoint frequency. The objective is to obtain the mean size and time of one checkpoint/recovery across different iterations from beginning to end for the three solutions. We set the checkpointing/recovering frequency to six times per hour and run each experiment for five times to ensure that the checkpoints/recoveries can cover the entire iteration. We calculate the average size and time of one checkpoint/recovery with different scales. We present the average checkpoint/recovery sizes for Jacobi, GMRES, and CG in Table 3. We present the average checkpoint/recovery time with different checkpointing solutions for Jacobi, GMRES, and CG in Figures 4, 5, and 6, respectively.

Table 3 illustrates that lossy compression can significantly reduce the checkpointing size compared with traditional and lossless checkpointing. Specifically, SZ lossy compression can reduce the checkpointing size to about $1/20 \sim 1/60$, whereas the lossless compression can achieve a compression ratio only up to about 6. Consequently, the checkpoint/recovery time can be reduced significantly for Jacobi, GMRES, and CG compared with the other two solutions,

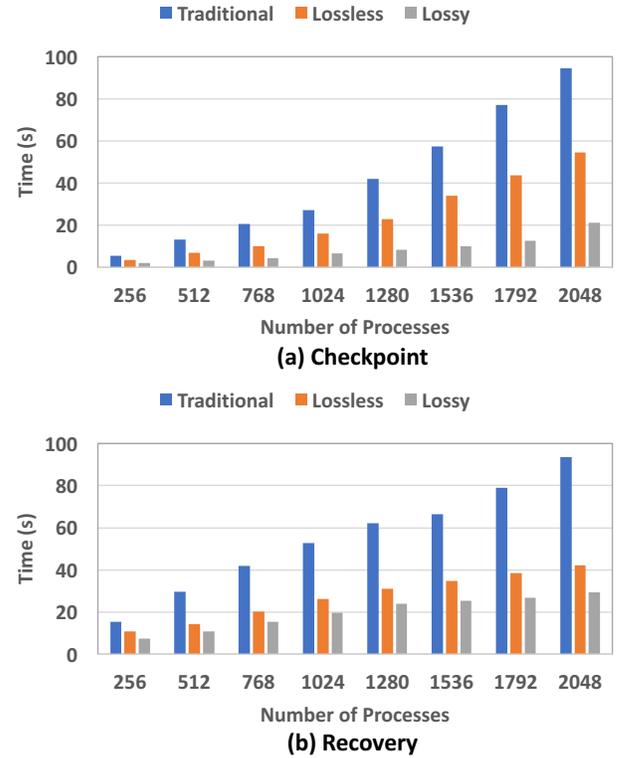


Figure 4: Average time of one checkpoint and recovery for Jacobi with different checkpointing techniques on Bebop.

as shown in Figure 4, 5, and 6. Comparing the three figures, we can see that the lossy checkpointing reduce checkpoint/recovery time more significantly for CG than for Jacob and GMRES. The reason is that the traditional and lossless checkpointing methods need to checkpoint/recover two vectors (x and p) for CG (as shown in line 4 in Algorithm 1) [16, 17], in that reinitializing p based x will lead to unknown delays (extra iterations). However, we have investigated the impact of lossy checkpoints on extra iterations of restarted CG, thus only the vector x needs to be checkpointed/recovered in our lossy checkpointing scheme. In addition, it is also observed that the checkpointing and recovery overhead both increase approximately linearly with the number of processes, because of linear increasing of the total checkpointing data size and the constant I/O bandwidth. In fact, such an I/O time increase is inevitable for any PFS considering the limited I/O bandwidth.

We include the compression and decompression time in the checkpoint/recovery time. The recovery time also contains the time to reconstruct static variables, including matrix A , preconditioner M , and right-hand side vector b . As shown in [48], because of no communication in parallel compression and decompression, the efficiency of parallel compression can stay at 90%, and the compression and decompression speed can reach 80 and 180 GB/s with 1,024 cores, respectively. Therefore, the compression and decompression take only a small portion of time in the checkpoint/recovery. Specifically, compressing and decompressing the 78.8 GB of checkpointing data take only about 0.5 seconds and 0.2 seconds, respectively.

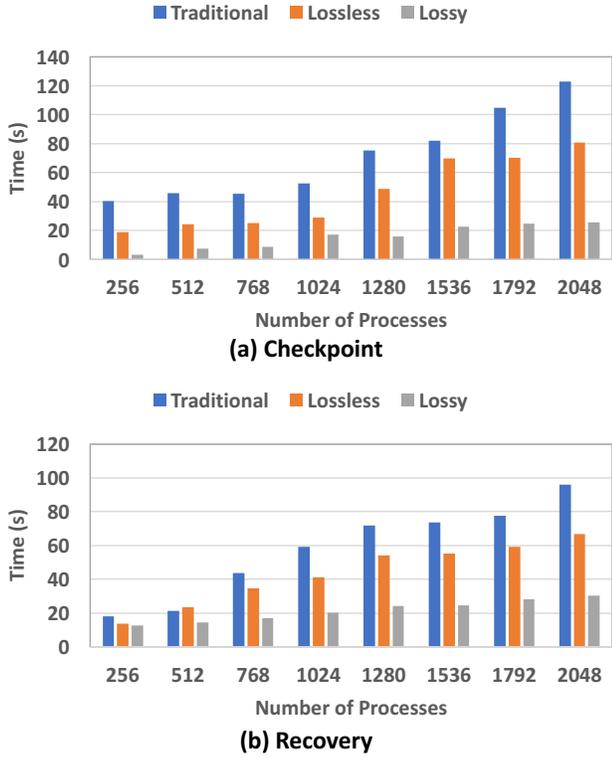


Figure 5: Average time of one checkpoint and recovery for GMRES with different checkpointing techniques on Bebop.

Such cost is nearly negligible compared with the average checkpoint/recovery time. Note that the time spent on I/O will increase with the number of processors, because of the inevitable bottleneck of the bandwidth when writing/reading data by many processes simultaneously (even with parallel I/O). By contrast, parallel compression/decompression time increases little with the number of processors, which means the performance gains by lossy checkpointing will increase with scales.

Based on the evaluated checkpointing time for different iterative methods (as shown in Figure 4, 5, and 6) and Equation (8), we can theoretically analyze the expected fault tolerance overhead for Jacobi, GMRES, and CG with two failure rates (i.e., MTTI = 1 hour and MTTI = 3 hours), as shown in Figure 7. Note that for the Jacobi method, the expectation of N' is about 6, which is calculated based on the interval $[\frac{N+1}{2} - \log_R(R^{\frac{N+1}{2}} + eb), N - \log_R(R^N + eb)]$, where $N = 3941$ and $eb = 10^{-4}$. We estimate the spectral radius R based on the final relative norm error and the number of convergence iterations. In our experiments, $R \approx 0.99998$. Following the discussion in Section 4.4, we set N' to 0 for GMRES and 594 for CG (i.e., 25% of the CG's total iterations).

Figure 7 illustrates that for both failure rates, the expected fault tolerance overhead of our proposed lossy checkpointing is always better than that of the other two solutions for Jacob and GMRES. For CG, the expected overhead of lossy checkpointing is better than that of the other two solutions when the number of processes is greater than 1536 and 768 for the two failure rates, respectively. We

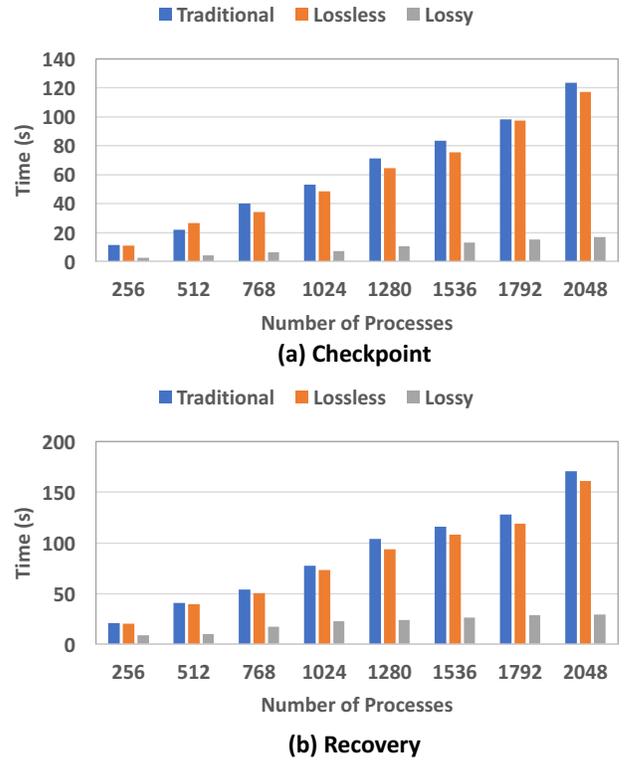


Figure 6: Average time of one checkpoint and recovery for CG with different checkpointing techniques on Bebop.

note that in Figure 7, the curves with lossy checkpointing increase much slowly than the curves with other two checkpointing solutions, thus demonstrating that our proposed lossy checkpointing is expected to achieve more performance gain as scale increases compared with the other two solutions. In the next subsection we will use the optimal checkpoint interval with given failure rate to experimentally prove this conclusion.

5.4 Experimental Evaluation

In this subsection, we evaluate the fault tolerance overhead experimentally for the three solutions with their corresponding optimal checkpointing intervals in the presence of injected failures. As described in Section 2, the MTTI can be almost hourly; hence, we inject failures with the rate being one failure per hour (i.e., $T_f = 3600$ seconds) in the experiment. Each failure may occur randomly at any time, including during computations of iterative methods and in the checkpoint/recovery period. The failure intervals follow an exponential distribution, because this is a common behavior of a system for most of its lifetime. According to Young's formula (as shown in Equation (1)), we can calculate the optimal checkpointing interval for the three solutions based on this failure rate and their checkpointing time as shown in Figure 5. Specifically, the calculated optimal checkpoint intervals for the traditional, lossless, and lossy checkpointing are 16 minutes, 12 minutes, and 7 minutes, respectively. We run each case with 2,048 processes/cores on Bebop ten times and investigate the average overall running time. The

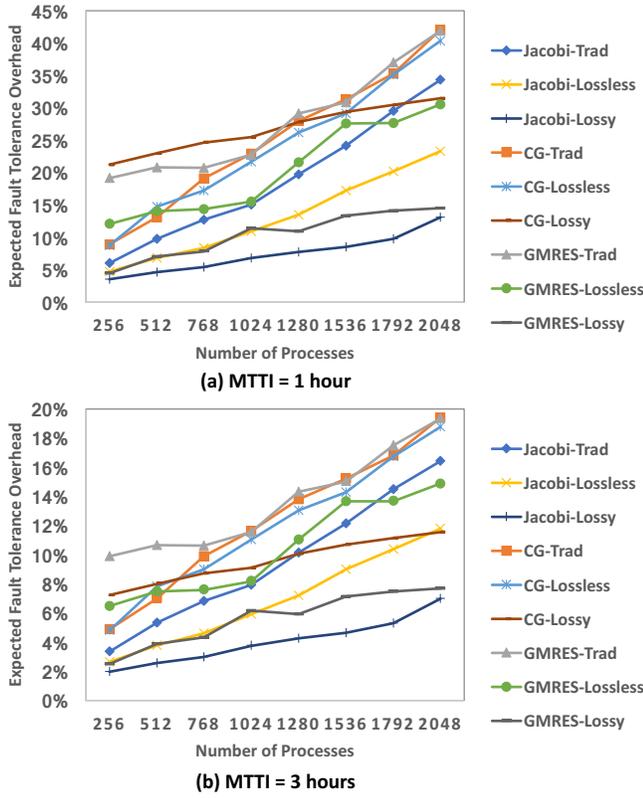


Figure 7: Expected overhead of fault-tolerant Jacobi, GMRES, & CG with different checkpointing techniques and failure rates on Bebop.

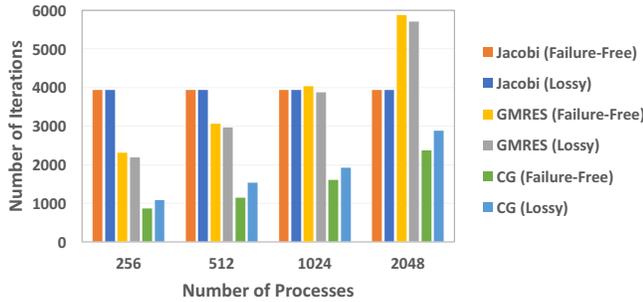


Figure 8: Number of convergence iterations with lossy checkpointing method for Jacobi, GMRES, and CG on Bebop.

baseline time of the iterative methods is the overall productive execution time of solving the 3D Poisson equation (as shown in Equation (15)) *once* without checkpointing and failure interruption. Specifically, the baseline times of Jacobi, GMRES, and CG are about 50 minutes, 120 minutes, and 35 minutes, respectively. We also compare the experimental overhead with the expected overhead derived theoretically by our performance model.

Figure 8 presents the numbers of convergence iterations with lossy checkpointing for Jacobi, GMRES, and CG compared with

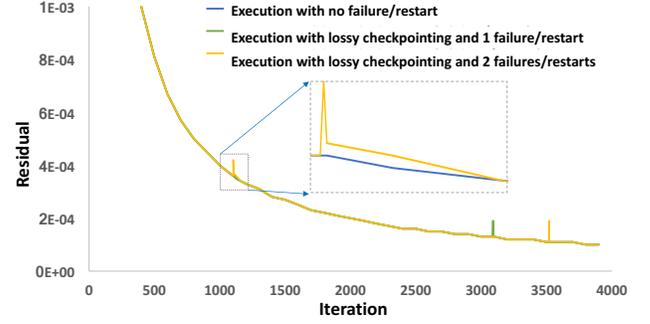


Figure 9: Typical example executions of Jacobi method with lossy checkpointing on Bebop.

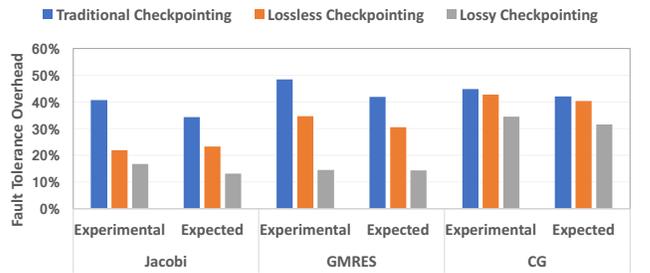


Figure 10: Experimental overhead versus expected overhead of fault-tolerant Jacobi, GMRES, and CG with different checkpointing techniques on Bebop.

their baseline executions (failure-free) on the Bebop cluster with 2,048 processes. The experiments illustrate that lossy checkpointing under our settings (including convergence accuracy, error bound and failure rate) introduces no delay (i.e., 0 extra iterations) on the convergence for Jacobi method, as shown in Figure 8. This is consistent with our theoretical analysis in Section 4.4.1: the upper bound of the number of extra iterations N' is 6, based on the interval $[\frac{N+1}{2} - \log_R(R^{\frac{N+1}{2}} + eb), N - \log_R(R^N + eb)]$ with $N = 3941$ and $eb = 10^{-4}$. Figure 8 shows that lossy checkpointing slightly accelerates the convergence of GMRES in the condition of bounding the jump of the residual (Theorem 3), which is consistent with the analysis and discussion presented in Section 4.4.2. Figure 8 also illustrates that the lossy checkpointing with $eb = 10^{-4}$ and $T_f = 3600$ seconds will delay the convergence of CG by 24.8% on average in terms of the convergence iterations, which is consistent with the analysis presented in Section 4.4.3 (as shown in Figure 2).

In Figure 9, we show two typical example executions of Jacobi method with lossy checkpointing. It shows that after a lossy recovery, Jacobi method can quickly converge to the same residual value as the failure-free Jacobi does, with no extra iterations.

In Figure 10 we present the average fault tolerance overhead of these three solutions with 2,048 processes on Bebop. Here the fault tolerance overhead refers to the overall running time taking away the baseline time. The figure illustrates that our proposed lossy checkpointing outperforms the other two solutions with Jacobi, GMRES, and CG. Specifically, for Jacobi, our solution reduces

the fault tolerance overhead by 59% compared with the traditional checkpointing and 24% compared with the lossless checkpointing. For GMRES, our solution outperforms the traditional checkpointing and the lossless checkpointing by 70% and 58%, respectively, in terms of the fault tolerance overhead. For CG, our solution reduce the fault tolerance overhead by 23% and 20% compared with the traditional and lossless checkpointing, respectively. Note that the experimental overheads for traditional and lossless checkpointing are higher than their expected overheads except for Jacobi with lossless checkpointing. The reason could be that the expected overhead formulas (Equations (4) and (8)) assume that the checkpointing time equals the recovery time, namely, $T_{rc} = T_{ckp}$. Except for Jacobi with lossless checkpointing, however, the other solutions' recovery time is higher than their checkpointing time, as shown in Figures 4, 5, and 6, because of reconstructing static variables. Hence, it will lead to a loss of accuracy between the experimental results and the expected analysis in terms of the fault tolerance overhead. Note that for our lossy checkpointing, there is only a small difference (up to about 10%) between the experimental overhead and the expected overhead, as shown in Figure 10.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel, efficient lossy checkpointing scheme, by exploring how to efficiently leverage the lossy compression technique to improve the overall checkpointing/restart performance for iterative methods in failure prone environment. We have four significant contributions: (1) we propose a lossy checkpointing scheme that can significantly improve the checkpointing performance for iterative methods; (2) we formulate the lossy checkpointing performance model and quantify the tradeoff between the reduced checkpointing overhead and the extra number of iterations caused by the compression errors; (3) we analyze the impact of the lossy checkpointing for multiple types of iterative methods; and (4) we evaluate the lossy checkpointing solution using a parallel environment with 2,048 cores. Our experiments show that our lossy checkpointing method can significantly reduce the fault tolerance overhead for the Jacobi, GMRES, and CG methods in the presence of failures, by 23%~70% compared with traditional checkpointing and 20%~58% compared with lossless compressed checkpointing. We plan to study how to extend our lossy checkpointing scheme to additional scientific application domains.

ACKNOWLEDGMENTS

This research was supported by the Exascale Computing Project (ECP), Project Number: 17-SC-20-SC, a collaborative effort of two DOE organizations – the Office of Science and the National Nuclear Security Administration, responsible for the planning and preparation of a capable exascale ecosystem, including software, applications, hardware, advanced system engineering and early testbed platforms, to support the nation's exascale computing imperative. The material was supported by the U.S. Department of Energy, Office of Science, under contract DE-AC02-06CH11357, and supported by the National Science Foundation under Grant No. 1305624, No. 1513201, and No. 1619253. We gratefully acknowledge the computing resources provided on Bebop, a high-performance computing cluster operated by the Laboratory Computing Resource Center at Argonne National Laboratory. We would like to thank Dr. Patrick Bridges for his helpful suggestions for the final paper.

REFERENCES

- [1] Bebop cluster. <https://www.lcr.anl.gov/systems/resources/bebop>, 2018. Online.
- [2] Summary of Sparse Linear Solvers Available from PETSc. <http://www.mcs.anl.gov/petsc/documentation/linearsolvvertable.html>, 2018. Online.
- [3] A. M. Agbaria and R. Friedman. Starfish: Fault-tolerant dynamic mpi programs on clusters of workstations. In *Proceedings of 8th International Symposium on High Performance Distributed Computing*, pages 167–176, 1999.
- [4] E. Agullo, L. Giraud, A. Guermouche, J. Roman, and M. Zounon. *Towards resilient parallel linear Krylov solvers: recover-restart strategies*. PhD thesis, INRIA, 2013.
- [5] W. Austin, G. Ballard, and T. G. Kolda. Parallel tensor compression for large-scale scientific data. In *2016 IEEE International Parallel and Distributed Processing Symposium*, pages 912–922, 2016.
- [6] J. M. Bahi, S. Contassot-Vivier, and R. Couturier. *Parallel iterative algorithms: from sequential to grid computing*. CRC Press, 2007.
- [7] S. Balay, S. Abhyankar, M. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. Gropp, D. Kaushik, et al. *Petsc users manual revision 3.8*. Technical report, Argonne National Lab.(ANL), Argonne, IL (United States), 2017.
- [8] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. Van der Vorst. *Templates for the solution of linear systems: building blocks for iterative methods*. SIAM, 1994.
- [9] R. Barrett, S. Borkar, S. Dossanjh, S. Hammond, M. Heroux, X. Hu, J. Luitjens, S. Parker, J. Shalf, and L. Tang. On the role of co-design in high performance computing. *Transition of HPC Towards Exascale Computing*, 24:141, 2013.
- [10] L. Bautista-Gomez, S. Tsuboi, D. Komatitsch, F. Cappello, N. Maruyama, and S. Matsuoka. Fti: High performance fault tolerance interface for hybrid systems. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, page 32, 2011.
- [11] U. Becciani, E. Sciacca, M. Bandieramonte, A. Vecchiato, B. Bucciarelli, and M. G. Lattanzi. Solving a very large-scale sparse linear system with a parallel algorithm in the gaia mission. In *High Performance Computing & Simulation (HPCCS), 2014 International Conference on*, pages 104–111. IEEE, 2014.
- [12] B. Bode, M. Butler, T. Dunning, W. Gropp, T. Hoefler, W.-m. Hwu, and W. Kramer. The Blue Waters super-system for super-science. *Contemporary High Performance Computing Architectures*, 2012.
- [13] P. G. Bridges, K. B. Ferreira, M. A. Heroux, and M. Hoemmen. Fault-tolerant linear solvers via selective reliability. pages 914–922, 2015.
- [14] J. Calhoun, F. Cappello, L. Olson, M. Snir, and W. Gropp. Exploring the feasibility of lossy compression for pde simulations. *The International Journal of High Performance Computing Applications*, 2018. To appear.
- [15] J. Chen, X. Liang, and Z. Chen. Online algorithm-based fault tolerance for cholesky decomposition on heterogeneous systems with gpus. In *2016 IEEE International Parallel and Distributed Processing Symposium*, pages 993–1002. IEEE, 2016.
- [16] Z. Chen. Algorithm-based recovery for iterative methods without checkpointing. In *Proceedings of the 20th International Symposium on High Performance Distributed Computing*, pages 73–84, 2011.
- [17] Z. Chen. Online-abft: An online algorithm based fault tolerance scheme for soft error detection in iterative methods. In *Proceedings of the 18th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, volume 48, pages 167–176, 2013.
- [18] Z. Chen, S. W. Son, W. Hendrix, A. Agrawal, W.-k. Liao, and A. Choudhary. Numarck: Machine learning algorithm for resiliency and checkpointing. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 733–744, 2014.
- [19] A. J. Chorin. Numerical solution of the navier-stokes equations. *Mathematics of computation*, 22(104):745–762, 1968.
- [20] T. A. Davis and Y. Hu. The university of florida sparse matrix collection. *ACM Transactions on Mathematical Software*, 38(1):1, 2011.
- [21] L. P. Deutsch. Gzip file format specification version 4.3. 1996.
- [22] S. Di, L. Bautista-Gomez, and F. Cappello. Optimization of a multilevel checkpoint model with uncertain execution scales. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 907–918, 2014.
- [23] S. Di, M. S. Bouguerra, L. Bautista-Gomez, and F. Cappello. Optimization of multilevel checkpoint model for large scale hpc applications. In *28th International Parallel and Distributed Processing Symposium*, pages 1181–1190, 2014.
- [24] S. Di and F. Cappello. Adaptive impact-driven detection of silent data corruption for hpc applications. *IEEE Transactions on Parallel and Distributed Systems*, 27(10):2809–2823, 2016.
- [25] S. Di and F. Cappello. Fast error-bounded lossy hpc data compression with sz. In *2016 IEEE International Parallel and Distributed Processing Symposium*, pages 730–739. IEEE, 2016.
- [26] J. Elliott, M. Hoemmen, and F. Mueller. Evaluating the impact of sdc on the gmres iterative solver. In *2014 IEEE 28th International Parallel and Distributed Processing Symposium*, pages 1193–1202, 2014.
- [27] H. Fu, J. Liao, J. Yang, L. Wang, Z. Song, X. Huang, C. Yang, W. Xue, F. Liu, F. Qiao, et al. The sunway taihulight supercomputer: system and applications. *Science*

- China Information Sciences*, 59(7):072001, 2016.
- [28] M. T. Heath. *Scientific computing*. McGraw-Hill New York, 2002.
- [29] T. Z. Islam, K. Mohror, S. Bagchi, A. Moody, B. R. De Supinski, and R. Eigenmann. MCrEngine: A scalable checkpointing system using data-aware aggregation and compression. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, page 17, 2012.
- [30] S. Lakshminarasimhan, N. Shah, S. Ethier, S.-H. Ku, C.-S. Chang, S. Klasky, R. Latham, R. Ross, and N. F. Samatova. Isabela for effective in situ compression of scientific data. *Concurrency and Computation: Practice and Experience*, 25(4):524–540, 2013.
- [31] J. Langou, Z. Chen, G. Bosilca, and J. Dongarra. Recovery patterns for iterative methods in a parallel unstable environment. *SIAM Journal on Scientific Computing*, 30(1):102–116, 2007.
- [32] G. Li, K. Pattabiraman, C.-Y. Cher, and P. Bose. An applicationspecific checkpointing technique for minimizing checkpoint corruption. In *International Symposium on Software Reliability Engineering*. IEEE, 2015.
- [33] X. Liang, J. Chen, D. Tao, S. Li, P. Wu, H. Li, K. Ouyang, Y. Liu, F. Song, and Z. Chen. Correcting soft errors online in fast fourier transform. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, page 30. ACM, 2017.
- [34] P. Lindstrom. Fixed-rate compressed floating-point arrays. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2674–2683, 2014.
- [35] P. Lindstrom and M. Isenburg. Fast and efficient compression of floating-point data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1245–1250, 2006.
- [36] R.-T. Liu and Z.-N. Chen. A large-scale study of failures on petascale supercomputers. *Journal of Computer Science and Technology*, 33(1):24–41, 2018.
- [37] A. Moody, G. Bronevetsky, K. Mohror, and B. R. d. Supinski. Design, modeling, and evaluation of a scalable multi-level checkpointing system. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–11, 2010.
- [38] J. Mora Acosta. *Numerical algorithms for three dimensional computational fluid dynamic problems*. Universitat Politècnica de Catalunya, 2001.
- [39] S. Patankar. *Numerical heat transfer and fluid flow*. CRC press, 1980.
- [40] J. S. Plank, K. Li, and M. A. Puening. Diskless checkpointing. *IEEE Transactions on Parallel and Distributed Systems*, 9(10):972–986, 1998.
- [41] M. J. D. Powell. Restart procedures for the conjugate gradient method. *Mathematical programming*, 12(1):241–254, 1977.
- [42] P. Ratanaworabhan, J. Ke, and M. Burtcher. Fast lossless compression of scientific floating-point data. In *2006 Data Compression Conference.*, pages 133–142, 2006.
- [43] Y. Saad and M. H. Schultz. Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.
- [44] P. Sao and R. Vuduc. Self-stabilizing iterative solvers. In *Proceedings of the Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems*, page 4, 2013.
- [45] N. Sasaki, K. Sato, T. Endo, and S. Matsuoka. Exploration of lossy compression for application-level checkpoint/restart. In *2015 IEEE International Parallel and Distributed Processing Symposium*, pages 914–922, 2015.
- [46] O. Schenk. Symmetric indefinite KKT matrices. <https://sparse.tamu.edu/Schenk>, 2018. Online.
- [47] O. Schenk, A. Wächter, and M. Weiser. Inertia-revealing preconditioning for large-scale nonconvex constrained optimization. *SIAM Journal on Scientific Computing*, 31(2):939–960, 2008.
- [48] D. Tao, S. Di, Z. Chen, and F. Cappello. Significantly improving lossy compression for scientific data sets based on multidimensional prediction and error-controlled quantization. In *2017 IEEE International Parallel and Distributed Processing Symposium*, pages 1129–1139. IEEE, 2017.
- [49] D. Tao, S. Di, X. Liang, Z. Chen, and F. Cappello. Lossy Checkpointing Library. <https://github.com/dingwentao/fti-sz>, 2018. Online.
- [50] D. Tao, S. L. Song, S. Krishnamoorthy, P. Wu, X. Liang, E. Z. Zhang, D. Kerbyson, and Z. Chen. New-sum: A novel online abft scheme for general iterative methods. In *Proceedings of the 25th ACM International Symposium on High-Performance Parallel and Distributed Computing*, pages 43–55, 2016.
- [51] R. Thakur, W. Gropp, and E. Lusk. On implementing mpi-io portably and with high performance. In *Proceedings of the Sixth Workshop on I/O in Parallel and Distributed Systems*, pages 23–32. ACM, 1999.
- [52] P. Wu and Z. Chen. Ft-scalapack: Correcting soft errors on-line for scalapack cholesky, qr, and lu factorization routines. In *Proceedings of the 23rd International Symposium on High-Performance Parallel and Distributed Computing*, pages 49–60. ACM, 2014.
- [53] P. Wu, N. DeBardeleben, Q. Guan, S. Blanchard, J. Chen, D. Tao, X. Liang, K. Ouyang, and Z. Chen. Silent data corruption resilient two-sided matrix factorizations. In *Proceedings of the 22nd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pages 415–427. ACM, 2017.
- [54] P. Wu, Q. Guan, N. DeBardeleben, S. Blanchard, D. Tao, X. Liang, J. Chen, and Z. Chen. Towards practical algorithm based fault tolerance in dense linear algebra. In *Proceedings of the 25th ACM International Symposium on High-Performance Parallel and Distributed Computing*, pages 31–42. ACM, 2016.
- [55] J. W. Young. A first order approximation to the optimum checkpoint interval. *Communications of the ACM*, 17(9):530–531, 1974.