

# Algorithm-Based Fault Tolerance for Dense Matrix Factorizations, Multiple Failures and Accuracy

AURELIEN BOUTEILLER, THOMAS HERAULT, GEORGE BOSILCA, PENG DU,  
and JACK DONGARRA, Innovative Computing Laboratory, University of Tennessee, Knoxville

10

Dense matrix factorizations, such as LU, Cholesky and QR, are widely used for scientific applications that require solving systems of linear equations, eigenvalues and linear least squares problems. Such computations are normally carried out on supercomputers, whose ever-growing scale induces a fast decline of the Mean Time To Failure (MTTF). This article proposes a new hybrid approach, based on Algorithm-Based Fault Tolerance (ABFT), to help matrix factorizations algorithms survive fail-stop failures. We consider extreme conditions, such as the absence of any reliable node and the possibility of losing both data and checksum from a single failure. We will present a generic solution for protecting the right factor, where the updates are applied, of all above mentioned factorizations. For the left factor, where the panel has been applied, we propose a scalable checkpointing algorithm. This algorithm features high degree of checkpointing parallelism and cooperatively utilizes the checksum storage leftover from the right factor protection. The fault-tolerant algorithms derived from this hybrid solution is applicable to a wide range of dense matrix factorizations, with minor modifications. Theoretical analysis shows that the fault tolerance overhead decreases inversely to the scaling in the number of computing units and the problem size. Experimental results of LU and QR factorization on the Kraken (Cray XT5) supercomputer validate the theoretical evaluation and confirm negligible overhead, with- and without-errors. Applicability to tolerate multiple failures and accuracy after multiple recovery is also considered.

Categories and Subject Descriptors: G.4 [Mathematical Software]: Reliability and robustness

General Terms: Algorithms

Additional Key Words and Phrases: ABFT, fault-tolerance, high performance computing, linear algebra

## ACM Reference Format:

Bouteiller, A., Herauld, T., Bosilca, G., Du, P., and Dongarra, J. 2014. Algorithm-based fault tolerance for dense matrix factorizations, multiple failures and accuracy. *ACM Trans. Parallel Comput.* 1, 2, Article 10 (January 2015), 28 pages.

DOI: <http://dx.doi.org/10.1145/2686892>

## 1. INTRODUCTION

Today's high-performance computers have paced into The Petaflops realm, through the increase of system scale. The number of system components, such as CPU cores, memory, networking, and storage grow considerably. One of the most powerful Petaflop scale machines, Kraken, from National Institute for Computational Sciences and University of Tennessee, harnessed as many as 112,800 cores to reach its peak performance of 1.17 Petaflops to rank No.11 on the November 2011 Top500 list.<sup>1</sup> Even

<sup>1</sup><http://www.top500.org>

---

This article is authored by employees of the United States Government and is in the public domain. Non-exclusive copying or redistribution is allowed, provided that the article citation is given and the authors and agency are clearly identified by its source.

Authors' address: A. Bouteiller, T. Herauld, G. Bosilca (corresponding author); email: [bosilca@icl.utk.edu](mailto:bosilca@icl.utk.edu), P. Du, and J. Dongarra, Innovative Computing Laboratory, University of Tennessee, Knoxville.

This paper is authored by an employee(s) of the United States Government and is in the public domain. Non-exclusive copying or redistribution is allowed, provided that the article citation is given and the authors and agency are clearly identified as its source.

© 2015 ACM 2329-4949/2015/01-ART10 \$15.00

DOI: <http://dx.doi.org/10.1145/2686892>

using accelerators, as in Titan, the current No.1, more than 18,688 accelerator boards and supporting host processors are still necessary. With the increase of system scale and chip density, the reliability and availability of such systems has declined. It has been shown that, under specific circumstances, adding computing units might hamper applications completion time, as a larger node count implies a higher probability of reliability issues. This directly translates into a lower efficiency of the machine, which equates to a lower scientific throughput [Streitz et al. 2006]. It is estimated that the MTTF of High Performance Computing (HPC) systems might drop to about one hour in the near future [Cappello 2009]. Without a drastic change at the algorithmic level, such a failure rate will certainly prevent capability applications from progressing.

Exploring techniques for creating a software ecosystem and programming environment capable of delivering computation at extreme scale, which are both resilient and efficient, will eliminate a major obstacle to scientific productivity on tomorrow's HPC platforms. In this work we advocate that in extreme scale environments, successful approaches to fault tolerance (e.g., those which exhibit acceptable recovery times and memory requirements) must go beyond traditional systems-oriented techniques and leverage intimate knowledge of dominant application algorithms, in order to create a middleware that is far more adapted and responsive to the application's performance and error characteristics.

While many types of failures can strike a distributed system [Schroeder and Gibson 2007], the focus of this article is on the most common representation: the fail-stop model. In this model, a failure is defined as a process that *completely* and *definitely* stops responding, triggering the loss of a critical part of the global application state. To be more realistic, we assume a failure could occur at any moment and can affect any part of the application's data. We introduce a new generic hybrid approach based on algorithm-based fault tolerance (ABFT) that can be applied to several ubiquitous one-sided dense linear factorizations. Using one of these factorizations, namely LU with partial pivoting, which is significantly more challenging due to pivoting, we theoretically prove that this scheme successfully applies to the three well known one-sided factorizations, Cholesky, LU and QR. To validate these claims, we implement and evaluate this generic ABFT scheme with both the LU and QR factorizations. A significant contribution of this work (which is an extension of Du et al. [2012]) is to protect the part of the matrix below the diagonal (referred to as "the left factor" in the rest of the text) during the factorization, which was hitherto never achieved. The extended content covers the case of multiple *simultaneous* failures and presents an effective protection scheme that uses  $2f$  checksum blocks to protect against  $f$  simultaneous failures. Additionally, the resulting accuracy after multiple recoveries during the lifetime of the application is experimentally assessed.

The rest of the article is organized as follows. Section 2 presents background and prior work in the domain; Section 3 reviews the features of full factorizations. Section 4 discusses the protection of the right factor using the ABFT method. Section 5 reviews the idea of vertical checkpointing and proposes the new checkpointing method to protect the left factor. Section 6 discusses an effective checksumming strategy to recover from multiple simultaneous failures. Section 7 evaluates the performance, overhead and accuracy of the proposed algorithm using the example of LU and QR, and Section 8 concludes the work.

## 2. ALGORITHM-BASED FAULT TOLERANCE BACKGROUND

The most well-known fault-tolerance technique for parallel applications, checkpoint-restart (C/R), encompasses two categories, the system and application level. At the system level, message passing middleware deals with faults automatically, without intervention from the application developer or user [Bouteiller et al. 2010; Burns et al.

1994]. At the application level, the application state is dumped to a reliable storage when the application code mandates it. Even though C/R bears the disadvantage of high overhead while writing data to stable storage, it is widely used nowadays by high end systems [Katz et al. 2009]. To reduce the overhead of C/R, diskless checkpointing [Lu 2005; Plank et al. 1998] has been introduced to store checksum in memory rather than stable storage. While diskless checkpointing has shown promising performance in some applications (for instance, FFT in Elnozahy et al. [1992]), it exhibits large overheads for applications modifying substantial memory regions between checkpoints [Plank et al. 1998], as is the case with factorizations.

In contrast, Algorithm Based Fault Tolerance (ABFT) is based on adapting the algorithm so that the application dataset can be recovered at any moment, without involving costly checkpoints. ABFT was first introduced to deal with silent error in systolic arrays [Huang and Abraham 1984]. Unlike other methods that treat the recovery data and computing data separately, ABFT approaches are based on the idea of maintaining consistency of the recovery data, by applying appropriate mathematical operations on both the original and recovery data. Typically, for linear algebra operations, the input matrix is extended with supplementary columns and/or rows containing checksums. This initial encoding happens only once; the matrix algorithms are designed to work on the encoded checksum along with matrix data, similar mathematical operations are applied to both the data and the checksum so that the checksum relationship is kept invariant during the course of the algorithm. Should some data be damaged by failures, it is then possible to recover the application by inverting the checksum operation to recreate missing data. The overhead of ABFT is usually low, since no periodical global checkpoint or rollback-recovery is involved during computation and the computation complexity of the checksum operations scales similarly to the related matrix operation (and the ratio of extra computation is small and asymptotically tends toward zero). ABFT and diskless checkpointing have been combined to apply to basic matrix operations like matrix-matrix multiplication [Bosilca et al. 2009; Chen and Dongarra 2006a, 2006b, 2008] and have been implemented on algorithms similar to those of ScaLAPACK [Choi et al. 1996], which is widely used for dense matrix operations on parallel distributed memory systems.

Recently, ABFT has been applied to the High Performance Linpack (HPL) [Davies et al. 2011] and to the Cholesky factorization [Hakkarinen and Chen 2010]. Both Cholesky and HPL have the same factorization structure, where only half of the factorization result is required, and the update to the trailing matrix is based on the fact that the left factor result is a triangular matrix. This approach, however, does not necessarily apply to other factorizations, like QR where the left factor matrix is full, nor when the application requires both the left and right factorization results. Also, LU with partial pivoting, when applied to the lower triangular  $L$ , potentially changes the checksum relation and renders basic checkpointing approaches useless, when checksums are computed along the intuitive column-wise direction.

The generic ABFT framework for matrix factorizations we introduce in this work can be applied not only to Cholesky and HPL, but also to LU and QR. The right factor is protected by a traditional ABFT checksum, while the left factor is protected by a novel vertical checkpointing scheme, making the resulting approach an hybrid between ABFT and algorithm driven checkpointing. Indeed, this checkpointing algorithm harnesses some of the properties of the factorization algorithm to exchange limited amount of rollback with the ability to overlap the checkpointing of several panel operations running in parallel. Other contributions of this work include correctness proofs and overhead characterization for the ABFT approach on the most popular 2D-block cyclic distribution (as opposed to the 1D distributions used in previous works). These proofs consider the effect of failures during critical phases of the

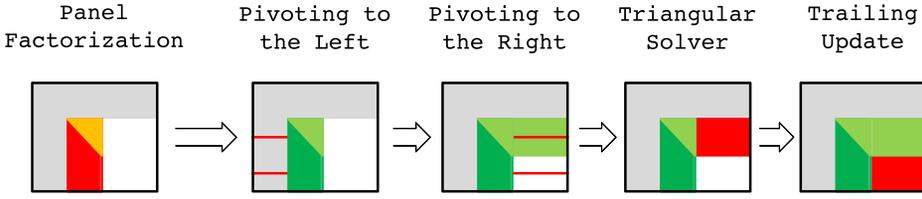


Fig. 1. Steps applied to the input matrix in an iteration of the LU factorization; Green: Just finished; Red & Orange: being processed; Gray: Finished in previous iterations.

algorithm, and demonstrate that recovery is possible without suffering from error propagation.

### 3. FULL FACTORIZATIONS OF MATRIX

In this work, we consider the case of factorizations where the lower triangular part of the factorization result matters, as is the case in QR and LU with pivoting. For example, the left factor  $Q$  is required when using  $QR$  to solve the least square problem, and so is  $L$  when solving  $A^k x = b$  with the “LU factorization outside the loop” method [Golub and Van Loan 1996]. In the remaining of this section, we recall the main algorithm of the most complex case of one-sided factorization, block LU with pivoting. Additionally, we highlight challenges specific to this type of algorithms, when compared to algorithms studied in previous works.

Figure 1 presents the diagram of the basic operations applied to the input matrix to perform the factorization. The block LU factorization algorithm can be seen as a recursive process. At each iteration, the panel factorization is applied on a block column. This panel operation factorizes the upper square (selecting adequate pivots and applying internal row swapping as necessary to ensure numerical stability), and scales the lower polygon accordingly. The output of this panel is used to apply row swapping to the result of previous iterations, on the left, and to the trailing matrix on the right. The triangular solver is applied to the right of the factored block to scale it accordingly, and then the trailing matrix is updated by applying a matrix-matrix multiply update. Then the trailing matrix is used as the target for the next iteration of the recursive algorithm, until the trailing matrix is empty. Technically, each of these basic steps is usually performed by applying a parallel Basic Linear Algebra Subroutine (PBLAS).

The structure of the other one-sided factorizations, Cholesky and QR, are similar with minor differences. Both QR and Cholesky can maintain adequate numerical stability without pivoting (and according swapping operations). The structure of the block QR and LU with pivoting factorization is identical, only the computational kernels differ. The structure of the block Cholesky factorization is similar, with the notable difference that the trailing matrix update involves only the upper right triangle (the lower left triangle is symmetric and is therefore omitted to spare space and computation). In the case of LU and QR, however, the upper left and lower right factors are distinct. There are a significant number of applications, like iterative refinement and algorithms for eigenvalue problems, where the entire factorization result is needed, including the lower left factor, henceforth commanding adequate protection for this part of the result as well.

### 4. PROTECTION OF THE RIGHT FACTOR MATRIX WITH ABFT

Our focus is on providing an algorithm that protects both the upper right and the lower left factors, as the factorization unfolds. In this section, we detail the ABFT approach that is used to protect the upper triangle from failures, while considering the intricacies of typical block cyclic distributions and failure detection delays.

#### 4.1. Checksum Relationship

ABFT approaches are based upon the principle of keeping an invariant bijective relationship between protective supplementary blocks and the original data through the execution of the algorithm, by the application of numerical updates to the checksum. In order to use ABFT for matrix factorization, an initial checksum is generated before the actual computation starts. In future references we use  $G$  to refer to the generator matrix, and  $A$  to the original input matrix. The checksum  $C$  for  $A$  is produced by

$$C = GA \text{ or } C = AG. \quad (1)$$

When  $G$  is all-1 vector, the checksum is simply the sum of all data items from a certain row or column. Referred to as the *checksum relationship*, (1) can be used at any step of the computation for checking data integrity (by detecting mismatching checksum and data) and recovery (inverting the relation builds the difference between the original and the degraded dataset). With the type of failures we consider (Fail-Stop), data cannot be corrupted, so we will use this relationship to implement the recovery mechanism only. This relationship has been shown separately for Cholesky [Hakkarinen and Chen 2010], and HPL [Davies et al. 2011], both sharing the property of updating the trailing matrix with a lower triangular matrix. However, in this work we consider the general case of matrix factorization algorithms, including those where the full matrix is used for trailing matrix updates (as is the case for QR and LU with partial pivoting). In this context, the invariant property has not been demonstrated; we will now demonstrate that it holds for full matrix based updates algorithms as well.

#### 4.2. Checksum Invariant with Full Matrix Update

In Luk and Park [1988],  $ZU$  is used to represent a matrix factorization (optionally with pairwise pivoting for LU), where  $Z$  is the left matrix (lower triangular in the case of Cholesky or full for LU and QR) and  $U$  is an upper triangular matrix. The factorization is then regarded as the process of applying a series of matrices  $Z_i$  to  $A$  from the left until  $Z_i Z_{i-1} \cdots Z_1 A$  becomes upper triangular.

**THEOREM 4.1.** *Checksum relationship established before  $ZU$  factorization is maintained during and after factorization.*

**PROOF.** Suppose data matrix  $A \in \mathbb{R}^{n \times n}$  is to be factored as  $A = ZU$ , where  $Z$  and  $U \in \mathbb{R}^{n \times n}$  and  $U$  is an upper triangular matrix.  $A$  is checkpointed using generator matrix  $G \in \mathbb{R}^{n \times nc}$ , where  $nc$  is the width of checksum. To factor  $A$  into upper triangular form, a series of transformation matrices  $Z_i$  is applied to  $A$  (with partial pivoting in LU).

*Case 1: No Pivoting.*

$$U = Z_n Z_{n-1} \dots Z_1 A. \quad (2)$$

Now the same operation is applied to  $A_c = [A, AG]$

$$\begin{aligned} U_c &= Z_n Z_{n-1} \dots Z_1 [A, AG] \\ &= [Z_n Z_{n-1} \dots Z_1 A, Z_n Z_{n-1} \dots Z_1 AG] \\ &= [U, UG]. \end{aligned} \quad (3)$$

For any  $k \leq n$ , using  $U^k$  to represent the result of  $U$  at step  $k$ ,

$$\begin{aligned} U_c^k &= Z_k Z_{k-1} \dots Z_1 [A, AG] \\ &= [Z_k Z_{k-1} \dots Z_1 A, Z_k Z_{k-1} \dots Z_1 AG] \\ &= [U^k, U^k G]. \end{aligned} \quad (4)$$

*Case 2:* With partial pivoting.

$$\begin{aligned} U_c^k &= Z_k P_k Z_{k-1} P_{k-1} \dots Z_1 P_1 [A, AG] \\ &= [Z_k P_k Z_{k-1} P_{k-1} \dots Z_1 P_1 A, \\ &\quad Z_k P_k Z_{k-1} P_{k-1} \dots Z_1 P_1 AG] \\ &= [U^k, U^k G]. \end{aligned} \quad (5)$$

Therefore the checksum relationship holds for LU with partial pivoting, Cholesky and QR factorizations.  $\square$

### 4.3. Checksum Invariant in Block Algorithms

Theorem 4.1 shows the mathematical checksum relationship in matrix factorizations. However, in production, HPC factorizations are performed in block algorithms, and execution is carried out in a recursive way. Linear algebra packages, like ScaLAPACK, consist of several function components for each factorization. For instance, LU has a panel factorization, a triangular solver and a matrix-matrix multiplication. We need to ensure that the checksum relationship also holds for block algorithms, both at the end of each iteration, and after the factorization is completed.

**THEOREM 4.2.** *For ZU factorization in block algorithm, checksum at the end of each iteration only covers the upper triangular part of data that has already been factored and are still being factored in the trailing matrix.*

**PROOF.** Input Matrix  $A$  is split into blocks of data of size  $nb \times nb$  ( $A_{ij}, Z_{ij}, U_{ij}$ ), and the following stands:

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{bmatrix} = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \end{bmatrix}, \quad (6)$$

where  $A_{13} = A_{11} + A_{12}$ , and  $A_{23} = A_{21} + A_{22}$ .

Since  $A_{13} = Z_{11}U_{13} + Z_{12}U_{23}$ , and  $A_{23} = Z_{21}U_{13} + Z_{22}U_{23}$ , and using the relation

$$\begin{cases} A_{11} = Z_{11}U_{11} \\ A_{12} = Z_{11}U_{12} + Z_{12}U_{22} \\ A_{21} = Z_{21}U_{11} \\ A_{22} = Z_{21}U_{12} + Z_{22}U_{22} \end{cases}$$

in (6), we have the following system of equations:

$$\begin{cases} Z_{21}(U_{11} + U_{12} - U_{13}) = Z_{22}(U_{23} - U_{22}) \\ Z_{11}(U_{11} + U_{12} - U_{13}) = Z_{12}(U_{23} - U_{22}). \end{cases}$$

This can be written as:

$$\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix} \begin{bmatrix} U_{11} + U_{12} - U_{13} \\ -(U_{23} - U_{22}) \end{bmatrix} = 0.$$

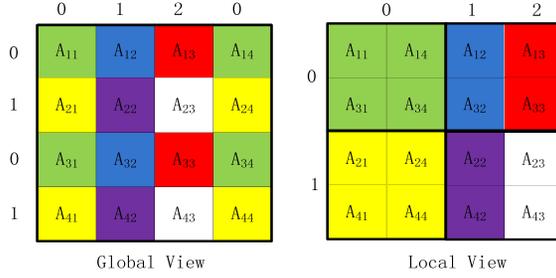


Fig. 2. Example of a 2D block-cyclic data distribution.

For LU, Cholesky and QR,  $\begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22} \end{bmatrix}$  is always nonsingular, so  $\begin{bmatrix} U_{11} + U_{12} - U_{13} \\ U_{23} - U_{22} \end{bmatrix} = 0$ , and  $\begin{cases} U_{11} + U_{12} = U_{13} \\ U_{23} = U_{22} \end{cases}$ .

This shows that after  $ZU$  factorization, checksum blocks cover the upper triangular matrix  $U$  only, even for the diagonal blocks. At the end of each iteration, for example the first iteration in (6),  $Z_{11}$ ,  $U_{11}$ ,  $Z_{21}$  and  $U_{12}$  are completed, and  $U_{13}$  is already  $U_{11} + U_{12}$ . The trailing matrix  $A_{22}$  is updated with

$$A_{22}' = A_{22} - Z_{21}U_{12} = Z_{22}U_{22},$$

and  $A_{23}$  is updated to

$$\begin{aligned} A_{23}' &= A_{23} - Z_{21}U_{13} \\ &= A_{21} + A_{22} - Z_{21}(U_{11} + U_{12}) \\ &= Z_{21}U_{11} + A_{22} - Z_{21}U_{11} - Z_{21}U_{12} \\ &= A_{22} - Z_{21}U_{12} = Z_{22}U_{22}. \end{aligned}$$

Therefore, at the end of each iteration, data blocks that have already been and are still being factored remain covered by checksum blocks.  $\square$

#### 4.4. Issues with Two-Dimensional Block-Cyclic Distribution

It has been well established that data layout plays an important role in the performance of parallel matrix operations on distributed memory systems [Choi et al. 1996; Kumar et al. 1994]. In 2D block-cyclic distributions, data is divided into equally sized blocks, and all computing units are organized into a virtual two-dimension grid  $P$  by  $Q$ . Each data block is distributed to computing units in round robin following the two dimensions of the virtual grid. Figure 2 is an example of a  $P = 2, Q = 3$  grid applied to a global matrix of  $4 \times 4$  blocks. The same color represents the same process while numbering in  $A_{ij}$  indicates the location in the global matrix. This layout helps with load balancing and reduces data communication frequency, because in each step of the algorithm, many computing units can be engaged in computations concurrently, and communications pertaining to blocks positioned on the same unit can be grouped. Thanks to these advantages, many prominent software libraries (like ScaLAPACK [Choi et al. 1996]) assume a 2D block-cyclic distribution.

However, with a 2D block-cyclic data distribution, the failure of a single process, usually a computing node which keeps several noncontiguous blocks of the matrix, results in holes scattered across the whole matrix. Figure 3 is an example of a  $5 \times 5$  blocks matrix (on the left) with a  $2 \times 3$  process grid. Red blocks represent holes caused

by the failure of the single process (1, 0). In the general case, these holes can impact both checksum and matrix data at the same time.

#### 4.5. Checksum Protection Against Failure

Our algorithm works under the assumption that any process can fail and therefore the data, including the checksum, can be lost. Rather than forcing checksum and data on different processes and assuming only one would be lost, as in Davies et al. [2011], we put checksum and data together in the process grid and design the checksum protection algorithm accordingly.

*4.5.1. Minimum Checksum Amount for Block Cyclic Distributions.* Theoretically, the sum-based checksum  $C$  of a series of  $N$  blocks  $A_i$ ,  $1 \leq i \leq N$ , where  $N$  is the total number of blocks in one row/column of the matrix, is computed by:

$$C = \sum_{i=1}^N A_i. \quad (7)$$

With the 2D block-cyclic distribution, a single failure punches multiple holes in the global matrix. With more than one hole per row/column,  $C$  in (7) is not sufficient to recover all lost data. A slightly more sophisticated checksum scheme is required.

**THEOREM 4.3.** *Using sum-based checkpointing, for  $N$  data items distributed in block-cyclic onto  $Q$  processes, the size of the checksum to recover from the loss of one process is  $\lceil \frac{N}{Q} \rceil$*

**PROOF.** With 2D block-cyclic, each process gets  $\lceil \frac{N}{Q} \rceil$  items. At the failure of one process, all data items in the group held by the process are lost. Take data item  $a_k$ ,  $1 \leq k \leq \lceil \frac{N}{Q} \rceil$ , from group  $j$ ,  $1 \leq j \leq Q$ . To be able to recover  $a_k$ , any data item in group  $j$  cannot be used, so at least one item from another group is required to create the checksum, and this generates one additional checksum item. Therefore for all items in group  $j$ ,  $\lceil \frac{N}{Q} \rceil$  checksum items are generated so that any item in group  $j$  can be recovered.  $\square$

Applying this theorem, we have the following checksum algorithm: Suppose  $Q$  processes are in a process column or row, and let each process have  $K$  blocks of data of size  $nb \times nb$ . Without loss of generality, let  $K$  be the largest number of blocks owned by any of the  $Q$  processes. From Theorem 4.3, the size of the checksum in this row is  $K$  blocks.

Let  $C_k$  be the  $k^{\text{th}}$  checksum item, and  $A_k^j$ , be the  $k^{\text{th}}$  data item on process  $j$ ,  $1 \leq k \leq \lceil \frac{N}{Q} \rceil$ ,  $1 \leq j \leq Q$ :

$$C_k = \sum_{j=1}^Q A_k^j. \quad (8)$$

Under (8), we have the following corollary.

**COROLLARY 4.4.** *The  $k^{\text{th}}$  block of checksum is calculated using the  $k^{\text{th}}$  block of data of each process having at least  $k$  blocks.*

*4.5.2. Checksum Duplicates.* Since ABFT checksum is stored by regular processors, it has to be considered as fragile as the matrix data. From Theorem 4.3 and using the same  $N$  and  $Q$ , the total number of checksum blocks is  $K = \lceil \frac{N}{Q} \rceil$ . These checksum

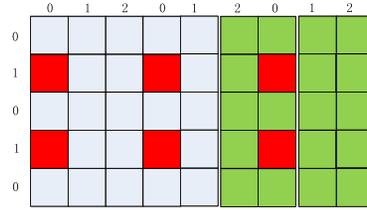


Fig. 3. Holes in a checksum protected matrix caused by a single failure and the naive checksum duplication protection scheme (3x2 process grid).

blocks can be appended to the bottom or to the right of the global data matrix accordingly, and since checksum is stored on computing processes, these  $K$  checksum blocks are distributed over  $\min(K, Q)$  processes (see Figure 3). If a failure strikes any of these processes, like (1, 0) in this example, some checksum is lost and cannot be recovered. Therefore, checksum itself needs protection; as a first approach, duplication is used to protect checksum from failure.

A straightforward way of performing duplication is to make a copy of the entire checksum block, as illustrated by the two rightmost columns in Figure 3. While simple to implement, this method suffers from two major defects. First, if the checksum width  $K$  is a multiple of  $Q$  (or  $P$  for column checksum), the duplicate of a checksum block is located on the same processor, defeating the purpose of duplication. This could be solved at the cost of introducing an extra empty column in the process grid to resolve the mapping conflict. More importantly, to maintain the checksum invariant property, it is required to apply the trailing matrix update on the checksum (and its duplicates) as well. From Corollary 4.4, once all the  $k^{th}$  block columns on each process have finished the panel factorization (in  $Q$  step), the  $k^{th}$  checksum block column is no longer active in any further computation (except pivoting) and should be excluded from the computing scope to reduce the ABFT overhead. This is problematic, as this results in updates applying to noncontiguous local matrices (containing now inactive checksum columns), and splitting the PBLAS calls, to avoid excluded columns, increases the communication volume and has a significant impact on the trailing matrix update efficiency.

**4.5.3. Reverse Neighboring Checksum Storage.** In the previous paragraph, we described the specific challenges posed by the maintenance of checksums during the factorization, hinting the crucial role of checksum storage location on efficiency. From this observation, we propose the following reverse neighboring checksum duplication method that allows for applying the update in a single PBLAS call without incurring extraneous computation.

Figure 4 is an example of the reverse neighboring checksum method on a  $2 \times 3$  grid. The data matrix has  $8 \times 8$  blocks and therefore the size of checksum is  $8 \times 3$  blocks with an extra  $8 \times 3$  blocks copy. The arrows indicate where checksum blocks are stored on the right of the data matrix, according to the reverse storage scheme. For example, in the LU factorization, the first 3 block columns produce the checksum in the last two block columns (hence making 2 duplicate copies of the checksum); the next 3 block columns then produce the next 2 rightmost checksum columns, etc. Because copies are stored in consecutive columns of the process grid, for any 2D grid with  $Q > 1$ , the checksum duplicates are guaranteed to be stored on different processors. The triangular solve (TRSM) and trailing matrix update (GEMM) are applied to the whole checksum area until the first three columns are factored. In following factorization steps, the two last block columns of checksum are excluded from the TRSM and GEMM scope. Since

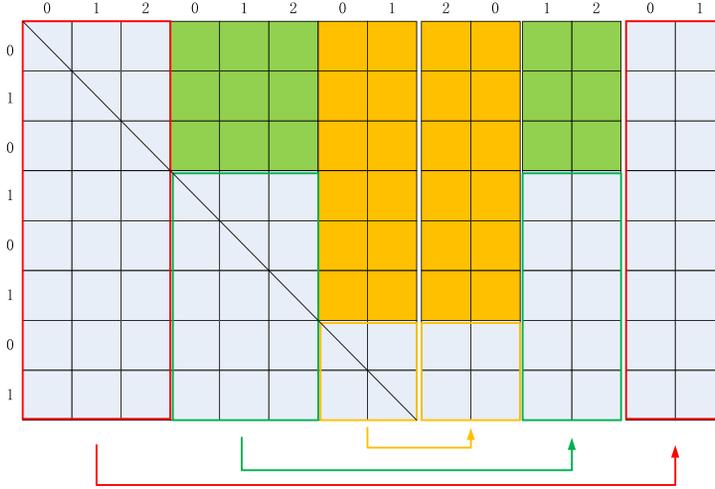


Fig. 4. Reverse neighboring checksum storage, with two checksum duplicates per  $Q$ -wide groups.

---

### Algorithm 1 Checksum Management

---

On a  $P \times Q$  grid, matrix is  $M \times N$ , block size is  $NB \times NB$

$C_k$  represents the  $k^{th}$  checksum block column

$A_k$  represents the  $k^{th}$  data block column

Before factorization:

Generate the initial checksum:

$$C_k = \sum_{j=(k-1) \times Q+1}^{(k-1) \times Q+Q} A_j, k = 1, \dots, \left\lceil \frac{N}{NB \times Q} \right\rceil$$

For each of  $C_k$ , make a copy of the whole block column and put right next to its original block column

Checksum  $C_k$  and its copy are put in the  $k^{th}$  position starting from the far right end

Begin factorization

Host algorithm starts with an initial scope of  $M$  rows and  $N + \left\lceil \frac{N}{Q} \right\rceil$  columns

For each  $Q$  panel factorizations, the scope decreases  $M$  rows and  $2 \times NB$  columns

End factorization

---

TRSM and GEMM claim most of the computation in the LU factorization, shrinking the update scope greatly reduces the overhead of the ABFT mechanism by diminishing the amount of (useless) extra computations; meanwhile, the efficiency of the update operation itself remains optimal as, thanks to the reverse storage scheme, the update still operates on a contiguous memory region and can be performed by a single PBLAS call.

Last, one can note that the iteration when the update scope is reduced is the same as when only the upper part of the checksum remains useful (for recovering the upper part of the matrix). The lower part of the checksum becomes invalid at this iteration, henceforth it is never referenced again during the following iterations (neither during computation or recovery). We will explain in the next section how this extra storage can be leveraged to protect the lower triangular part of the matrix.

#### 4.6. Delayed Recovery and Error Propagation

In this work, we assume that a failure can strike at any moment during the life span of factorization operations or even the recovery process. Theorem 4.2 proves that at the moment where the failure happens, the checksum invariant property is satisfied,

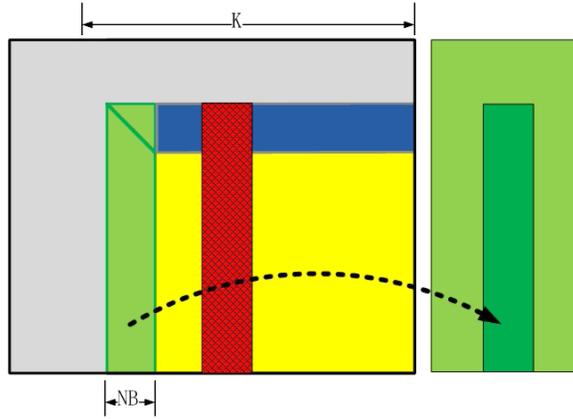


Fig. 5. Ghost pivoting Issue. Gray: Result in previous steps; Light Green: Panel factorization result in current step; Deep Green: The checksum that protects the light green; Blue: TRSM zone; Yellow: GEMM zone; Red: one of the columns affected by pivoting.

meaning that the recovery can proceed successfully. However, in large scale systems, which are asynchronous by nature, the time interval between the failure and the moment when it is detected by other processes is unknown, leading to delayed recoveries, with opportunities for error propagation.

The *ZU* factorization is composed of several subalgorithms that are called on different parts of the matrix. Matrix multiplication, which is used for trailing matrix updates and claims more than 95% of the execution time, has been shown to be ABFT compatible [Bosilca et al. 2009], that is to compute the correct result even with delayed recovery. One feature that has the potential to curb this compatibility is pivoting, in *LU*, especially when a failure occurs between the panel factorization and the row swapping updates, there is a potential for destruction of rows in otherwise unaffected blocks.

Figure 5 shows an example of such a case. Suppose the current panel contributes to the  $i^{th}$  column of checksum. When panel factorization finishes, the  $i^{th}$  column becomes intermediate data which does not cover any column of matrix. If a failure at this instant causes holes in the current panel area, then lost data can be recovered right away. Pivoting for this panel factorization has only been applied within the light green area. Panel factorization is repeated to continue on the rest of the factorization. However, if failure causes holes in other columns that also contribute to the  $i^{th}$  column of checksum, these holes cannot be recovered until the end of the trailing matrix update. To make it worse, after the panel factorization, pivoting starts to be applied outside the panel area and can move rows in holes into healthy area or vice versa, extending the recovery area to the whole column, as shown in red in Figure 5 including triangular solving area. To recover from this case, in addition to matrix multiplication, the triangular solver is also required to be protected by ABFT.

**THEOREM 4.5.** *Failure in the right-hand sides of triangular solver can recover from fail-stop failure using ABFT.*

**PROOF.** Suppose  $A$  is the upper or lower triangular matrix produced by LU factorization (nonblocked in ScaLAPACK LU),  $B$  is the right-hand side, and the triangular solver solves the equation  $Ax = B$ .

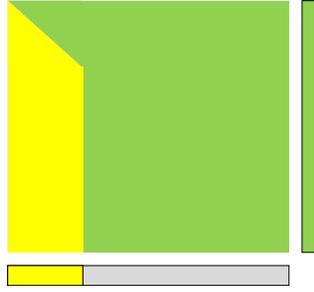


Fig. 6. Separation of lower and upper areas protected by checksum (green) and checkpoint (yellow) during the course of the factorization algorithm.

Supplement  $B$  with checksum generated by  $B_c = B * G_r$  to extended form  $\hat{B} = [B, B_c]$ , where  $G_r$  is the generator matrix. Solve the extended triangular equation:

$$\begin{aligned}
 Ax_c &= B_c = [B, B_c] \\
 \therefore x_c &= A^{-1} \times [B, B_c] \\
 &= [A^{-1}B, A^{-1}B_c] \\
 &= [x, A^{-1}BG_r] \\
 &= [x, xG_r].
 \end{aligned}$$

Therefore, data in the right-hand sides of the triangular solver is protected by ABFT.  $\square$

With this theorem, if failure occurs during triangular solving, lost data can be recovered when the triangular solver completes. Since matrix multiplication is also ABFT compatible, the whole red region in Figure 5 can be recovered after the entire trailing matrix update is done, leaving the opportunity for failure detection and recovery to be delayed at a convenient moment in the algorithm.

## 5. PROTECTION OF THE LEFT FACTOR MATRIX WITH Q-PARALLEL CHECKPOINT

It has been proven in Theorem 4.2 that the checksum covers only the upper triangular part of the matrix until the current panel, and the trailing matrix is subject to future updates. This is depicted in Figure 6, where the green checksum on the right of the matrix protects exclusively the green part of the matrix. Another mechanism must be added for the protection of the left factor (the yellow area).

### 5.1. Impracticability of ABFT for Left Factor Protection

The most straightforward idea, when considering the need of protecting the lower triangle of the matrix, is to use an approach similar to the one described above, but column-wise. Unfortunately, such an approach is difficult, if not impossible in some cases, as proved in the remaining of this section.

*5.1.1. Pivoting and Vertical Checksum Validity.* In LU, partial pivoting prevents the left factor from being protected through ABFT. The most immediate reason is as follow: The PBLAS kernel used to compute the panel factorization (see Figure 1) performs simultaneously the search for the best pivot in the column and the scaling of the column with that particular pivot. If applied directly on the matrix and the checksum blocks, similarly to what the trailing update approach does, checksum elements are at risk of being selected as pivots, which results in exchanging checksum rows into the matrix.

This difficulty could be circumvented by introducing a new PBLAS kernel that does not search for pivots in the checksum.

Unfortunately, legitimate pivoting would still break the checksum invariant property, due to row swapping. In LU, for matrix  $A$ ,

$$\begin{aligned} A &= \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{pmatrix} \\ &= \begin{pmatrix} L_{11}U_{11} & L_{11}U_{12} \\ L_{21}U_{11} & L_{21}U_{12} + L_{22}U_{22} \end{pmatrix}. \end{aligned} \quad (9)$$

Then panel factorization is

$$\begin{pmatrix} A_{11} \\ A_{21} \end{pmatrix} = \begin{pmatrix} L_{11}U_{11} \\ L_{21}U_{11} \end{pmatrix} = \begin{pmatrix} L_{11} \\ L_{21} \end{pmatrix} U_{11}. \quad (10)$$

To protect  $L_{11}$  and  $L_{21}$ , imagine that we maintain a separate checksum, stored at the bottom of the matrix, as shown in the yellow bottom rectangle of Figure 6, that we plan on updating by scaling it accordingly to the panel operation. In this vertical checksum, each  $P$  tall group of blocks in the 2D block cyclic distribution is protected by a particular checksum block. Suppose rows  $i_1$  and  $i_2$  reside on blocks  $k_{i_1}$  and  $k_{j_1}$  of two processes. It is not unusual that  $k_{i_1} \neq k_{j_1}$ . By Corollary 4.4, block  $k_{i_1}$  and  $k_{j_1}$  contribute to column-wise checksum block  $k_{i_1}$  and  $k_{j_1}$  respectively in the column that local blocks  $k_{i_1}$  and  $k_{j_1}$  belong to. This relationship is expressed as

$$\begin{aligned} \text{row } i_1 &\mapsto \text{checksum block } k_{i_1} \\ \text{row } j_1 &\mapsto \text{checksum block } k_{j_1}, \end{aligned}$$

where  $\mapsto$  reads ‘‘contributes to’’. After the swapping, the relationship should be updated to

$$\begin{aligned} \text{row } i_1 &\mapsto \text{checksum block } k_{j_1} \\ \text{row } j_1 &\mapsto \text{checksum block } k_{i_1}. \end{aligned}$$

This requires a regeneration of checksum blocks  $k_{i_1}$  and  $k_{j_1}$  in order to maintain the checkpoint validity. Considering there are  $nb$  potential pivoting operations per panel, hence a maximum of  $nb+1$  checksum blocks to discard, this operation has the potential to be as expensive as computing a complete vertical checkpoint.

*5.1.2. QR Factorization.* Although QR has no pivoting, it still cannot benefit from ABFT to cover  $Q$ , as we prove in the following text.

**THEOREM 5.1.**  *$Q$  in Householder QR factorization cannot be protected by performing factorization along with the vertical checksum.*

**PROOF.** Append a  $m \times n$  nonsingular matrix  $A$  with checksum  $GA$  of size  $c \times n$  along the column direction to get matrix  $A_c = \begin{bmatrix} A \\ GA \end{bmatrix}$ .  $G$  is  $c \times m$  generator matrix. Suppose  $A$  has a QR factorization  $Q_0R_0$ .

Obtain the QR factorization of  $A_c$ :

$$\begin{bmatrix} A \\ GA \end{bmatrix} = Q_c R_c = \begin{bmatrix} Q_{c11} & Q_{c12} \\ Q_{c21} & Q_{c22} \end{bmatrix} \begin{bmatrix} R_{c11} \\ \emptyset \end{bmatrix}.$$

$Q_{c11}$  is  $m \times m$  and  $Q_{c21}$  is  $c \times m$ .  $R_c$  is  $m \times n$  and  $\emptyset$  represents  $c \times n$  zero matrix.  $R_c \neq 0$  and is full rank. Because  $R_c$  is upper triangular with nonzero diagonal elements and therefore nonsingular.

$$Q_c Q_c^T = \begin{bmatrix} Q_{c11} & Q_{c12} \\ Q_{c21} & Q_{c22} \end{bmatrix} \begin{bmatrix} Q_{c11}^T & Q_{c21}^T \\ Q_{c12}^T & Q_{c22}^T \end{bmatrix} = I.$$

Therefore,

$$Q_{c11} Q_{c11}^T + Q_{c12} Q_{c12}^T = I. \quad (11)$$

Since  $A = Q_{c11} R_{c11}$  and  $R_{c11}$  is nonsingular, then  $Q_{c11} \neq 0$  and is nonsingular.

Assume  $Q_{c12} = 0$ .  $Q_{c11} Q_{c21}^T + Q_{c12} Q_{c22}^T = 0$ , therefore  $Q_{c11} Q_{c21}^T = 0$ . We have shown that  $Q_{c11}$  is nonsingular, so  $Q_{c21}^T = 0$  and this conflicts with  $GA = Q_{c21} R_{c11} \neq 0$ , so the assumption  $Q_{c12} = 0$  does not hold. From Equation 11,  $Q_{c11} Q_{c11}^T \neq I$ . This means even though  $A = Q_{c11} R_{c11}$ ,  $Q_{c11} R_{c11}$  is not a QR factorization of  $A$ .  $\square$

## 5.2. Panel Checkpointing

Given that the  $ZU$  factorization cannot protect  $Z$  by applying ABFT in the same way as for  $U$ , separate efforts are needed. For the rest of this article, we use the term “checksum” to refer to the ABFT checksum, generated before the factorization, that is maintained by the application of numerical updates during the course of the algorithm, in contrast to “checkpointing” for the operation that creates a new protection block during the course of the factorization. LU factorization with partial pivoting being the most complex problem, it is used here for the discussion. The method proposed in this section can be applied to the QR factorization with minimal efforts nonetheless.

In a  $ZU$  block factorization using 2D cyclic distribution, once a panel of  $Z$  is generated, it is stored into the lower triangular region of the original matrix. For example, in  $LU$ , vectors of  $L$ , except the diagonal ones, are stored in  $L$ . In QR, vectors  $v$  that are used to generate the elementary reflectors are stored. These lower triangular parts from the panel factorization are final results, and are not subject to further updates during the course of the algorithm, except for partial pivoting row swapping in LU. Therefore only one vertical checkpointing “should be” necessary to maintain each panel’s safety, as is discussed in Davies et al. [2011]. We will show how this idea, while mathematically trivial, needs to be refined to support partial pivoting. We will then propose a novel checkpointing scheme, leveraging properties of the block algorithm to checkpoint  $Z$  in parallel, that demonstrates a much lower overhead when compared to this basic approach.

## 5.3. Postponed Left Pivoting

Although once a panel is factored, it is not changed until the end of the computation, row swaps incurred by pivoting are still to be applied to the left factor as the algorithm progresses in the trailing matrix, as illustrated in Figure 1. The second step (pivoting to the left) swaps two rows to the left of the current panel. The same reasoning as presented in Section 5.1.1 holds, meaning that the application of pivoting row swaps to the left factor has the potential to invalidate checkpoint blocks. Since pivoting to the left is carried out in every step of LU, this causes significant checkpoint maintenance overhead.

Unlike pivoting to the right, which happens during updates and inside the panel operation, whose result are reused in following steps of the algorithm, pivoting to the left can be postponed. The factored  $L$  is stored in the lower triangular part of the

matrix without further usage during the algorithm. As a consequence, we delay the application of all left pivoting to the end of the computation, in order to avoid expensive checkpoint management. We keep track of all pivoting that should have been applied to the left factor, and when the algorithm has completed, all row swaps are applied just in time before returning the end-result of the routine.

#### 5.4. Q-Parallel Checkpointing of $Z$

The vertical checkpointing of the panel result requires a set of reduction operations immediately after each panel factorization. Panel factorization is on the critical path and has lower parallelism, compared to other routines of the factorization (such as trailing matrix update). The panel factorization works only on a single block column of the matrix, hence benefits from only a  $P$  degree of parallelism, in a  $P \times Q$  process grid. Checkpointing worsens this situation, because it applies to the same block column, and is bound to the same low level of exploitable parallelism. Furthermore, the checkpointing cannot be overlapped with the computation of the trailing matrix update: all processes who do not appear on the same column of the process grid are waiting in the matrix-matrix multiply PBLAS, stalled because they require the panel column to enter the call in order for the result of the panel to be broadcasted. If the algorithm enters the checkpointing routine before going into the trailing update routine, the entire update is delayed. If the algorithm enters the trailing update before starting the checkpointing, the checksum is damaged in a way that prevents recovering that panel, leaving it vulnerable to failures.

Our proposition is then twofold: we protect the content of the blocks before the panel, which then enables starting immediately the trailing update without jeopardizing the safety of the panel result. Then, we wait until sufficient checkpointing is pending to benefit from the maximal parallelism allowed by the process grid.

*5.4.1. Enabling Trailing Matrix Update Before Checkpointing.* The major problem with enabling the trailing matrix update to proceed while the checkpointing of the panel is not finished is that the ABFT protection of the update modifies the checksum in a way that disables protection for the panel blocks. To circumvent this limitation, in a  $P \times Q$  grid, processes are grouped by sections of width  $Q$ , that are called a *panel scope*. When the panel operation starts applying to a new section, the processes of this panel scope make a local copy of the impending column and the associated checksum, called a *snapshot*. This operation involves no communication, and features the maximum  $P \times Q$  parallelism. The memory overhead is limited, as it requires only the space for at most two extra columns to be available at all time, one for saving the state before the application of the panel to the target column, and one for the checksum column associated to these  $Q$  columns. The algorithm then proceeds as usual, without waiting for checkpoints before entering the next  $Q$  trailing updates. Because of the availability of this extra protection column, the original checksum can be modified to protect the trailing matrix without threatening the recovery of the panel scope, which can roll back to that previous dataset should a failure occur.

*5.4.2. Q-Parallel Checkpointing.* When a panel scope is completed, the  $P \times Q$  group of processes undergo checkpointing simultaneously. Effectively,  $P$  simultaneous checkpointing reductions are taking place along the block rows, involving the  $Q$  processes of that row to generate a new protection block. This scheme enables the maximum parallelism for the checkpoint operation, hence decreasing its global impact on the failure free overhead. Another strong benefit is that it scales with the process grid perfectly, whereas regular checkpointing suffers from scaling with the square root of the number of processes (as it involves only one dimension of the process grid).

**5.4.3. Optimized Checkpoint Storage.** According to Corollary 4.4, starting from the first block column on the left, every  $Q$  block columns contribute to one block column of checksum, which means that once the factorization is done for these  $Q$  block columns, the corresponding checksum block column becomes useless (it does not protect the trailing matrix anymore, it has never protected the left factor, see Theorem 4.2). Therefore, this checksum storage space is available for storing the resultant checkpoint block generated to protect the panel result. Following the same policy as the checksum storage, discussed in Section 4.5.2, the checkpoint data is stored in reverse order from the right of the checksum (see Figure 4). As this part of the checksum is excluded from the trailing matrix update, the checkpoint blocks are not modified by the continued operation of the algorithm.

**5.4.4. Recovery.** The hybrid checkpointing approach requires a special recovery algorithm. Two cases are considered. First, when failure strikes during the trailing update, immediately after a panel scope checkpointing. For this case, the recovery is not attempted until the current step of the trailing update is done. When the recovery time comes, the checksum/checkpointing on the right of the matrix matches the matrix data as if the initial ABFT checksum had just been performed. Therefore any lost data blocks can be recovered by the simple reverse application of the ABFT checksum relationship.

The second case is when a failure occurs during the  $Q$  panel factorization, before the checkpointing for this panel scope can successfully finish. In this situation, all processes revert the panel scope columns to the snapshot copy. Holes in the snapshot data are recreated by using the snapshot copy of the checksum, applying the usual ABFT recovery. The algorithm is resumed in the panel scope, so that panel and updates are applied again within the scope of the  $Q$  wide section; updates outside the panel scope are discarded, until the prefailure iteration has been reached. Outside the panel scope, regular recovery mechanisms are deployed (ABFT checksum inversion for the trailing matrix, checkpoint recovery for the left factor). When the refactorization of panels finishes, the entire matrix, including the checksum, is recovered back to the correct state. The computation then resumes from the next panel factorization, after the failing step.

Figure 7 shows an example of the recovery when the process (1,0) in a  $2 \times 3$  grid failed. It presents the difference between the correct matrix dataset and the current dataset during various steps of failure recovery as a “temperature map,” brighter colors meaning large differences and black insignificant differences. The matrix size is  $80 \times 80$  and  $NB = 10$ , therefore the checksum size is  $80 \times 60$ . Failure occurs after the panel factorization starting at (41,41) is completed, within the  $Q = 3$  panel scope. First, using a fault tolerant MPI infrastructures, like FT-MPI [Fagg et al. 2004], the failed process (0,1) is replaced and reintegrates the process grid with a blank dataset, showing as evenly distributed erroneous blocks (A). Then the recovery process starts by mending the checksum using duplicates (B). The next step recovers the data which is outside the current panel scope (31:80,31:60), using the corresponding checksum for the right factor, and the checkpoints for the left factor (C). At this moment, all the erroneous blocks are repaired, except those in the panel scope (41:80, 41:50). Snapshots are applied to the three columns of the panel scope (31:80,31:60). Since these do not match the state of the matrix before the failure, but a previous state, this area appears as very different (D). Panel factorization is relaunched in the panel scope, in the area (31:80,31:60), with the trailing update limited within this area. This refactorization continues until it finishes panel (41:80,41:50) and by that time the whole matrix is recovered to the correct state (not presented, all black). The LU factorization can then proceed normally.

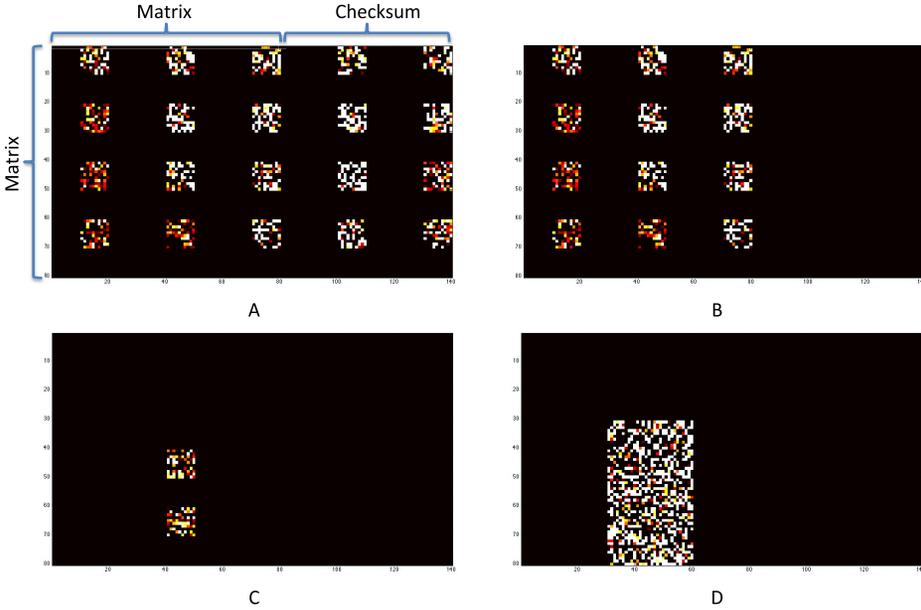


Fig. 7. Recovery example (matrix size  $800 \times 800$ , grid size  $2 \times 3$ , failure at process (0,1), step:41). A: Failure occurs; B: Checksum recovered; C: Data recovered using ABFT checksum and checkpointing output; D: Three panels restored using snapshots.

## 6. EXTENSION TO MULTIPLE FAILURES

To evaluate the extension of this ABFT technique to multiple failures we need to consider the definition of multiple failures from a timing perspective. If we consider a failure and the resulting reaction of the ABFT algorithm, we can define a failure period as the time between the moment where the failure is detected, and the moment where the algorithm has completely recovered its data, and is ready to continue the factorization. This time interval contains the recovery of the MPI processes, as well as the recovery of the user level data using the checksums, as presented in Section 5.4.4. It is worth mentioning, that once the system has recovered, it reaches a state similar to a failure-free execution, and it is again ready to face subsequent failures. Thus, as long as the failure-recovery periods do not overlap, the single failure scheme described in the previous sections can survive any number of faults. We will show in Section 7.5 that the number of subsequent faults has little impact on the accuracy of the factorization.

In the rest of this section, we focus on the more demanding case of simultaneous, overlapping failures. We use the following notations: an initial matrix of size  $M \times N$  is distributed on a process grid of  $P \times Q$ , in a 2D Block Cyclic way, where blocks are of size  $mb \times nb$  (see Section 4.4). The multiple failure ABFT protocol takes an additional parameter,  $F$ , denoting the maximal number of overlapping failures to tolerate, and we consider the case of  $f \leq F$  overlapping failures.

First, we note that the recovery algorithm works on block rows, so if failures happen on processors that are located on different rows of the process grid, the recovery algorithm can recover the missing blocks independently. However, in the worst case, all failures happen on the same process row, resulting in  $f \leq F$  missing blocks that must be reconstructed from their checksums.

The idea of the algorithm is still to complement the initial data with a checksum. As a first approximation, let us consider that failures do not hit the checksum part. To

tolerate at most  $f$  failures, we need  $f$  checksum functions  $ck_i, 1 \leq i \leq f : mb \times (nb \cdot Q) \mapsto mb \times nb$  that maps  $Q$  data blocks to a checksum block, so that the  $f$  combinations  $(b_1, b_2, \dots, b_Q) \rightarrow ck_i(b_1, b_2, \dots, b_Q), 1 \leq i \leq f$  form a linear system of  $f$  independent equations. In the case of one fault, we usually take  $ck_1 = \text{sum}$ . For  $ck_i, i > 1$ , we can take weighted sums (see section 4.1).

When a failure occurs, the missing blocks can be recovered by solving the linear system of equations, assuming that the checksum blocks are not lost. When failures introduce the loss of a checksum block, its data must be recovered. In the case of a single failure, we used the replication strategy: the checksum blocks are replicated on consecutive (thus independent) processors of the process grid, allowing to recover the missing checksum data when necessary.

In the case of multiple overlapping failures, the best strategy consists in ensuring that for any combination of failures, enough checksum blocks will survive to solve the linear system of equations with matrix data blocks (i.e., nonchecksum blocks) as unknown. Once this system is solved, the missing checksum blocks are then restored by recomputing the checksum value from the recovered data. For protection against  $F$  failures, at least  $F$  checksum blocks must remain available per group of  $Q$  blocks, so that the system remains fully determined.

If it were possible to store the checksum blocks on another row in the process grid, then  $F$  checksums per  $Q$  blocks would be sufficient. If  $f_p$  failures hit processes in the row  $p$  of the process grid, then, at most  $F - f_p$  checksum blocks would be damaged, meaning that  $f_p$  checksum blocks would in all cases remain available to restore the  $f_p$  lost blocks. Unfortunately, we recall that this approach is not possible, as the checksums have to be stored on the same process row as the block they protect for applying ABFT updates: the same mathematical transformation (as well as pivoting) has to be applied on the trailing matrix blocks and the checksum blocks during the update. If the checksums are offset, a different set of transformations would be applied (pertaining to the update of the next rows in the real matrix), and the checksum invariant property would be lost.

As a consequence, let us consider only the case where the  $Q$  processes of a particular row in the process grid store the data blocks and their associated checksums. Since  $F \geq f \geq 0, 2F - f \geq f$ ; thus, by considering  $2F$  consecutive processes to store the different  $ck_i$  checksums, one can ensure that at least  $f$  checksum blocks will survive, allowing to solve the linear system of  $f$  equations to recover the matrix data. This also implies that  $2F \leq Q$ , to prevent checksum blocks to be lost simultaneously by cycling over the process grid. Note that with  $K < 2F$  checksum blocks to cover  $Q$  processes, this strategy would be subject to a risk of nonrecoverable failure scenarios: if  $f = F$ , then less than  $K - f < f$  checksum blocks can survive the overlapping failures, and simultaneously,  $f$  matrix blocks belonging to the same segment of  $Q$  blocks can be lost. Because  $f$  equations are necessary to solve the corresponding system, and less than  $K - f < f$  blocks would be available,  $2F$  checksum blocks are sufficient and necessary.

The protocol stores one such set of  $2F$  checksum blocks per  $Q$  blocks of original data in the row, so in total, the memory overhead of ABFT to tolerate at most  $F \leq Q/2$  faults on a single block-row is  $\frac{2F}{Q} \frac{N}{nb}$  blocks of  $mb \times nb$  elements (that is  $\frac{2F \cdot N \cdot mb}{Q}$  elements). On the whole matrix, this corresponds to  $\frac{2F \cdot M \times N}{Q}$  elements.

Compared to the case with one failure,  $f$  failures also introduce a computational overhead since we must solve the linear equation systems to recover the original data. Fortunately, these systems are small:  $f$  equations of  $(Q + 1)$  blocks each, and can be solved locally, where the blocks must be stored, after the corresponding data has been gathered.

## 7. EVALUATION

In this section, we evaluate the performance of the proposed fault tolerant algorithm based on ABFT and reverse neighboring checkpointing. For a fault tolerant algorithm, a crucial consideration is the overhead added to failure free execution rate, due to various fault tolerance mechanisms such as checksum generation, checkpointing and extra flops. An efficient and scalable algorithm will incur a minimal overhead over the original algorithm while enabling scalable reconstruction of lost dataset in case of failure. A fault-tolerant strategy that has minimal failure free impact is more prone to being deployed by default, it then has the opportunity to help users survive occasional failures and improve effective throughput.

For experimental runs, we use the NSF Kraken supercomputer, hosted at the National Institute for Computational Science (NICS, Oak Ridge, TN) as our testing platform. This machine features 112,896 2.6GHz AMD Opteron cores, 12 cores per node, with the Seastar interconnect. At the software level, to serve as a comparison base, we use the non fault tolerant ScaLAPACK LU and QR in double precision with block size  $NB = 100$ . The fault tolerance functions are implemented and inserted as drop-in replacements for ScaLAPACK routines.

In this section, we first evaluate the theoretical overhead in the form of extra memory usage and computations, then show experimental results on Kraken to assess the effective performance of the method.

### 7.1. Complexity Analysis

**7.1.1. Extra Storage.** Checksum takes extra storage (memory), but on large scale systems, memory usage is usually maximized for computing tasks. Therefore, it is preferable to have a small ratio of checksum size over matrix size, in order to minimize the impact on the memory available to the application itself. For the sake of simplicity, and because of the small impact in term of memory usage, neither the pivoting vector nor the column shift are considered in this evaluation.

Different protection algorithms require different amounts of memory. In the following, we consider the duplication algorithm presented in Section 4.5.2 for computing the upper memory bound. The storage of the checksum includes the row-wise and column-wise checksums.

For an input matrix of size  $M \times N$  on a  $P \times Q$  process grid, the memory used for checksum (including duplicates) is  $M \times \frac{N}{Q} \times 2$ . The ratio  $R_{mem}$  of checksum memory over the memory of the input matrix, equals to  $\frac{2}{Q}$ , becomes negligible with the increase in the number of processes used for the computation.

**7.1.2. Extra Computations.** Due to the introduction of checksum, operations counts and communication have been increased, as update operation span on a larger matrix comprised of the original trailing matrix and the checksums. During checkpointing and recovery, extra workload is performed and this all together leads to higher computing complexity than the original implementation in ScaLAPACK.

For simplicity of description, we first consider square data matrices of size  $N \times N$  distributed on a square grid  $Q \times Q$ . The operation count ration for LU factorization without and with checksum is

$$\begin{aligned}
 R &= \frac{\frac{2}{3}N^3 - \frac{1}{2}N^2 + \frac{5}{6}N}{\frac{2}{3}(N + \frac{N}{Q})^3 - \frac{1}{2}(N + \frac{N}{Q})^2 + \frac{5}{6}(N + \frac{N}{Q})} \\
 &= \frac{\frac{2}{3} - \frac{1}{2N} + \frac{5}{6N^2}}{\frac{2}{3}(1 + \frac{1}{Q})^3 - \frac{1}{2N}(1 + \frac{1}{Q})^2 + \frac{5}{6N^2}(1 + \frac{1}{Q})}. \tag{12}
 \end{aligned}$$

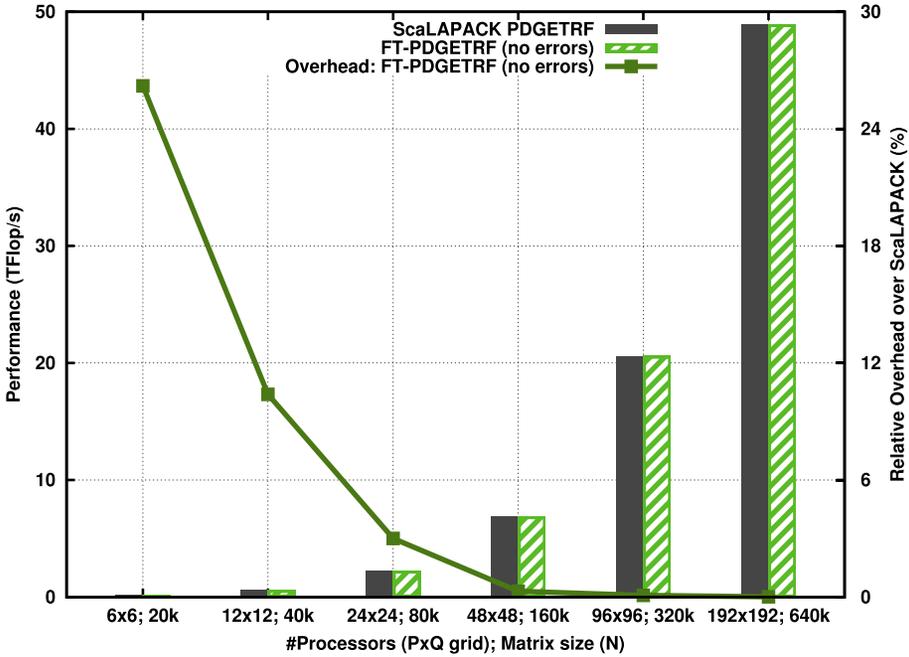


Fig. 8. Weak scalability of FT-LU: performance and overhead on Kraken, w.r.t. non fault tolerant LU.

Clearly  $\lim_{Q \rightarrow +\infty} R = 1$ . Hence for systems with high number of processes, the extra flops for updating checksum columns is negligible with respect to the normal flops realized to compute the result.

In addition, checksums must be generated, once at the start of the algorithm, the second time at the completion of a  $Q$ -wide panel scope. Both these activities account for  $O(N^2)$  extra computations (but enjoy excellent parallelism).

## 7.2. Overhead without Failures

Figure 8 evaluates the completion time overhead and performance, using the LU factorization routine PDGETRF. The performance in Tflop/s of both the original and fault tolerant version are presented in bar graphs. This experiment is carried out to test the weak scalability, where both the matrix and grid dimension doubles. The superimposed curves present the same data as overhead relative to the non-fault-tolerant algorithm in ScaLAPACK. As a reminder, the same overall number of processors is employed with or without fault tolerance: the same processors that carry the factorization computation also carry the checksum storage and maintenance operations. The result outlines that as the problem size and grid size increases, the overhead drops quickly and eventually becomes negligible. At the matrix size of  $640,000 \times 640,000$ , on 36,864 ( $192 \times 192$ ) cores, both versions achieved over 48Tflop/s, with an overhead of 0.016% for the ABFT algorithm.

As a side experiment, we implemented the naive vertical checkpointing method discussed in Section 5.2. As the left factor is touched only once during the computation, the approach of checkpointing the result of a panel synchronously can, *a-priori*, look sound. However, as the checkpointing of a particular panel suffers from its inability to exploit the full parallelism of the platform, it is subject to a derivative of Amdahl's law, its parallel efficiency is bound by  $P$ , while the overall computation enjoys a  $P \times Q$  parallel efficiency: its importance is bound to grow when the number of computing

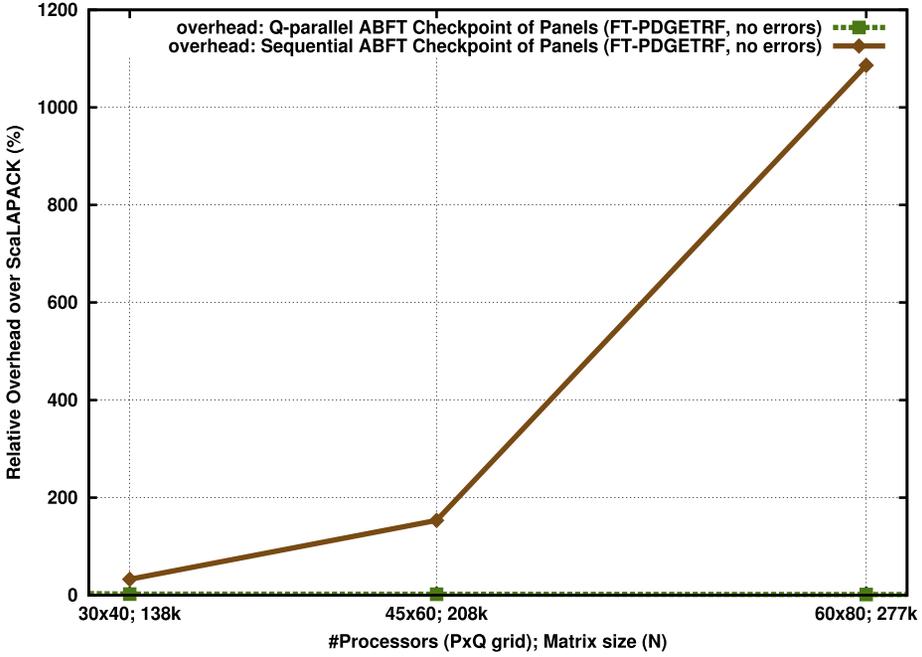


Fig. 9. Comparison between Q-parallel and per-panel L-checkpoint overhead.

resources increases. As a consequence, in the experiments, presented in Figure 9, the time to compute the naive checkpoint dominates the computation time and impose unacceptable overhead even for modest number of cores. On the other hand, the hybrid checkpointing approach exchanges the risk of a Q-step rollback with the opportunity to benefit from a  $P \times Q$  parallel efficiency for the panel checkpointing. Because of this improved parallel efficiency, the hybrid checkpointing approach benefits from a competitive level of performance, that follows the same trend as the original non fault tolerant algorithm.

### 7.3. Recovery Cost

Beside the “curb” cost of protection against failures, another interesting performance metric is the cost of recovering from a failure. In the Q-parallel hybrid scheme, there are two cases for the recovery. The first one is when failure occurs right after the reverse neighboring checkpointing of  $Q$  panels. At this moment the matrix is well protected by the checksum and therefore the lost data can be recovered directly from the checksum. We refer to this case as “failure on Q panels border.” The second case is when the failure occurs during the reverse neighboring checkpointing and therefore local snapshots have to be used along with refactorization to recover the lost data and restore the matrix state. This is referred to as the “failure within Q panels.”

Figure 10 shows the overhead from these two cases for the LU factorization, along with the no-error overhead as a reference. In the “border” case, the failure is simulated to strike when the  $96^{th}$  panel (which is a multiple of grid columns, 6, 12,  $\dots$ , 48) has just finished. In the “nonborder” case, failure occurs during the  $(Q + 2)^{th}$  panel factorization. For example, when  $Q = 12$ , the failure is injected when the trailing update for the step with panel (1301,1301) finishes. From the result in Figure 10, the recovery procedure in both cases adds a small overhead that also decreases when scaled to large

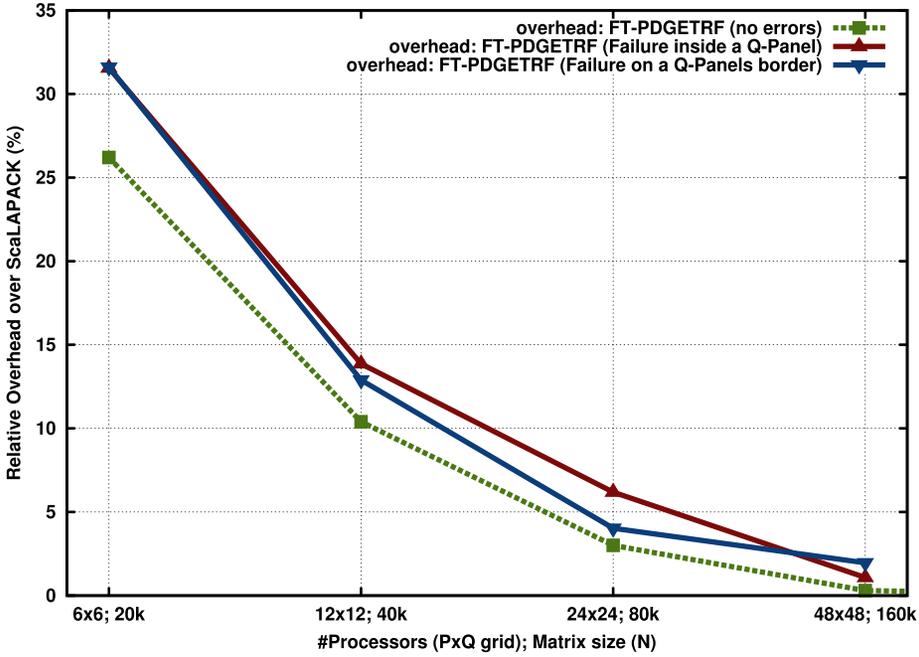


Fig. 10. Weak scalability of FT-LU: runtime overhead on Kraken when failures strike at different steps.

problem size and process grid. For largest setups, only 2-3 percent of the execution time is spent recovering from a failure.

#### 7.4. Extension to Other factorization

The algorithm proposed in this work can be applied to a wide range of dense matrix factorizations other than LU. As a demonstration we have extended the fault tolerance functions to the ScaLAPACK QR factorization in double precision. Since QR uses a block algorithm similar to LU (and also similar to Cholesky), the integration of fault tolerance functions is mostly straightforward. Figure 11 shows the performance of QR with and without recovery. The overhead drops as the problem and grid size increase, although it remains higher than that of LU for the same problem size. This is expected: as the QR algorithm has a higher complexity than LU ( $\frac{4}{3}N^3$  v.s.  $\frac{2}{3}N^3$ ), the ABFT approach incurs more extra computation when updating checksums. Similar to the LU result, recovery adds an extra 2% overhead. At size 160,000 a failure incurs about 5.7% penalty to be recovered. This overhead becomes lower, the larger the problem or processor grid size considered.

*About tall and skinny matrices.* One of the major use cases of the QR factorization involves computing the factorization of tall and skinny matrices, for which the vertical dimension  $M$  is much larger than  $N$ . In this section we consider the computational overhead of ABFT protection in this case.

When  $M \geq N$ , the normal operation count of the QR factorization is

$$2MN^2 - \frac{2}{3}N^3 + MN + N^2.$$

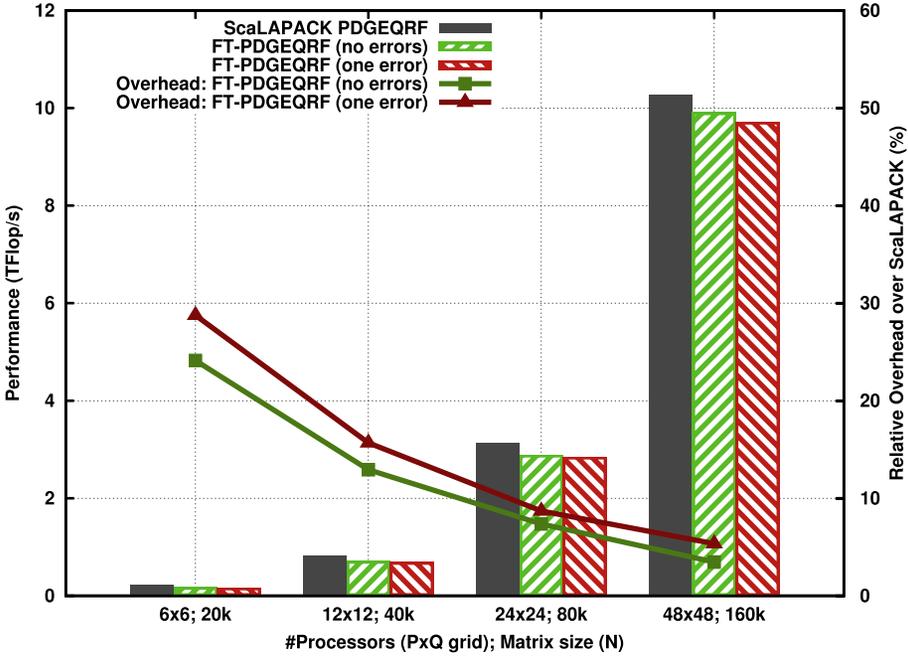


Fig. 11. Weak scalability of FT-QR: runtime overhead on Kraken when failures strike.

An overestimate of the number of flops performed to update the checksum can be obtained by substituting  $N$  by its extended value including checksum blocks  $N + \frac{N}{Q}$ .  $M$  remains unchanged. The ratio is then bound by

$$R = \frac{2MN^2 - \frac{2}{3}N^3 + MN + N^2}{2M(N + \frac{N}{Q})^2 - \frac{2}{3}(N + \frac{N}{Q})^3 + M(N + \frac{N}{Q}) + (N + \frac{N}{Q})^2}$$

Similarly to LU,  $\lim_{Q \rightarrow +\infty} R = 1$ . However, when  $M$  is much larger than  $N$ ,  $P$  must also be larger than  $Q$  in order to preserve an adequate load balance between the processes. It is therefore expected that the overhead ratio of ABFT for the block QR factorization applied on tall and skinny matrices would decrease at a slower rate when increasing the size of the target system. At the extreme, if one scales only the  $M$  dimension of the matrix, the ratio of additional computation incurred by ABFT protection would remain stable (instead of decreasing with system size).

One observation that diminish the practical impact of this issue lies in the fact that the standard block QR factorization underperforms in such extreme conditions; as the matrix becomes increasingly tall, more time is spent in the poorly parallel PDGEQR2 operation on the panel, while the time spent in flop rich trailing matrix updates becomes insignificant. Different algorithms, such as the hierarchical QR factorization [Cosnard et al. 1986] or communication avoiding QR [Demmel et al. 2012] should be employed when facing very tall and skinny matrices; rendering these significantly different algorithms fault tolerant is beyond the scope of this article.

### 7.5. Numerical Stability

Direct factorizations are backward stable, meaning that slight perturbation introduced by round-off errors are not amplified by the algorithm, and that the accuracy of the

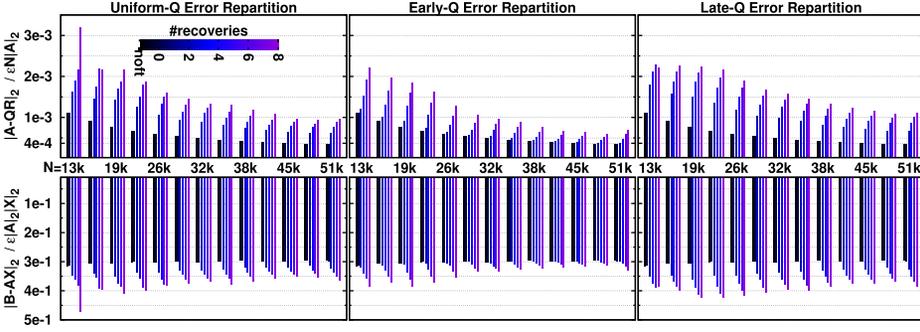


Fig. 12. Variation of result accuracy depending on temporal distribution of an increasing number of independent failures. 16x8 process grid, normalized to machine precision  $\varepsilon=1.1e-16$ .

final result mostly depends on the condition number of the input matrix. ABFT recovery is based on matrix summation, which is also backward stable, henceforth, the results obtained from execution containing failures and recovery are theoretically of good quality as well. In this section, we illustrate with some empirical measurements that not only recovery is backward stable, but that under the impact of massive number of failures, the factorization continues to deliver excellent stability (as predicted by the theory), but also that the solution of a linear system using the resultant factorization is not impacted in any meaningful way.

In the following experiments, we compute the QR factorization of a matrix, under the protection of ABFT checksums (with Q-parallel checkpoint of L). We introduce an increasing number of failures. Failures are nonsimultaneous, a new failure is introduced only after the previous recovery is entirely completed. We observe the behavior when failure position and timing varies.

We measure the impact on numerical quality of the factorization by computing four different metrics. The QR factorization is applied on an input matrix  $A$ , randomly generated with the `pdmag` routine from ScaLAPACK testings, the condition number of the resultant matrix is around  $1e6$ . The norm of the element-wise difference between the QR factorization of  $A$  and its initial state,  $\frac{\|A-QR\|_2}{N\|A\|_2}$ , gives the first evaluation of the quality of the resultant factorization under the stress of errors. Then, we generate  $X_0$  and  $B$  so that  $B = A \times X_0$ . The solution  $X$  of the linear system  $AX = B$  is computed using the result of the QR factorization of  $A$ , so that we can compute the backward error,  $\frac{\|B-AX\|_2}{\|A\|_2\|X\|_2}$ . To put the results in perspective, all norms are divided by the machine precision  $\varepsilon = 1.1102e - 16$ . We also compute the number of identical digits between the corrected solution  $X$  and the solution without ABFT  $X_{noft}$ , which is given by the log of the forward error  $-\log(\frac{\|X_{noft}-X\|_2}{N\|X_{noft}\|_2})$ , using  $X_{noft}$  as the “correct” solution rather than the real solution  $X_0$ .

In Figure 12, we consider three temporal repartitions of failures. In the *Uniform* strategy, the iterations of the algorithm are divided into  $f+1$  equal intervals, separated by  $f$  failures. In the *Early* strategy, a failure is introduced in every panel scope (said otherwise every  $Q \times NB$  iterations), starting from iteration 1, until  $f$  failures have been introduced. The *Late* strategy is symmetrical:  $f$  failures are introduced during the  $f$  latest panel scope at the end of the execution.

Each group of bar presents the results for a given matrix size, and within such a group, each bar denotes a different number of injected non-overlapping failures (and corresponding recoveries). The first general observation is that accuracy of result improve with matrix size, which is a known property of direct factorization algorithms.

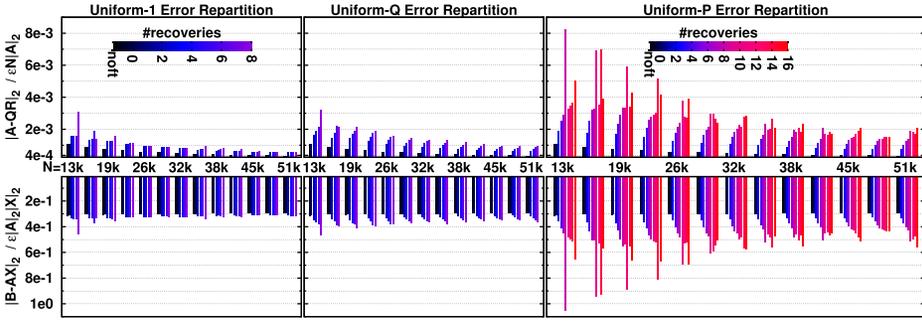


Fig. 13. Variation of result accuracy depending on spatial distribution (process grid position of the impacted processor) of an increasing number of independent failures.  $16 \times 8$  process grid, normalized to machine precision  $\varepsilon = 1.1e-16$ .

Even when injecting failures, this property continues to hold. Similarly, both the quality of the QR factorization and the solution of the linear system continue to behave similarly under the stress of failures. For a given problem size, each individual failure (and resultant round-off error during recovery) adds a small perturbation to the accuracy of the result; generally, the accuracy loss increases linearly with the number of recoveries during the run. However, the growth rate of the error when the number of failures increases is very low, and all obtained solutions are still of excellent quality.

Regarding the effect of temporal repartition of failures, one can note that failures happening early during the factorization have less impact than failures happening at the end of the factorization. Especially more so for small matrices. The explanation comes from the fact that the QR algorithm absorbs round-off error during its execution without failures, so in effect, round-off errors introduced early during the run get corrected by the normal unfolding of the QR algorithm on the remainder of the matrix. For failures that arise at the end of the computation, there is no such opportunity for self correction; the case is analogous to concentrating errors in the factorization of a smaller matrix (that has not yet been factorized), which, as demonstrated by the result, yields more impact on accuracy (see the case for  $N < 25K$ ). The Uniform repartition, which more closely depicts a random distribution of failures, is subject to a medium level of perturbation, compared to Late error distribution. The difference is most important on small matrices. For larger matrices, results tend to converge to a very small amount of precision loss, especially in terms of backward error.

We also consider three positional repartitions of failures, in Figure 13. In the *One* strategy, the failure is always introduced on processor  $(1, 1)$  in the  $P \times Q = 16 \times 8$  process grid. As this strategy repeatedly damages the same blocks for each subsequent failures, it minimizes the number of blocks where recovery injects round-off errors, for a given number of failures. In the *P* strategy, failures are affecting processors along the vertical direction of the process grid: a new failure is introduced on processor  $(p, 1)$  when the previous one impacted  $(p - 1, 1)$ , starting from  $(1, 1)$ . Similarly, in the *Q* strategy, failures affect processors along a row of the process grid, damaging processors  $(1, q)$ . In these strategies, with a sufficient number of failures, all blocks from some column or row of the matrix are destroyed and recovered at least once, a stressful situation.

As expected, the *One* failure distribution imparts a very limited accuracy loss. After the initial loss of accuracy resulting from the first failures, adding supplementary failures impacting the same blocks has no effect on the quality of both the factorization and solution. At the other extreme, the *P* repartition of errors is the most stressful scenario. While other scenarios impart a maximum of 8 failures, we inject up to

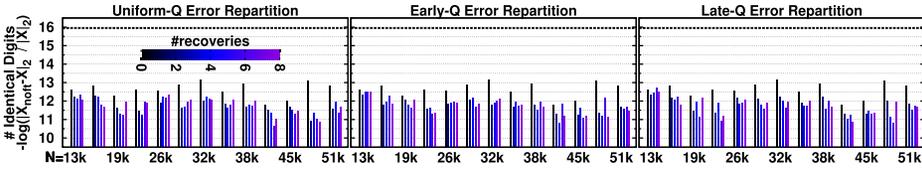


Fig. 14. “Forward error” assuming the correct solution is the approximate solution computed w/o fault tolerance; this is an estimate of the number of identical digits in both solutions. 16x8 process grid;  $\varepsilon=1.1e-16$ .

16 failures in the  $P$  scenario, so that all blocks of the impacted columns are destroyed and restored at least once over the course of the factorization. If we limit the analysis to less than 8 failures, the accuracy loss is comparable to that experienced by the  $Q$  scenario, however, linear growth factor per failure is higher. As failures concentrate on the same columns, the computation of the householder reflectors is impacted, which further disseminates the rounding errors. When injecting over 8 failures, the growth of the round-off error stops being linear with the number of failure, but oscillates in a more chaotic fashion. However, as observed for larger matrix sizes, a predictable behavior can be achieved again for matrices as small as  $N=32K$ , for which 16 failures during the small runtime (up to 310s, depending on the scenario), and at a small scale (128 nodes), is a very pessimistic scenario. In all cases, even when considering failure frequency well beyond reasonable expectations, the quality of the solution is well within acceptable range, two orders of magnitudes below the typical quality threshold of 100 for this type of random matrices. For larger matrices, the backward errors is at most doubled when introducing 16 errors in the worst spatial distribution.

The last set of experiments, presented in Figure 14, investigates the difference between the solution as computed by the ABFT QR, and by the original QR algorithm. Some science disciplines rely on algorithms that are chaotic in nature. As a consequence, a small deviation in the quality of the solution can result in correct, but widely different outcomes for the complete simulation at hand (for example, climate simulation is very sensitive to initial conditions). The closer the ABFT recovered result is to the original solution, the more reproducible and predictable such type of unstable simulations are. The set of graphs present the distance between  $X_{noft}$ , the solution without fault tolerance, and  $X$  the solution computed while sustaining failures. As a reminder, the condition number of the input matrices  $K(A)$  is in the order of  $10^6$ . As a consequence, it is expected that the result looses up to 6 digits of accuracy. In practice, the forward error is smaller, and the typical accuracy loss is only 4 digits. For 0 failures, the difference between  $X_{noft}$  and  $X$  is also the forward error, and illustrates the typical difference to be expected by running the same algorithm on a different hardware platform or software version (the difference in this case comes from the ordering of commutative operations inside MKL GEMM, which is different when the update is applied to a different matrix size). It is indeed notable that the quality of the solution without failures can vary depending on matrix size, number of processors, data distribution, and even underlying communication patterns employed within the MPI implementation of collective operations, and chaotic algorithms already have to cope with this possibility to obtain reproducible results on different deployments.

When failures are recovered, the difference between  $X_{noft}$  and the recovered  $X$  is in most cases very small, in the same order of magnitude as the forward error. For some matrix sizes, the recovered solution is sometimes closer to the real solution than the non fault tolerant algorithm (as is the case  $N=48K$  on the Early-Q error repartition). This may seem advantageous at first glance, but it may indicate that the difference between the recovered solution and the non-fault-tolerant solution is actually large. In

some rare instances, the distance between the two computed solutions reaches over a full significant digit. While the recovered solution is always perfectly acceptable, it is more variable than the original algorithm solution would be. There is no clear pattern driving the appearance of more distant solutions. Overall, the supplementary uncertainty on the result is measurable, but is only slightly larger in magnitude compared to the uncertainty from running on a different hardware platform. In this set of experiments, trimming all results to the 10 significant digits expected from the condition number of the matrices (by padding the remainder of the mantissa with 0, as an example) would achieve perfectly reproducible results in all cases.

## 8. CONCLUSION

In this article, by assuming a failure model in which fail-stop failures can occur anytime on any process during a parallel execution, a general scheme of ABFT algorithms for protecting one-sided matrix factorizations is proposed. This scheme can be applied to a wide range of dense matrix factorizations, including Cholesky, LU and QR. A significant property of the proposed algorithms is that both the left and right factorization results are protected. ABFT is used to protect the right factor with checksum generated before, and carried along during the factorizations. A highly scalable checkpointing method is proposed to protect the left factor. This method cooperatively reutilizes the memory space originally designed to store the ABFT checksum, and has minimal overhead by strategically coalescing checkpoints of many iterations. An extended method to tolerate multiple simultaneous failures is described. Large scale experimental results validate the design of the proposed fault tolerance method by highlighting a decreasing overhead for both LU and QR, and thus a highly scalable approach. The accuracy of the method is also investigated, and demonstrated to remain similar to the original version of the algorithm, even when multiple failures arise in a short interval. In future works, models could be designed to investigate the number of protective checksum blocks that optimizes the runtime expectation considering the trade-off between the cost of extra computation and the increased probability of multiple failures when the number of nodes is very large.

## REFERENCES

- G. Bosilca, R. Delmas, J. J. Dongarra, and J. Langou. 2009. Algorithm-based fault tolerance applied to high performance computing. *J. Parallel Distrib. Comput.* 69, 4, 410–416.
- A. Bouteiller, G. Bosilca, and J. J. Dongarra. 2010. Redesigning the message logging model for high performance. *Concurrency Computat.: Practice Exp.* 22, 16, 2196–2211. DOI:<http://dx.doi.org/10.1002/cpe.1589>.
- G. Bums, R. Daoud, and J. Vaigl. 1994. LAM: An open cluster environment for MPI. In *Proceedings of the Conference on Supercomputing (SC'94)*. IEEE/ACM, 379–386.
- F. Cappello. 2009. Fault tolerance in petascale/exascale systems: Current knowledge, challenges and research opportunities. *Int. J. High Perform. Comput. Appl.* 23, 3.
- Z. Chen and J. J. Dongarra. 2006a. Algorithm-based checkpoint-free fault tolerance for parallel matrix computations on volatile resources. In *Proceedings of the 20th International Conference on Parallel and Distributed Processing (IPDPS'06)*. IEEE, 97. <http://dl.acrn.org/citation.cfrn?id=1898953.1899028>.
- Z. Chen and J. J. Dongarra. 2006b. Scalable techniques for fault tolerant high performance computing. Ph.D. Dissertation, University of Tennessee, Knoxville, TN.
- Z. Chen and J. J. Dongarra. 2008. Algorithm-based fault tolerance for fail-stop failures. *IEEE Trans. Parallel Distrib. Syst.* 19, 12, 1628–1641.
- J. Choi, J. Demrnel, I. Dhillon, J. J. Dongarra, S. Ostrouchov, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley. 1996. ScaLAP ACK: A portable linear algebra library for distributed memory computers-design issues and performance. *Comput. Phys. Commun.* 97, 1–2, 1–15.
- M. Cosnard, J.-M. Muller, and Y. Robert. 1986. Parallel QR decomposition of a rectangular matrix. *Numer. Math.* 48, 2, 239–249. DOI:<http://dx.doi.org/10.1007/BF01389871>.

- T. Davies, C. Karlsson, H. Liu, C. Ding, and Z. Chen. 2011. High performance linpack benchmark: A fault tolerant implementation without checkpointing. In *Proceedings of the International Conference on Supercomputing (ICS'11)*. ACM, New York, 162–171. DOI:<http://dx.doi.org/10.1145/1995896.1995923>.
- J. Demmel, L. Grigori, M. Hoemmen, and J. Langou. 2012. Communication-optimal parallel and sequential QR and LU factorizations. *SIAM J. Sci. Comput.* 34, 1, 206–239. DOI:<http://dx.doi.org/10.1137/080731992>.
- P. Du, A. Bouteiller, G. Bosilca, T. Herault, and J. J. Dongarra. 2012. Algorithm-based fault tolerance for dense matrix factorizations. In *Proceedings of the 17th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP'12)*. ACM, New York, 225–234. DOI:<http://dx.doi.org/10.1145/2145816.2145845>.
- E. N. Elnozahy, D. B. Johnson, and W. Zwaenepoel. 1992. The performance of consistent checkpointing. In *Proceedings of the 11th IEEE Symposium on Reliable Distributed Systems*. 39–47. DOI:<http://dx.doi.org/10.1109/RELDIS.1992.235144>.
- G. E. Fagg, E. Gabriel, G. Bosilca, T. Angskun, Z. Chen, J. Pjesivac-grbovic, K. London, and J. J. Dongarra. 2004. Extending the MPI Specification for process fault tolerance on high performance computing systems. In *Proceedings of the 19th International Supercomputer Conference*.
- G. H. Golub and C. F. Van Loan. 1996. *Matrix Computations* 3rd Ed. Johns Hopkins University Press, Baltimore, MD.
- D. Hakkarinen and Z. Chen. 2010. Algorithmic Cholesky factorization fault recovery. In *Proceedings of the IEEE International Symposium on Parallel Distributed Processing*. IEEE, 1–10. DOI:<http://dx.doi.org/10.1109/IPDPS.2010.5470436>.
- K. H. Huang and J. A. Abraham. 1984. Algorithm-based fault tolerance for matrix operations. *IEEE Trans. Comput.* 100, 6, 518–528.
- D. S. Katz, J. Daly, N. DeBardleben, E. N. Elnozahy, B. Kramer, S. Lathrop, N. Nystrom, K. Milfeld, S. Sanielevici, S. Scott, and L. Votta. 2009. Fault tolerance for extreme-scale computing workshop report. Tech. Rep. ANL/MCS-TM-312, Argonne National Lab., Albuquerque, NM. [http://www.teragridforum.org/mediawiki/images/8/8c/FT\\_workshop\\_report.pdf](http://www.teragridforum.org/mediawiki/images/8/8c/FT_workshop_report.pdf).
- V. Kumar, A. Grama, A. Gupta, and G. Karypis. 1994. *Introduction to Parallel Computing: Design and Analysis of Algorithms*. Benjamin-Cummings Publishing Co., Redwood City, CA.
- Lu, C. 2005. Scalable diskless checkpointing for large parallel systems. Ph.D. Dissertation, University of Illinois at Urbana-Champaign, IL.
- F. T. Luk and H. Park. 1988. An analysis of algorithm-based fault tolerance techniques. *J. Parallel Distrib. Comput.* 5, 2, 172–184.
- J. S. Plank, K. Li, and M. A. Puening. 1998. Diskless checkpointing. *IEEE Trans. Parallel Distrib. Syst.* 9, 10, 972–986.
- B. Schroeder and G. A. Gibson. 2007. Understanding failures in Petascale computers. *J. Physics: Conference Series* 78, 1, 012022. <http://stacks.iop.org/1742-6596/78/i=1/a=012022>.
- F. H. Streitz, J. N. Glosli, M. V. Patel, B. Chan, R. K. Yates, B. R. Supinski, J. Sexton, and J. A. Gunnels. 2006. Simulating solidification in metals at high pressure: The drive to petascale computing. *J. Physics: Conference Series* 46, 254.

Received July 2013; revised April 2014; accepted June 2014