Fault Tolerant Matrix Operations for Networks of Workstations Using Diskless Checkpointing

James S. Plank †

Youngbae Kim †

Jack J. Dongarra †‡

† Department of Computer Science
University of Tennessee
Knoxville, TN 37996
[plank,youngbae,dongarra]@cs.utk.edu

‡ Mathematical Science Section
Oak Ridge National Laboratory
PO Box 2008, Building 6012
Oak Ridge, TN 37821-6367
dongarra@msr.epm.ornl.gov

Submitted for Publication. February, 1996

Fault Tolerant Matrix Operations for Networks of Workstations Using Diskless Checkpointing

James S. Plank† Youngbae Kim† Jack J. Dongarra†‡

†Department of Computer Science, University of Tennessee

‡Mathematical Science Section, Oak Ridge National Laboratory

Abstract

Networks of workstations (NOWs) offer a cost effective platform for high-performance, long-running parallel computations. However, these computations must be able to tolerate the changing and often faulty nature of NOW environments. We present high-performance implementations of several fault-tolerant algorithms for distributed scientific computing. The fault-tolerance is based on diskless checkpointing, *a paradigm that uses processor redundancy rather than stable storage as the fault-tolerant medium. These algorithms are able to run on clusters of workstations that change over time due to failure, load or availability. As long as there are at least n processors in the cluster, and failures occur singly, the computation will complete in an efficient manner. We discuss the details of how the algorithms are tuned for fault-tolerance and present the performance results on a PVM network of Sun workstations connected by a fast, switched ethernet.*

1 Introduction

Scientific computation has been a driving force behind parallel and distributed computing. Traditionally such computations have been performed on the largest and most expensive supercomputers: the Cray C90, Intel Paragon, and Maspar MP-2. Recently the price and performance of uniprocessor workstations and off-the-shelf networking has improved to the point that networks of workstations (NOWs) provide a parallel processing platform that is competitive with the supercomputers. The popularity of NOW programming environments like PVM [19] and MPI [32, 40], and the availability of high-performance libraries for scientific computing on NOWs like ScaLAPACK [13] show that networks of workstations are already in heavy use for scientific programming.

The major problem with programming on a NOW is the fact that it is prone to change. Idle workstations may be available for computation at one moment, and gone the next due to failure, load or ownership. We term any such event a failure. Thus, on the wish list of scientific programmers is a way to perform computation on a NOW whose components may change over time.

This paper provides a solution to this problem, especially tailored to the needs of scientific programmers. The solution is based on *diskless checkpointing*, a means of providing fault-tolerance without any dependence on disk. The end result is that as long as there are n processors available in the NOW (where n is defined by the user), and as long as failures come singly, the computation can progress reliably.

We describe our approach of incorporating diskless checkpointing in four well-known algorithms in linear algebra: Cholesky factorization, LU factorization, QR factorization and Preconditioned Conjugate Gradient (PCG) [4, 16]. Subroutines like these at are the heart of scientific computations. We show the performance of these subroutines on a cluster 17 Sun Sparc5 workstations connected by a fast (100 megabit) switched ethernet.

The importance of this work is that it demonstrates a novel technique of executing high-performance scientific computations on a changing pool of resources.

2 Supercomputers vs. NOWs

A supercomputer is a single computing resource. We usually think of each processor in a supercomputer as being identical — every node is a uniform part of the whole. Typically, a supercomputer is allocated exclusively for a single application, such as a grand challenge. If it can be partitioned, then each partition is allocated exclusively. The file system is often implemented using special disks and processors at the periphery of the supercomputer so that files are uniformly available, regardless of the partition being used. If one processor or part of the network fails, the whole computational platform is rendered useless until the faulty part is fixed.

For this reason, fault-tolerance in supercomputers is straightforward. *Consistent checkpointing* can be used to save the state of a parallel program. In consistent checkpointing, all processors cooperate to save a global checkpoint in stable storage. This checkpoint is composed of uniprocessor checkpoints for every processor in the system, and a log of messages that are in transit during checkpointing. Many algorithms exist for taking consistent checkpoints [11, 25, 29] and implementations have shown that the simplest of these, a two-phase commit called "Sync-and-stop" yields performance on par with the most complex [36]. Checkpointing performance is dependent on the size of the individual checkpoints, the speed of the file system, and the amount of physical memory available for buffering [17, 36]. These conclusions are not likely to change as new machines are released unless the model of exclusive node partitioning and wholesale partition failures is changed.

In contrast, a NOW is a distributed computing resource that is highly shared. Processors usually run a general-purpose time-sharing operating system, and each is often owned by a different user. Although the processing capacity of the NOW as a whole may be consistently large, individual processors can run the gamut from idle to unavailable (e.g. in use by the owner) back to idle in a relatively small time frame [33]. Programs for NOWs are generally written using some NOW programming environment like PVM [19] that provides convenient primitives for message passing. Such programming environments allow individual processors to enter or leave the NOW dynamically due to availability, load or failure. We term all such events failures. Thus, NOWs present a far more flexible failure model than supercomputers.

In such systems, consistent checkpointing to disk is overkill. If one processor becomes unavailable, the whole collection of processors must restart themselves from stable storage. Moreover, if the failed processor cannot be brought back online, then its checkpoint file will be unavailable unless it has been saved on a central file server which will then be a source of contention during checkpointing [28]. Therefore, a more relaxed model of checkpointing is needed — one that is tailored to the dynamic nature of NOWs.

3 A Model for Scientific Programs that Live on a NOW

Ideally, a scientific program executing on a NOW should be able to "live" on whatever pool of processors is currently available. Processors should be able to leave the NOW whenever they fail, and they should be able to join the NOW when they become functional. We describe a model of scientific computation that approaches this ideal.

We assume that we are running a high-performance scientific program, like electromagnetic scattering or atomic structure calculation. The bulk of the work in such programs is composed of well known subproblems: solving partial differential equations and linear systems. These subproblems are typically solved using high-performance libraries, such as ScaLAPACK [13], which are designed to get maximum performance out of the computing platform. An important performance consideration is *domain decomposition*, which is how the problem is partitioned among the available processors to minimize cache misses and the effects of message transmission. To perform domain decomposition properly, the number of processors is usually fixed at some n, often a perfect square or power of two.

To retain high-performance, we assume that the program is optimized to run on exactly n processors. Our computing platform is assumed to be a NOW, which can contain any number of processors at any one time. Our model of computation for fault-tolerance is as follows.

Whenever the NOW contains at least n processors, the computation should be running on n of the processors. Whenever the NOW contains fewer than n processors, the computation is *swapped off* the NOW. This can be done by some sort of consistent checkpointing scheme that saves a global checkpoint to a central file server at very coarse intervals (for example once every hour or day). Such checkpointing schemes are straightforward and have been discussed and implemented elsewhere [10, 17, 18, 25, 28, 36, 41].

Whenever the NOW contains *more than* n processors, then the computation should be running in such a manner that if any processor that is running the computation drops out of the NOW, due to failure, load or ownership, then it can be replaced quickly by another processor in the NOW. This is the important part of the computing model, because it means that as long as the pool of processors in the NOW numbers greater than n members, then even if the pool itself changes, the computation should be progressing efficiently, while still maintaining fault-tolerance to wholesale failures.

4 The Checkpointing Algorithm

The algorithm is based on *diskless checkpointing* [35]. If the program is executing on n processors, then there is a $n+1$-st processor called the *parity processor*. At all points in time, a consistent checkpoint is held in the n processors in memory. Moreover, the bitwise exclusive-or ($\oplus$) of the n checkpoints is held in the parity processor. This is called the *parity checkpoint*. If any processor fails, then its state can be reconstructed on the parity processor as the exclusive-or of the parity checkpoint and the remaining $n-1$ processors' checkpoints.

Diskless checkpointing has been shown to be effective at providing fault-tolerance for single processor failures as long as there is enough memory to hold single checkpoints in memory. To reduce the memory requirements, incremental checkpointing can be used, and compression can be helpful in reducing the load on network bandwidth [35].

To make checkpointing as efficient as possible, we implement algorithm-based checkpointing. In other words, rather than implement checkpointing transparently as in MIST [10], Fail-Safe PVM [28] or CoCheck [41], we hard-wire it into the program. This is beneficial for several reasons. First, the checkpointing can be placed at synchronization points in the program, which means that checkpoint consistency (defining network state [11]) is not a worry. Second, the checkpointed state can be minimized because the checkpointer knows exactly what to save and how to reconstruct state. This is as opposed to a transparent checkpointer that must save all program state because it knows nothing about the program. Third, with transparent checkpointing, checkpoints are binary memory dumps, which rules out a heterogeneous recovery. With algorithm-based checkpointing, the recovery routines can plan for recovery by a different kind of processor. In short, algorithm-based checkpointing is good because it enables the checkpointing to be as efficient as possible [27]. Its major drawback is programmer effort, since the fault-tolerance must be encorporated carefully into the program. However, if the algorithms being checkpointed can be put into frequently-used library calls, then the extra work is justifiable [39].

5 Checkpointing High-Performance Distributed Matrix Operations

We focus on two classes of matrix operations: direct, dense factorizations and an iterative equation solver. The factorizations (Cholesky, LU and QR) are operations for solving systems of simultaneous linear equations and finding least squares solutions of linear systems. All have been implemented in LAPACK [1] and ScaLAPACK [13], which are public-domain libraries providing high-performance implementations of linear algebra operations for uniprocessors and all kinds of parallel processing platforms. The iterative equation solver called Preconditioned Conjugate Gradient (PCG) is a well-known technique to solve sparse systems of linear equations [4].

We have implemented fault-tolerant versions of Cholesky, LU, QR, and PCG. In the sections that follow, we provide an overview of how each operation works, and how we make it fault-tolerant. Further details on the ScaLAPACK implementations may be found in books by Dongarra [16], and Golub [22].

5.1 Cholesky Factorization and the Basic Checkpointing Scheme

Of the three factorizations, Cholesky is the simplest. In Cholesky factorization, a dense, symmetric, positive definite matrix A is factored into two matrices L and L^T (i.e. $A = LL^T$) such that L is lower triangular. The algorithm for performing Cholesky factorization in ScaLAPACK is called "top-looking", and works as follows.

First, the matrix A is partitioned into square "blocks" of a user-specified block size b. Then A is distributed among the processors P_0 through P_{n-1}, logically reconfigured as a $p \times q$ mesh, as in Figure 1. For obvious reasons, a row of blocks is called a "row-block" and a column of blocks is called a "column-block." If there are N processors and A is an $N \times N$ matrix, then each processor holds $\frac{N}{bp}$ row-blocks and $\frac{N}{bq}$ column-blocks, where it is assumed that b, p and q divide N.

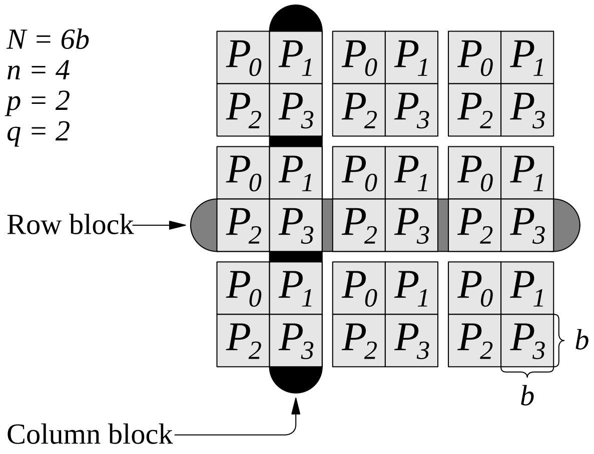
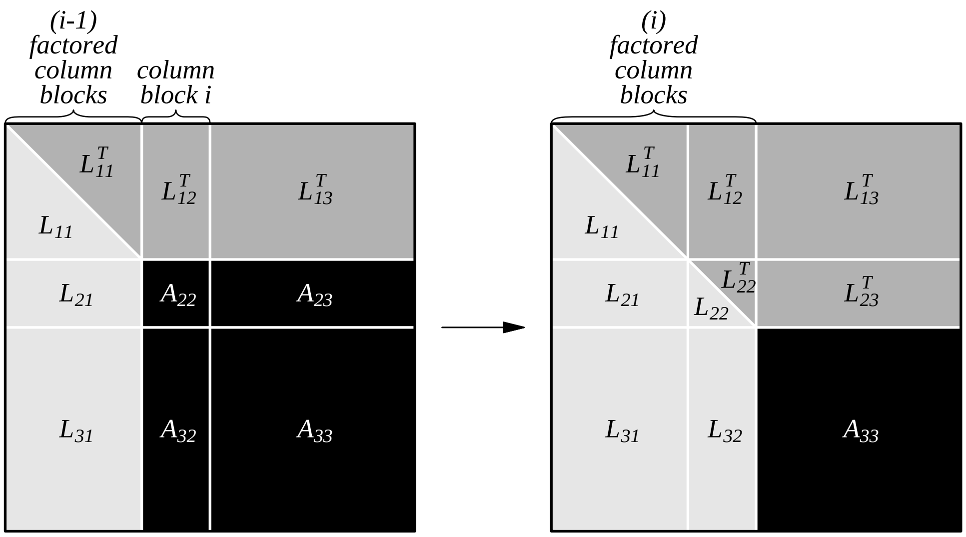


Figure 1: Data distribution of a matrix with 6×6 blocks over a 2×2 mesh of processors

The factorization of A is performed in place, and proceeds in N/b steps, one for each column-block of the matrix. At the beginning of step i, the leftmost $i-1$ column-blocks are assumed to be factored, and the remaining column-blocks are unchanged. In step i, the i-th column-block gets factored using a multiplication, subtraction and factorization.

Thus, each step looks as in Figure 2. Inherent in this picture is the communication — for example, to perform $\hat{A}_{22} \leftarrow A_{22} - L_{21}L_{12}^T$, all the involved blocks must be sent to the processor holding A_{22}. Note also that Figure 2 is a logical representation of the system. Since A is symmetric and L^T is the transpose of L, only half of A and none of L^T need to be stored.

The key fact to notice from Figure 2 is that at step i, only A_{22} and A_{32} get modified. The rest of the blocks in the factorization remain the same. Thus, only blocks from column-block i are modified during step i.



(I – Multiplication/Subtraction) $$\begin{pmatrix} \hat{A}_{22} \\ \hat{A}_{32} \end{pmatrix} \leftarrow \begin{pmatrix} A_{22} \\ A_{32} \end{pmatrix} - \begin{pmatrix} L_{21} \\ L_{31} \end{pmatrix} L_{12}^T$$

(II – Factorization) $$\begin{pmatrix} \hat{A}_{22} \\ \hat{A}_{32} \end{pmatrix} = \begin{pmatrix} L_{22} \\ L_{32} \end{pmatrix} L_{22}^T$$

Figure 2: Step i for Cholesky factorization

To make the Cholesky factorization fault-tolerant, we first allocate a parity processor P_n. For each panel of n blocks in the matrix, there is one block in P_n containing the bitwise exclusive-or of each block in the panel. This is depicted in Figure 3 for the example system of Figure 1.

Each processor P_j (including P_n) allocates room for an extra column-block called CB_j. Now, the algorithm for performing fault-tolerant Cholesky factorization is as follows:

- Initialize the global state of the system.
- For each step i:
  - Let P_j be a processor with blocks in column-block i. P_j copies these blocks to CB_j.
  - P_n also copies its blocks corresponding to blocks in column-block i to CB_n.
  - The processors perform step i.
  - The processors P_j $(0 \leq j < n)$ cooperate with P_n to update the exclusive-or for the newly-modified blocks in column-block i.
  - The processors synchronize, and go to step $i + 1$.

Thus, at the beginning of each step, the processors hold the state of the factorization as depicted in Figure 3. If any one processor P_j fails, then it can be replaced by P_n, or by a new processor. This new processor calculates

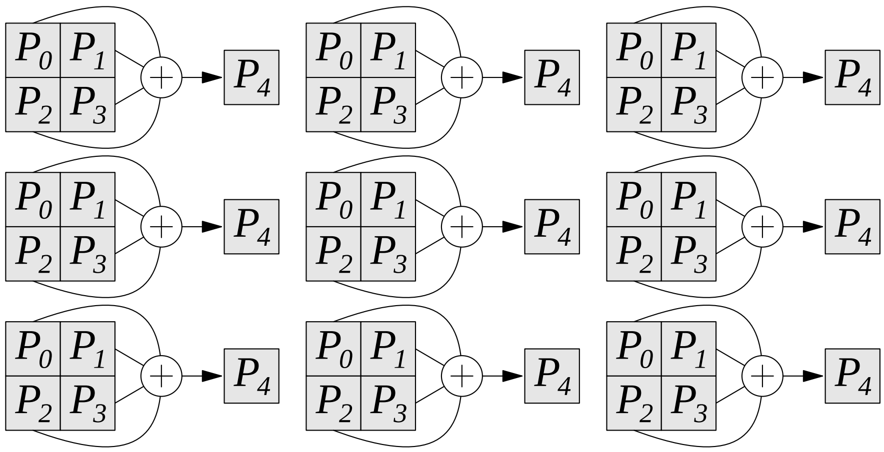


Figure 3: Configuring the system for checkpointing

P_j's state from the bitwise exclusive-or of the remaining processors. Obviously, P_n can be replaced in a similar manner.

If any one processor P_j fails in the middle of a step, then the remaining processors can roll back to the beginning of the step by copying CB back to column-block i. Then P_j can be recovered as described in the preceding paragraph.

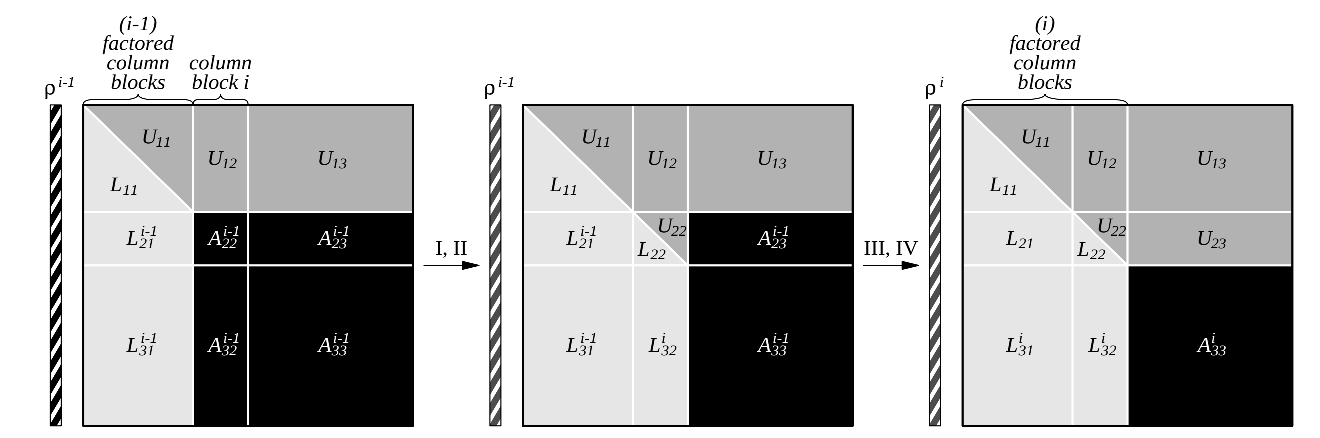
It is assumed here that failure detection is provided by the computing platform. For example, PVM detects processor and certain network failures, and a resource manager like CARMI [37] can be added to PVM to detect failures due to load and ownership.

5.2 LU Factorization

In LU factorization, a dense matrix A is factored using a sequence of elementary eliminations with pivoting such that $\rho A = LU$, where L is a lower triangular matrix with 1's on the diagonal and U is an upper triangular matrix. ρ is a permutation matrix necessary for numerical stability: a proper permutation of the rows of A minimizes the growth of roundoff error during the elimination. LU factorization involves a general non-symmetric matrix, and is computationally more complex than Cholesky factorization.

There are three well-known algorithm variants for implementing LU factorization on parallel machines: left-looking, right-looking and Crout. These variants differ in the regions of data that are accessed and computed during each step (see [16] for details). Below, we describe the Crout variant and how it is checkpointed. We discuss the ramifications of the algorithm selection and checkpointing performance in section 7.3.

Like Cholesky factorization, LU factorization is performed in place, replacing A with L and U. Moreover, the permutation matrix ρ is included as output to the subroutine. Since a permutation matrix is simply the identity



(I – Multiplication/Subtraction)
$$\begin{pmatrix} A_{22}^{i} \\ A_{32}^{i} \end{pmatrix} \leftarrow \begin{pmatrix} A_{22}^{i-1} \\ A_{32}^{i-1} \end{pmatrix} - \begin{pmatrix} L_{21}^{i-1} \\ L_{31}^{i-1} \end{pmatrix} U_{12}$$

(II – Factorization / Definition of $\hat{\rho}$)
$$\hat{\rho} \begin{pmatrix} A_{22}^{i} \\ A_{32}^{i} \end{pmatrix} = \begin{pmatrix} L_{22} \\ L_{32}^{i} \end{pmatrix} U_{22}$$

(III – Pivoting)
$$\begin{pmatrix} L_{21}^{i} \\ L_{31}^{i} \end{pmatrix} \leftarrow \hat{\rho} \begin{pmatrix} L_{21}^{i-1} \\ L_{31}^{i-1} \end{pmatrix}, \quad \begin{pmatrix} A_{23}^{i} \\ A_{33}^{i} \end{pmatrix} \leftarrow \hat{\rho} \begin{pmatrix} A_{23}^{i-1} \\ A_{33}^{i-1} \end{pmatrix}, \quad \rho^{i} \leftarrow \hat{\rho}\rho^{i-1}$$

(IV – Solve/Multiplication/Subtraction)
$$U_{23} \leftarrow (L_{32}^{i})^{-1} A_{23}^{i} - L_{21} U_{12}$$

Figure 4: Step i for Crout LU factorization

matrix I with rows permuted, it may be represented by a one dimensional array, where the i^{th} entry contains the index of the non-zero element in row i of ρ. Like A, ρ is distributed among the processors. Each processor P_j contains its portion of ρ in ρ_j.

As before, the matrix is partitioned into blocks and distributed among the processors. The factorization proceeds in steps, one for each column/row block in A. In step i, the i-th column-block is factored, and the result of this factoring is used to factor the i-th row-block. The details are in Figure 4.

The memory update patterns in Crout LU are more complex than in Cholesky factorization. In step i, both column-block i and row-block i are modified. Moreover, the permutation matrix ρ is altered, and at most b rows in L_{31} and A_{33} are swapped with rows in row-block i due to pivoting. Thus, the algorithm to make Crout LU fault-tolerant, though similar to Cholesky factorization, is necessarily more complex.

To be specific, P_n starts as in Cholesky factorization, with blocks containing the exclusive-or of panels of blocks of A. Moreover, P_n has some memory ρ_n, which contains the bitwise exclusive-or of each processor's ρ_j. Each processor P_j (including P_n) allocates room for an extra column-block, CB_j, an extra row-block RB_j, and a cache of ρ_j called ρ'_j. Finally, each processor P_j (including P_n) allocates room for a row-block's worth of pivoting rows PR_j. The fault-tolerant LU factorization proceeds as follows:

- Initialize the global state of the system (including P_n).
- For each step i:
  - Let P_j be a processor with blocks in column-block i. P_j copies these blocks to CB_j.
  - Let P_j be a processor with blocks in row-block i. P_j copies these blocks to RB_j.
  - P_n copies its blocks corresponding to blocks in column-block i and row-block i to CB_n and RB_n.
  - All P_j $(0 \leq j \leq n)$ copy ρ_j to ρ'_j.
  - The processors perform substeps I and II of step i.
  - In substep III, b rows of the matrix are swapped with rows in row-block i. Before doing so, the processors P_j that own these rows copy them to PR_j. P_n copies its rows corresponding to these rows to PR_n.
  - Now, the processors perform substeps III and IV.
  - The processors P_j $(0 \leq j < n)$ cooperate with P_n to update the exclusive-or for the newly-modified blocks in column-block i, row-block i, the swapped pivot rows, and ρ.
  - The processors synchronize, and go to step $i+1$.

As in Cholesky factorization, if a processor P_j fails during step i of the computation it can be replaced by P_n, or by a new processor. The replacement proceeds as follows:

- For all remaining processors P_k (this includes P_n), if P_k had started substep III, then it copies any rows back from PR_k to their original position.
- All P_k copy their data from CB_k, RB_k and ρ'_k back to column-block i, row-block i, and ρ_j respectively.
- P_j's state is reconstructed from the bitwise exclusive or of the blocks in the other P_k.
- The computation proceeds from the beginning of step i.

5.3 QR Factorization

In QR factorization a real $M \times N$ matrix A is factored such that

$$A \;=\; Q \begin{pmatrix} R \\ 0 \end{pmatrix},$$

where Q is an $M \times N$ orthogonal matrix and R an $N \times N$ upper triangular matrix. In the ScaLAPACK implementation of the QR factorization, the matrix Q is not generated explicitly since it would require too much extra storage. Instead, Q can be applied or manipulated through the identity $Q = I - VTV^T$, where V is a lower triangular matrix of "Householder" vectors, and T is an upper triangular matrix constructed from information in V. When the factorization is complete, the matrix A is transformed into V, T and R, where V is in the lower triangle of the original matrix A, R is in the upper triangle, and T is stored in a one-dimensional array.

Like LU factorization, there are multiple algorithms for QR factorization. We focus on the left-looking algorithm. Complete details of the implementation of this algorithm are beyond the scope of this paper, but may be found in Dongarra's book [16]. A high-level picture is provided in Figure 5.

It should be clear from Figure 5 that only column-block i of matrix A is changed during factoring step i. Therefore the fault-tolerant version of QR works exactly like the fault-tolerant version of Cholesky — each

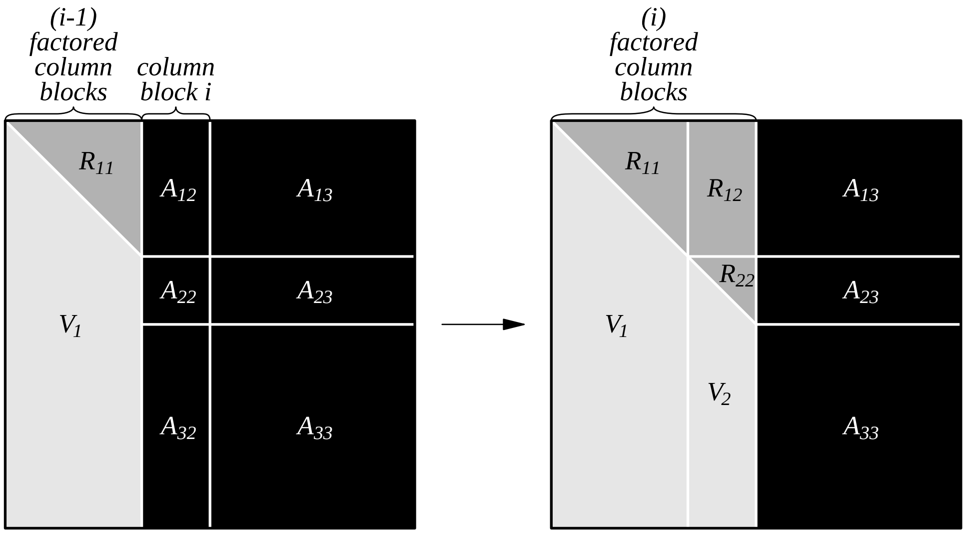


Figure 5: Step i of QR factorization

processor P_j allocates an extra column-block CB_j to hold the initial value of column-block i during step i, so that the computation can be rolled back to the beginning of step i if there is a failure.

5.4 Iterative equation solver (PCG)

Iterative equation solvers are used for the following problem. Given a large sparse matrix A and a vector b, find the vector x such that $Ax = b$. Iterative equation solvers work as follows. Given an initial approximation to x, the method iteratively refines this approximation until $Ax = b$ to within some error tolerance. Unfortunately, no single iterative method is robust enought to solve all sparse linear systems accurately and efficiently. Therefore, we limit our scope to one such method, known as "Preconditioned Conjugate Gradient" (PCG).

If A is positive definite symmetric, then PCG can be used to solve the system $Ax = b$ by projecting A onto a "Krylov subspace" and then solving the system in this subspace. The details of the algorithm are beyond the scope of this paper [4, 16, 22]. However its mechanics as they impact fault-tolerance are simple. First, the sparse matrix A is represented in a dense form, and is then distributed along with b and two preconditioners M_1 and M_2 to the processors P_0 through P_{n-1}. M_1 and M_2 are diagonal matrices and thus may be represented by linear arrays. After this point, A, b, M_1 and M_2 are not altered.

Now, the vectors p_0, r_0, w_0 and ξ_0 are calculated from A, b, M_1 and M_2. These intermediate vectors are used to calculate the vector x_0, which is the first approximation to x. The algorithm then iterates as follows: The values of A, b, M_1, M_2, x_{i-1}, p_{i-1}, r_{i-1}, w_{i-1} and ξ_{i-1} are used to calculate p_i, r_i, w_i and ξ_i. These are then used to calculate x_i, the i-th approximation to x. The iterations continue until $Ax_i = b$ to within a given error tolerance.

Adding fault-tolerance to the PCG algorithm is straightforward. First, the processors distribute A, b, M_1 and

M_2, and allocate memory for x_i, p_i, r_i, w_i, and ξ_i. The extra processor P_n is initialized to contain the bitwise exclusive-or of all these variables. Now, each processor (including P_n) must include extra vectors for each of x, p, r, w and ξ. These extra vectors are maintained like CB in the factorization examples. They hold the values of x_{i-1}, p_{i-1}, r_{i-1}, w_{i-1} and ξ_{i-1} during step i so that the step can be rolled back following a failure.

Note that in PCG, we can checkpoint every k steps by copying x_i, p_i, r_i, w_i and ξ_i to the extra vectors and computing the bitwise exclusive-or of x, p, r, w and ξ only when i is a multiple of k. The result is that processors may roll back up to k steps to the previous checkpoint upon a failure. However, since checkpoints are only taken every k steps, the overhead of checkpointing will be reduced by a factor of k.

6 Implementation Results

We implemented and executed these programs on a network of Sparc-5 workstations running PVM [19]. This network consists of 24 workstations, each with 96 Mbytes of RAM, connected by a switched 100 megabit ethernet. The peak measured bandwidth in this configuration is 40 megabits per second between two random workstations. These workstations are generally allocated for undergraduate classwork, and thus are usually idle during the evening and busy executing I/O-bound and short CPU-bound jobs during the day. We ran our experiments on these machines when we could allocate them exclusively for our own use.

The results presented here are for a network of 17 processors, where 16 are running the program ($n = 16$, $p = q = 4$) and one is calculating the parity. We ran three sets of tests for each instance of each problem. In the first there is no checkpointing. In the second, the program checkpoints, but there are no failures, and in the third, a processor failure is injected randomly to one of the processors, and the program completes with 16 processors. In the results that follow, we present only the time to perform the recovery, since there is no checkpointing after recovery.

6.1 Cholesky Factorization

We ran ten different instances of the Cholesky factorization, one for each of ten matrix sizes from $N = 1,000$ to $N = 10,000$. In each run, the block size was 50. The data for this experiment is in Figure 6.

As displayed in the leftmost graph of Figure 6, Cholesky factorization has a running time of $O(N^3)$. The total overhead of checkpointing consists of the following portions:

- *The time to take the initial checkpoint of matrix A* (T_{init}). In our calculations below, we assume that message bandwidth dominates the overhead of message-passing enough that we can ignore message latency. Each entry of A is a double precision floating point number (8 bytes). As stated in section 5.1 above, since A is symmetric, only half of it needs to be stored. Therefore the total amount of storage needed for A is $4N^2$ bytes. These bytes are distributed evenly among the n processors, which perform the `XOR` using a binary

Matrix Size		Running	With Checkpointing							Recovery
N	Total Size (Mbyte)	Time (sec)	# of Check-points	Checkpoint Interval (sec)	Running Time (sec)	Total Overhead (sec)	T_{init} (sec)	T_{rest}: Avg. Per Ckp. (sec)	Overhead %	Time (sec)
1000	4	12	20(+1)	0.7	15	3	2	0.1	25.0	2
2000	16	58	40(+1)	1.6	70	12	6	0.1	20.7	6
3000	36	163	60(+1)	2.9	189	26	14	0.2	16.0	13
4000	64	331	80(+1)	4.4	379	48	26	0.3	14.5	23
5000	100	602	100(+1)	6.4	682	80	38	0.4	13.3	36
6000	144	992	120(+1)	8.8	1112	120	52	0.6	12.1	51
7000	196	1516	140(+1)	11.5	1681	165	70	0.7	10.9	69
8000	256	2244	160(+1)	14.9	2494	250	102	0.9	11.1	98
9000	324	3057	180(+1)	18.0	3367	310	134	1.0	10.1	132
10000	400	4112	200(+1)	21.7	4489	377	152	1.1	9.2	148

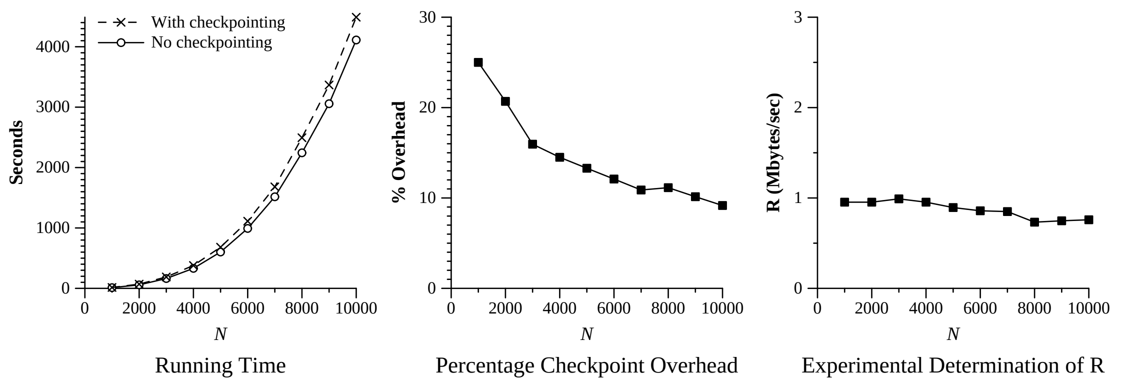


Figure 6: Results for Cholesky factorization

tree algorithm. Thus, the first checkpoint takes $\frac{4N^2(\log n)}{nR}$ seconds, where R is the rate of sending a message and XOR-ing it, expressed in bytes per second.

- *The time to take the column-block checkpoints* (T_{rest}). There are $\frac{N}{b}$ of these checkpoints, with an average size of $4bN$. Since p processors cooperate for each of these checkpoints, the total overhead of these checkpoints is $\frac{N}{b}\left(4bN\frac{(\log p)}{pR}\right) = \frac{4N^2(\log p)}{pR}$.

Thus, the total overhead of checkpointing is:

$$OV_{\text{cholesky}} = \frac{4N^2}{R}\left(\frac{\log n}{n} + \frac{\log p}{p}\right). \tag{1}$$

Since OV_{cholesky} is $O(N^2)$ and Cholesky factorization is $O(N^3)$, we expect the percentage overhead of checkpointing to decrease as N increases. This is plotted in the middle graph of Figure 6.

The rightmost graph of Figure 6 plots R as determined by equation (1) for each value of N. Since the peak observed network performance is 40 megabits per second, we expect that R will be somewhat lower than 5 Mbytes/sec to take account of synchronization and the XOR time. This is shown to be the case.

Matrix Size		Running	With Checkpointing							Recovery
	Total Size	Time	# of Check-	Checkpoint Interval	Running Time	Total Overhead	T_{init}	T_{rest}: Avg. Per Ckp.	Overhead %	Time
N	(Mbyte)	(sec)	points	(sec)	(sec)	(sec)	(sec)	(sec)		(sec)
1000	8	56	20(+1)	3.3	66	10	3	0.3	17.9	3
2000	32	190	40(+1)	5.5	221	31	12	0.5	16.3	11
3000	72	451	60(+1)	8.7	519	68	25	0.7	15.1	24
4000	128	853	80(+1)	12.2	978	125	50	0.9	14.7	44
5000	200	1509	100(+1)	17.1	1715	206	79	1.3	13.7	69
6000	288	2294	120(+1)	21.7	2605	311	104	1.7	13.6	99
7000	392	3545	140(+1)	28.4	3978	433	148	2.0	12.2	136
8000	512	4959	160(+1)	34.5	5521	562	195	2.3	11.3	178
9000	648	6877	180(+1)	42.4	7627	750	253	2.8	10.9	248
10000	800	8850	200(+1)	49.0	9793	943	295	3.2	10.7	283

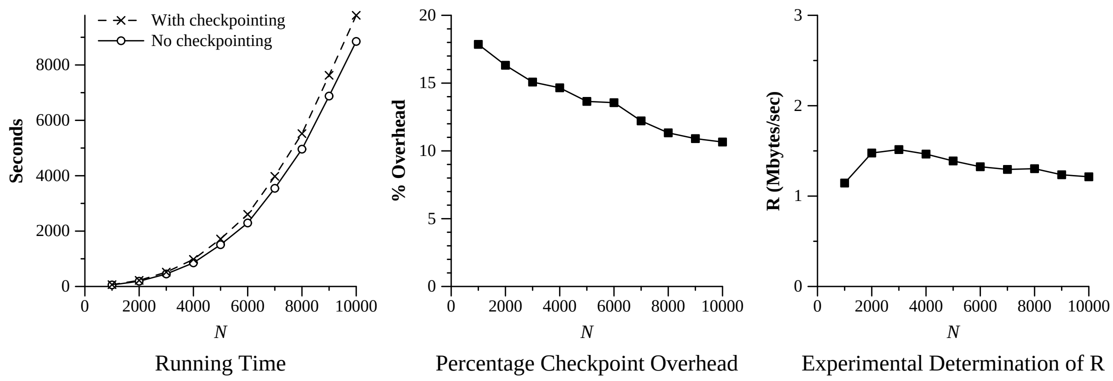


Figure 7: Results for LU factorization.

Recovery consists of taking the bitwise exclusive-or of every processor's matrix A. Thus, the overhead of recovery should equal T_{init}, which is reflected in the last column of Figure 6. Notice that the time it takes to recover is irrespective of the location of the failure.

6.2 LU Factorization

The results from the LU factorization are in Table 7. Again, the block size was 50. The results are very similar to the results from the Cholesky factorizations. Like Cholesky, Crout LU consumes $O(N^3)$ floating point operations but its constants are greater (by a factor of two), resulting in longer running times.

Since the matrix is not symmetric, the first checkpoint of A takes twice as long as in Cholesky factorization. Moreover, the calculation of T_{rest} is more complex. Each checkpoint consists of a portion of a column block, two row-blocks, and the pivot vector ρ. The overhead of sending the row-blocks is $\frac{16Nb(\log q)}{qR}$, and the overhead of sending ρ is $\frac{4b}{R}$, because only b elements of ρ are altered per iteration, and these b elements are contained entirely in one processor. The average size of L_{32}^i is $\frac{N}{2b}$. Therefore the average overhead of sending the column block is

Matrix Size		Running	With Checkpointing							Recovery
	Total Size	Time	# of Check-	Checkpoint Interval	Running Time	Total Overhead	T_{init}	T_{rest}: Avg. Per Ckp.	Overhead %	Time
N	(Mbyte)	(sec)	points	(sec)	(sec)	(sec)	(sec)	(sec)		(sec)
1000	8	376	20(+1)	19.4	388	12	5	0.3	3.2	4
2000	32	1210	40(+1)	31.1	1246	36	14	0.6	3.0	12
3000	72	2909	60(+1)	49.8	2988	79	32	0.8	2.7	29
4000	128	6052	80(+1)	77.5	6197	145	55	1.1	2.4	52
5000	200	10934	100(+1)	111.6	11162	228	74	1.5	2.1	70

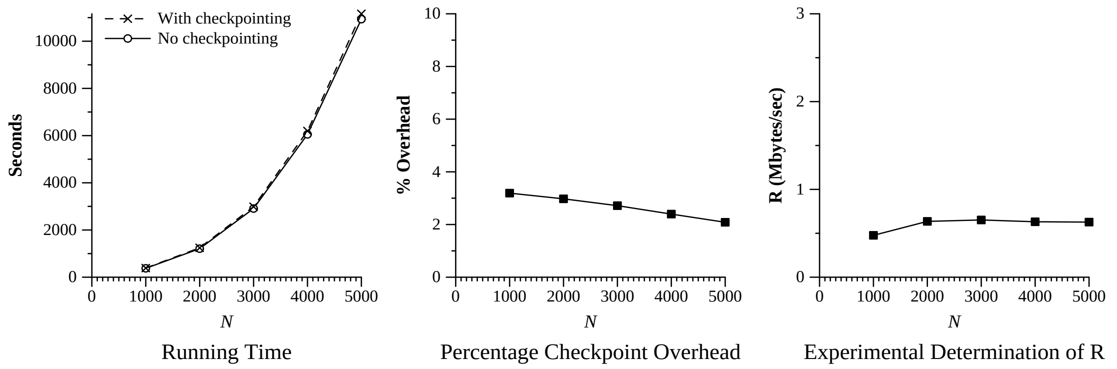


Figure 8: Results for QR factorization.

$\frac{4N(\log p)}{bpR}$. This yields the following equation for the overhead of checkpointing the Crout LU factorization:

$$OV_{\text{crout}} = \frac{4N}{R} + \frac{4N^2}{R}\left(\frac{(\log p)}{p} + \frac{4(\log q)}{q} + \frac{2(\log n)}{n}\right). \tag{2}$$

The rightmost graph of Figure 7 shows that equation (2) yields similar values of R as Cholesky factorization. The recovery time once again is roughly equal to T_{init}.

6.3 QR Factorization

The results from the QR factorization are in Table 8. Once more, the block size was 50. QR factorization is another $O(N^3)$ algorithm whose constants are greater than the LU factorizations. As such, only five values of N were tested because of the large running times.

QR checkpointing is exactly like Cholesky checkpointing, except that all of A is checkpointed initially, and every column-block is checkpointed in its entirety. Therefore, the overhead of QR checkpointing is exactly twice that of Cholesky checkpointing:

$$OV_{\text{QR}} = \frac{8N^2}{R}\left(\frac{\log n}{n} + \frac{\log p}{p}\right). \tag{3}$$

QR factorization has the lowest checkpointing overhead percentage of all the factorizations.

6.4 Preconditioned Conjugate Gradient

k	Running Time (sec)	With Checkpointing: # of Check-points	Checkpoint Interval (sec)	Running Time (sec)	Total Overhead (sec)	T_{init} (sec)	T_{rest}: Avg. Per Ckp. (sec)	Overhead %	Recovery Time (sec)
25	5789	200(+1)	47	9436	3647	28.9	18.1	63.0	41
50	5789	100(+1)	76	7607	1818	28.6	17.9	31.4	40
100	5789	50(+1)	134	6746	957	29.7	18.5	16.5	41
500	5789	10(+1)	600	6004	215	29.7	18.5	3.7	40
1000	5789	5(+1)	1181	5909	120	29.1	18.2	2.1	41
2500	5789	2(+1)	2926	5853	64	28.4	17.8	1.1	41
5000	5789	1(+1)	5835	5835	46	28.3	17.7	0.8	42

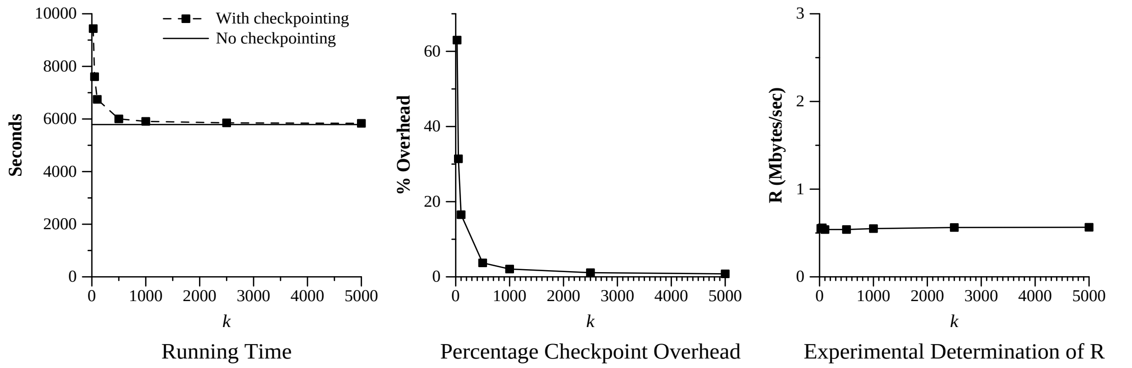


Figure 9: Results for PCG

We executed an instance of PCG with a $N \times N$ matrix A for I iterations, where $N = 1,048,576$, and $I = 5,000$. This calculated x to within a tolerance of 10^{-8}. The results of this instance with varying values of k (iterations per checkpoint) is in Table 9.

As in the factorizations, the overhead of checkpointing is broken into two parts: T_{init}, which is the time to checkpoint A, b, M_1 and M_2, and T_{rest}, which accounts for all of the checkpoints of x, p, r, w and ξ. The dense representation of A is a $5 \times N$ matrix, yielding a value of $\frac{64N(\log n)}{nR}$ for T_{init}. The remaining checkpoints comprise five vectors of length N. These should take $\frac{40N(\log n)}{nR}$ each. Since there are $\frac{I}{k}$ of these checkpoints, the overhead of checkpointing PCG is:

$$OV_{\text{PCG}} = \frac{8N(\log n)}{nR}\left(8 + \frac{5I}{k}\right). \tag{4}$$

To recover, matrix A and all eight vectors need to be reconstructed. Thus, the overhead of recovery time should be $\frac{104N(\log n)}{nR}$, which is the sum of the T_{init} and T_{rest} columns of Figure 9.

7 Discussion

7.1 Checkpointing Overhead and Interval

The results presented in the previous section show that on current NOWs, the performance of this method for fault-tolerant computation is surprisingly good. In the Cholesky and LU factorizations, checkpoints are taken less than a minute apart, yet the overhead is low — in the long-running instances, the total checkpointing overhead is under 15 percent. In the QR factorizations, the overhead of checkpointing is under two percent in all instances, while the checkpointing interval is less than two minutes. In all the factorizations, the overhead of checkpointing is $O(N^2)$, while the running time complexity is $O(N^3)$. Thus, the percentage of checkpointing overhead decreases as the problem size increases.

One interesting thing to notice is that there is no term for the block size b in equations (1), (2) and (3). This means that the block size, and therefore the checkpointing interval has little impact on the overhead of checkpointing. Also of interest is the fact that there is no term for q in equations (1) and (3). Thus, the best checkpointing performance in Cholesky and QR factorization should be realized when $p = n$ and $q = 1$. Of course, the selection of p and q also impacts the performance of the factorization [12].

In the PCG implementation, there is a tradeoff between the checkpointing interval and the overhead of checkpointing. This tradeoff is controlled by the variable k. In our example, the overhead of checkpointing is roughly $29 + 18\left(\frac{5000}{k}\right)$. Therefore, one can choose a value of k to achieve a desired checkpointing overhead or interval. For example, to achieve a 10 percent checkpointing overhead, one can choose k to be 164. This would yield roughly 578 seconds of checkpointing overhead and checkpoints every 207 seconds.

7.2 Extra Parity Processors

The choice of one parity processor P_n was made simply to present the concept of diskless checkpointing. If the NOW executing the computation contains $n + m$ processors, then there is no reason why $m - 1$ of them should be idle. Instead of having all n processors checkpoint to P_n, we can partition the n processors into m groups $G_0, \ldots G_{m-1}$, and have P_{n+j} be responsible for checkpointing the processors in G_j, for $0 \leq j < m$. This is basically a 1-dimensional parity scheme, which can tolerate up to m simultaneous processor failures, as long as each failure occurs in a different group [21].

The extreme we have presented is $m = 1$. At the other extreme are systems like Isis [5] or Targon [7] where $m = n$, and every processor has a backup processor to which it sends checkpoints. As m grows, the overhead of checkpointing and recovery will decrease because there is less contention for the parity processors.

To tolerate *any* combination of m processor failures, m parity processors must be combined with more sophisticated error-correction techniques [6, 8]. This means that every processor's checkpoint must be sent to multiple parity processors. In the absence of broadcast hardware, this kind of fault-tolerance will likely impose too great an overhead.

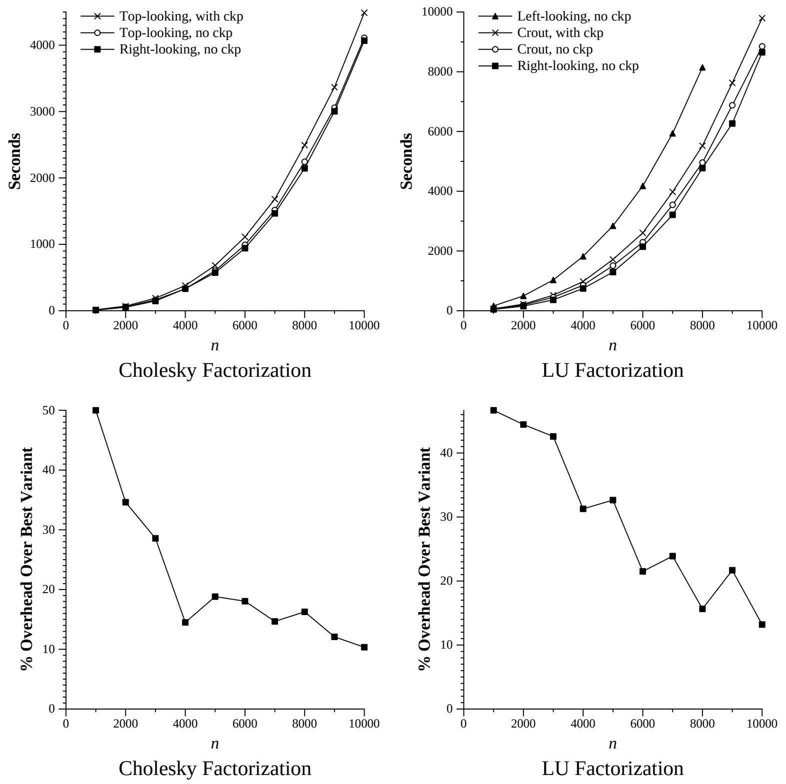


Figure 10: Performance of checkpointing as compared to all factorization variants

7.3 Choice of Factorization Algorithm

As a final remark, the choice of the algorithms for the factorization has an impact on both the performance of the factorization and the performance of checkpointing. Specifically, for Cholesky factorization, there are top-looking and right-looking algorithm variants and for LU factorization, there are left-looking, right-looking and Crout variants [16]. In general, the right-looking algorithms perform the best because they minimize communication overhead. However, when factoring column-block i, they modify all blocks in column-block j and row-block k such that $j \geq i$ and $k \geq i$. Thus, the average iteration modifies $\sum_{i=1}^{\frac{N}{b}}(bi)^2 \approx \frac{N^3}{3b}$ matrix elements that would have to be checkpointed. This would lead to prohibitively high overheads in terms of both time and memory.

To assess the impact of our algorithm selection, we implemented all algorithm variants of Cholesky and LU factorization without checkpointing. Figure 10 plots the running times of these variants for all problem sizes, and includes the results of checkpointing. The lower row of graphs plots the overhead of checkpointing as compared to the fastest of the factorization variants.

Figure 10 shows that the checkpointing overhead of both Cholesky and LU factorizations is low even compared to the right-looking variants. They too exhibit the trend of decreasing percentage of overhead as the problem size increases.

8 Related Work

There has been much research on algorithm-based fault-tolerance for matrix operations on parallel platforms where (unlike the above platform) the computing nodes are not responsible for storage of the input and output elements [23, 31, 38]. These methods concentrate mainly on fault-detection, and in some cases, correction. It is future research to see whether these techniques can be used to further improve diskless checkpointing. It should be noted that in the absence of a system that performs detection of data corruption, our scheme can be used for that purpose. In terms of coding theory, our mechanism is one that can provide detection in an error-prone environment, or correction in an erasure environment [34].

Checkpointing on parallel and distributed systems has been studied and implemented by many people [7, 14, 15, 17, 24, 26, 30, 36, 39, 42, 43, 44]. All of this work however focuses on either checkpointing to disk or process replication. The technique of using a collection of extra processors to provide fault-tolerance with no reliance on disk comes from Plank and Li [35] and is unique to this work.

There are efforts to provide programming platforms for heterogeneous computing that can adapt to changing load. These can be divided into two groups – those presenting new paradigms for parallel programming that facilitate fault-tolerance/migration [2, 3, 15, 20], and migration tools based on consistent checkpointing [9, 37, 41]. In the former group, the programmer must make his or her program conform to the programming model of the platform. None are garden variety message-passing environments like PVM or MPI. Those in the latter group achieve transparency, but cannot migrate a process without that process's participation. Thus, they cannot handle processor failures or revocation due to ownership without checkpointing to a central disk.

9 Conclusions

We have given a method for executing certain scientific computations on a changing or faulty Network of Workstations. This method enables a computation designed to execute on n processors to run on a NOW platform where individual processors may leave and enter the NOW due to failures or load. As long as the number of processors in the NOW is greater than n, and as long as processors leave the NOW singly, the computation can proceed efficiently.

We have implemented this method on four scientific calculations and shown performance results on a fast network of Sparc-5 workstations. The results show that our methods exhibit low overhead while checkpointing at a fine-grained interval (in most cases less than 5 minutes).

Our continuing progress with this work has been in three directions. First, we are adding the ability for processors to join the NOW in the middle of a calculation and participate in the fault-tolerant operation of the program. Currently, once a processor quits, the system merely completes with exactly n processors and no checkpointing. Second, we have added the capacity for multiple parity processors as outlined in Section 7.2. Preliminary results have shown that this improves both the reliability of the computation and the performance of checkpointing. Third, we are designing a technique to checkpoint the right-looking factorizations using a checksum approach and reverse computation to restore the bulk of processor state upon failure. The result is a mechanism that checkpoints at somewhat larger intervals, but with lower overhead than the algorithms described in this paper.

For the future, we would like to integrate our scheme with general load-balancing. In other words, if a few processors are added to or deleted from the NOW, then the system continues running using the mechanisms outlined in this paper. However, if the size of the processor pool changes by an order of magnitude, then it makes sense to reconfigure the system with a different value of n. Such an integration would represent a truly adaptive, high-performance methodology for scientific computations on NOWs.

Acknowledgements

The authors thank the following people for their help concerning this research: Richard Barrett, Micah Beck, Randy Bond, Jaeyoung Choi, Brian Davis, Chris Jepeway, Mark Jones, Kai Li, Bob Manchek, Nitin Vaidya and Clint Whaley.

James Plank is supported by National Science Foundation grant CCR-9409496 and the ORAU Junior Faculty Enhancement Award. Jack Dongarra is supported by the Defense Advanced Research Projects Agency under contract DAAL03-91-C-0047, administered by the Army Research Office by the Office of Scientific Computing, U.S. Department of Energy, under Contract DE-AC05-84OR21400 by the National Science Foundation Science and Technology Center Cooperative Agreement CCR-8809615.

References

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK User's Guide.* SIAM, Philadelphia, PA, 1992.

[2] J. N. C. Arabe, A. Beguelin, B. Lowekamp, E. Seligman, M. Starkey, and P. Stephan. Dome: Parallel programming in a distributed computing environment. In *10th International Parallel Processing Symposium*, April 1996.

[3] D. E. Bakken and R. D. Schilchting. Supporting fault-tolerant parallel programming in linda. *IEEE Transactions on Parallel and Distributed Systems*, 6(3):287–302, March 1995.

[4] R. Barrett *et al. Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods.* SIAM Press, Philadelphia, Pennsylvania, 1994.

[5] K.P. Birman and K. Marzullo. ISIS and the meta project. *Sun Technology*, Summer 1989.

[6] M. Blaum, J. Brady, J. Bruck, and J. Menon. EVENODD: An optimal scheme for tolerating double disk failures in RAID architectures. In *21st Annual International Symposium on Computer Architecture*, pages 245—254, Chicago, IL, April 1994.

[7] A. Borg, W. Blau, W. Graetsch, F. Herrman, and W. Oberle. Fault tolerance under UNIX. *ACM Transactions on Computer Systems*, 7(1):1–24, Feb 1989.

[8] W. A. Burkhard and J. Menon. Disk array storage system reliability. In *23rd International Symposium on Fault-Tolerant Computing*, pages 432–441, Toulouse, France, June 1993.

[9] J. Casas, D. Clark, R. Konuru, S. Otto, R. Prouty, and J. Walpole. MPVM: A migration transparent version of PVM. *Computing Systems*, 8(2):171–216, Spring 1995.

[10] J. Casas, D. L. Clark, P. S. Galbiati, R. Konuru, S. W. Otto, R. M. Prouty, and J. Walpole. MIST: PVM with transparent migration and checkpointing. In *3rd Annual PVM Users' Group Meeting*, Pittsburgh, PA, May 1995.

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

[12] J. Choi, J. Demmel, I. Dhillon, J. Dongarra, S. Ostrouchov, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley. ScaLAPACK: A portable linear algebra library for distributed memory. Technical Report CS-95-283, University of Tennessee, 1995.

[13] J. Choi, J. Dongarra, R. Pozo, and D. Walker. ScaLAPACK: A scalable linear algebra library for distributed memory concurrent computers. In *Proceedings of the Fourth Symposium on the Frontiers of Massively Parallel Computation*, pages 120–127. IEEE Publishers, 1992.

[14] F. Cristian and F. Jahanain. A timestamp-based checkpointing protocol for long-lived distributed computations. In *10th Symposium on Reliable Distributed Systems*, pages 12–20, October 1991.

[15] D. Cummings and L. Alkalaj. Checkpoint/rollback in a distributed system using coarse-grained dataflow. In *24th International Symposium on Fault-Tolerant Computing*, pages 424–433, Austin, TX, June 1994.

[16] J. J. Dongarra, I. S. Duff, D. C. Sorensen, and H. A. van der Vorst. *Solving Linear Systems on Vector and Shared Memory Computers*. SIAM, Philadelphia, Pennsylvania, 1991.

[17] E. N. Elnozahy, D. B. Johnson, and W. Zwaenepoel. The performance of consistent checkpointing. In *11th Symposium on Reliable Distributed Systems*, pages 39–47, October 1992.

[18] E. N. Elnozahy and W. Zwaenepoel. On the use and implementation of message logging. In *24th International Symposium on Fault-Tolerant Computing*, pages 298–307, Austin, TX, June 1994.

[19] A. Geist, A. Beguelin, J. Dongarra, R. Manchek, W. Jaing, and V. Sunderam. *PVM — A Users' Guide and Tutorial for Networked Parallel Computing.* MIT Press, Boston, 1994.

[20] D. Gelernter and D. Kaminsky. Supercomputing out of recycled garbage: Preliminary experience with piranha. In *International Conference on Supercomputing*, pages 417–427, Washington, D.C., June 1992. ACM.

[21] G. A. Gibson, L. Hellerstein, R. M. Karp, R. H. Katz, and D. A. Patterson. Failure correction techniques for large disk arrays. In *Third International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 123–132, Boston, MA, April 1989.

[22] G. H. Golub and C. V. Van Loan. *Matrix Computations, 2nd ed.* The Johns Hopkins University Press, Baltimore, Mayland, 1989.

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

[24] D. B. Johnson and W. Zwaenepoel. Recovery in distributed systems using optimistic message logging and checkpointing. *Journal of Algorithms*, 11(3):462–491, September 1990.

[25] R. Koo and S. Toueg. Checkpointing and rollback-recovery for distributed systems. *IEEE Transactions on Software Engineering*, SE-13(1):23–31, January 1987.

[26] T. H. Lai and T. H. Yang. On distributed snapshots. *Information Processing Letters*, 25:153–158, May 1987.

[27] L. A. Laranjeira, M. Malek, and R. M. Jenevein. Space/time overhead analysis and experiments with techniques for fault tolerance. *Dependable Computing and Fault-Tolerant Systems*, 8(3):303–318, 1993.

[28] J. León, A. L. Fisher, and P. Steenkiste. Fail-safe PVM: A portable package for distributed programming with transparent recovery. Technical Report CMU-CS-93-124, Carnegie Mellon University, February 1993.

[29] K. Li, J. F. Naughton, and J. S. Plank. An efficient checkpointing method for multicomputers with wormhole routing. *International Journal of Parallel Processing*, 20(3):159–180, June 1992.

[30] K. Li, J. F. Naughton, and J. S. Plank. Low-latency, concurrent checkpointing for parallel programs. *IEEE Transactions on Parallel and Distributed Systems*, 5(8):874–879, August 1994.

[31] F. T. Luk and H. Park. An analysis of algorithm-based fault tolerance techniques. *Journal of Parallel and Distributed Computing*, 5:172–184, 1988.

[32] Message Passing Interface Forum. MPI: A message-passing interface standard. *International Journal of Supercomputer Applications*, 8(3/4), 1994.

[33] M. W. Mutka and M. Livny. The available capacity of a privately owned workstation environment. *Performance Evaluation*, August 1991.

[34] W. W. Peterson and E. J. Weldon, Jr. *Error-Correcting Codes, Second Edition*. The MIT Press, Cambridge, Massachusetts, 1972.

[35] J. S. Plank and K. Li. Faster checkpointing with $N+1$ parity. In *24th International Symposium on Fault-Tolerant Computing*, pages 288–297, Austin, TX, June 1994.

[36] J. S. Plank and K. Li. Ickp — a consistent checkpointer for multicomputers. *IEEE Parallel & Distributed Technology*, 2(2):62–67, Summer 1994.

[37] J. Pruyne and M. Livny. Parallel processing on dynamic resources with CARMI. In *First IPPS Workshop on Job Scheduling Strategies for Parallel Processing*, April 1995.

[38] A. Roy-Chowdhury and P. Banerjee. Algorithm-based fault location and recovery for matrix computations. In *24th International Symposium on Fault-Tolerant Computing*, pages 38–47, Austin, TX, June 1994.

[39] L. M. Silva, J. G. Silva, S. Chapple, and L. Clarke. Portable checkpointing and recovery. In *Proceedings of the HPDC-4, High-Performance Distributed Computing*, pages 188–195, Washington, DC, August 1995.

[40] M. Snir, S. W. Otto, S. Huss-Lederman, D. W. Walker, and J. J. Dongarra. *MPI: The Complete Reference*. MIT Press, Boston, 1996.

[41] G. Stellner. CoCheck: Checkpointing and process migration for MPI. In *10th International Parallel Processing Symposium*, April 1996.

[42] R. E. Strom and S. Yemini. Optimistic recovery in distributed systems. *ACM Transactions on Computer Systems*, 3(3):204–226, August 1985.

[43] G. Sure, R. Janssens, and W. K. Fuchs. Reduced overhead logging for rollback recovery in distributed shared memory. In *25th International Symposium on Fault-Tolerant Computing*, pages 279–288, June 1995.

[44] Y. M. Wang and W. K. Fuchs. Lazy checkpoint coordination for bounding rollback propagation. In *12th Symposium on Reliable Distributed Systems*, pages 78–85, October 1993.