CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 0000; **00**:1–13
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cpe

Extending the Scope of the Checkpoint-on-Failure Protocol for Forward Recovery in Standard MPI

Wesley Bland* Peng Du, Aurelien Bouteiller, Thomas Herault, George Bosilca, Jack J. Dongarra

Innovative Computing Laboratory, University of Tennessee
1122 Volunteer Blvd., Knoxville, TN 37996-3450, USA

SUMMARY

Most predictions of Exascale machines picture billion way parallelism, encompassing not only millions of cores, but also tens of thousands of nodes. Even considering extremely optimistic advances in hardware reliability, probabilistic amplification entails that failures will be unavoidable. Consequently, software fault tolerance is paramount to maintain future scientific productivity. Two major problems hinder ubiquitous adoption of fault tolerance techniques: 1) traditional checkpoint based approaches incur a steep overhead on failure free operations and 2) the dominant programming paradigm for parallel applications (the MPI Standard) offers extremely limited support of software-level fault tolerance approaches. In this paper, we present an approach that relies exclusively on the features of a high quality implementation, as defined by the current MPI Standard, to enable algorithmic based recovery, without incurring the overhead of customary periodic checkpointing. The validity and performance of this approach are evaluated on large scale systems, using the QR factorization as an example. Copyright © 0000 John Wiley & Sons, Ltd.

Received . . .

KEY WORDS: Fault Tolerance, Message Passing Interface, ABFT, Checkpoint-on-Failure

1. INTRODUCTION

The insatiable processing power needs of domain science has pushed High Performance Computing (HPC) systems to feature a significant performance increase over the years, even outpacing "Moore's law" expectations. Leading HPC systems, whose architectural history is listed in the Top500[†] ranking, illustrate the massive parallelism that has been embraced in the recent years; current number 1 – Titan – has over half a million cores (including accelerators), number 2 – Sequoia – has over 1.5 million cores, and even with the advent of accelerators, it requires no less than 98,000 cores for the DiRAC system (#23) to breach the Petaflop barrier. Indeed, the International Exascale Software Project, a group created to evaluate the challenges on the path toward Exascale, has published a public report outlining that a massive increase in scale will be necessary when considering probable advances in chip technology, memory and interconnect speeds, as well as limitations in power consumption and thermal envelope [1]. According to these projections, as early as 2014, billion way parallel machines, encompassing millions of cores, and tens of thousands of nodes, will be necessary to achieve the desired level of performance. Even considering extremely optimistic advances in hardware reliability, probabilistic amplification entails that failures will be

*Correspondence to: Innovative Computing Laboratory, University of Tennessee
1122 Volunteer Blvd., Knoxville, TN 37996-3450, USA. E-mail: wbland@icl.utk.edu
[†]www.top500.org

unavoidable, becoming common events. Hence, fault tolerance is paramount to maintain scientific productivity.

Already, for Petaflop scale systems the issue has become pivotal. On one hand, the capacity type workload, composed of a large amount of medium to small scale jobs, which often represent the bulk of the activity on many HPC systems, is traditionally left unprotected from failures, resulting in diminished throughput when failures occur. On the other hand, selected capability applications, whose significance is motivating the construction of supercomputing systems, are protected against failures by ad-hoc, application-specific approaches, at the cost of straining engineering efforts, translating into high software development expenditures. Traditional approaches based on periodic checkpointing and rollback recovery, incurs a steep overhead, as much as 25% [2], on failure-free operations. Forward recovery techniques, most notably Algorithm-Based Fault Tolerant techniques (ABFT), use mathematical properties to reconstruct failure-damaged data and do exhibit significantly lower overheads [3]. However, and this is a major issue preventing their wide adoption, the resiliency support ABFT demands from the MPI library largely exceeds the specifications of the MPI Standard [4] and has proven to be an unrealistic requirement, considering that only a handful of MPI implementations provide it. Several propositions have emerged during the efforts of the MPI forum toward the MPI-3 standard[‡]. However, these propositions are still in their infancy and it is expected that several years will pass before they are blessed by the forum in a future revision and become generally deployed and available.

The current MPI-3 standard leaves open an optional behavior regarding failures to qualify as a "high quality implementation." According to this specification, when using the `MPI_ERRORS_RETURN` error handler, the MPI library should return control to the user when it detects a failure. In this paper, we propose the idea of Checkpoint-on-Failure (CoF) as a minimal impact feature to enable MPI libraries to support forward recovery strategies. Despite the default application-wide abort action that all notable MPI implementations undergo in case of a failure, we demonstrate that an implementation that enables CoF is simple and yet effectively supports ABFT recovery strategies that completely avoid costly periodic checkpointing.

This paper is an extended version of the distinguished work published in [5]. It completes the analysis by considering the broader case of general applications where only part of the computations are handled by MPI routines. In Section 5, we explain how such applications, for which periodic checkpoint restart is generally not practical, can still integrate efficiently the subset of their MPI operations with the CoF approach. Additionally, this type of deployment also reduces the checkpoint overhead to an insignificant proportion of the runtime: the non-MPI part of the application can remain dormant during the redeployment of MPI, so that the dataset remains resident in memory without reloading from checkpoints. We then discuss the evaluation of this application scheme with an additional evaluation in the experimental section.

The paper is organized as follows: the next section presents typical fault tolerant approaches and related works to discuss their requirements and limitations. Then in Section 3 we present the CoF approach, and the minimal support required from the MPI implementation. Section 4 presents a practical use case: the ABFT QR algorithm and how it has been modified to fit the proposed paradigm. Section 5 introduces a technique for the integration of CoF-enabled operations in broader applications, and Section 6 presents an experimental evaluation of the implementation, followed by our conclusions.

2. BACKGROUND & RELATED WORK

Message passing is the dominant form of communication used in parallel applications, and MPI is the most popular library used to implement it. In this context, the primary form of fault tolerance today is rollback recovery with periodical checkpoints to disk. While this method is effective in allowing applications to recover from failures using a previously saved state, it causes serious

[‡]`http://meetings.mpi-forum.org/mpi3.0_ft.php`

DOI: 10.1002/cpe

scalability concerns [6]. Moreover, periodic checkpointing requires precise heuristics for fault frequency to minimize the number of superfluous, expensive protective actions [7, 8, 9, 10, 11]. In contrast, the work presented here focuses on enabling *forward recovery*. Checkpoint actions are taken only *after* a failure is detected; hence the checkpoint interval is optimal by definition, as there will be one checkpoint interval per effective fault.

Forward recovery leverages algorithms' properties to complete operations despite failures. In naturally fault tolerant applications, the algorithm can compute the solution while totally ignoring the contributions of failed processes. In ABFT applications, a recovery phase is necessary, but failure damaged data can be reconstructed only by applying mathematical operations on the remaining dataset [12]. A recoverable dataset is usually created by initially computing redundant data, dispatched so as to avoid unrecoverable loss of information from failures. At each iteration, the algorithm applies the necessary mathematical transformations to update the redundant data (at the expense of more communication and computation). Despite great scalability and low overhead [3, 13], the adoption of such algorithms has been hindered by the requirement that the support environment must continue to consistently deliver communications, even after being crippled by failures.

The current MPI Standard (MPI-3.0, [4]) does not provide significant help to deal with the required type of behavior. Section 2.8 states in the first paragraph: "*MPI does not provide mechanisms for dealing with failures in the communication system. [. . .] Whenever possible, such failures will be reflected as errors in the relevant communication call. Similarly, MPI itself provides no mechanisms for handling processor failures.*" Failures, be they due to a broken link or a dead process, are considered resource errors. Later, in the same section: "*This document does not specify the state of a computation after an erroneous MPI call has occurred. The desired behavior is that a relevant error code be returned, and the effect of the error be localized to the greatest possible extent.*" So, for the current standard, process or communication failures are to be handled as errors, and the behavior of the MPI application after an error has been returned is left unspecified by the standard. However, the standard does not prevent implementations from going beyond its requirements, and on the contrary, encourages high-quality implementations to *return* errors once a failure is detected. Unfortunately, most of the implementations of the MPI Standard have taken the path of considering process failures as unrecoverable errors, and the processes of the application are most often killed by the runtime system when a failure hits any of them, leaving no opportunity for the user to mitigate the impact of failures.

In the past, some efforts have been undertaken to enable ABFT support in MPI. FT-MPI [14] was an MPI-1 implementation which proposed changes to the MPI semantics to enable repairing communicators, thus re-enabling communications for applications damaged by failures. This approach has proven successful and applications have been implemented using FT-MPI. However, these modifications were not adopted by the MPI standardization body, and the resulting lack of portability undermined user adoption for this fault tolerant solution.

During the process that recently resulted in the MPI-3 Standard, a specific working group was assembled to investigate the issues of Fault Tolerance support in MPI. Some of the early results are outlined in the following publication [15]. Late in the process, promising results had been demonstrated toward effective support of process failures and continued MPI operations with acceptable overhead [16]. However, these propositions were in too early a state to meet the calendar requirements of MPI-3 and their adoption (and according availability in production systems) is, at best, postponed to the restart of the MPI Forum toward the next version of the MPI standard.

In [17], the authors discuss alternative or slightly modified interpretations of the MPI Standard that enable some forms of fault tolerance. One core idea is that process failures happening in another MPI world, connected only through an inter-communicator, should not prevent the continuation of normal operations. The complexity of this approach, for both the implementation and users, has prevented these ideas from having a practical impact.

In the CoF approach, the only requirement from the MPI implementation is that it does not forcibly kill the living processes without returning control. No stronger support from the MPI stack is required, and the state of the library is left undefined. This simplicity has enabled us to actually

DOI: 10.1002/cpe

implement our proposition, and then experimentally support and evaluate a real ABFT application. Similarly, little effort would be required to extend MPICH-2 to support CoF (see Section 7 of the Readme[§]).

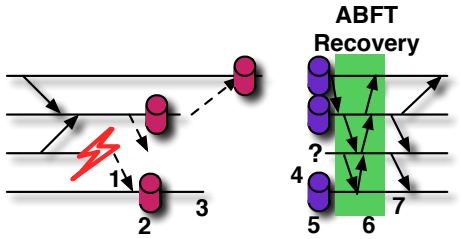
3. ENABLING ALGORITHM-BASED FAULT TOLERANCE IN MPI

3.1. The Checkpoint-on-Failure Protocol

In this paper, we advocate that an extremely efficient form of fault tolerance can be implemented, strictly based on the MPI Standard, for applications capable of taking advantage of forward recovery. ABFT methods are an example of forward recovery algorithms, capable of restoring missing data from redundant information located on other processes. This forward recovery step requires communication between processes, and we acknowledge that, in light of the current standard, requiring the MPI implementation to maintain service after failures is too demanding. However, a high-quality MPI library should at least allow the application to regain control following a process failure. We note that this control gives the application the opportunity to save its state and exit gracefully, rather than the usual behavior of being aborted by the MPI implementation.

Algorithm 1 The Checkpoint-on-Failure Protocol

1. MPI returns an error on surviving processes
2. Surviving processes checkpoint
3. Surviving processes exit
4. A new MPI application is started
5. Processes load from checkpoint (if any)
6. Processes enter ABFT dataset recovery
7. Application resumes



Based on these observations, we propose a new approach for supporting ABFT applications, called Checkpoint-on-Failure (CoF). Algorithm 1 presents the steps involved in the CoF method. In the associated explanatory figure, horizontal lines represent the execution of processes in two successive MPI applications. When a failure eliminates a process, other processes are notified and regain control from ongoing MPI calls (1). Surviving processes assume the MPI library is dysfunctional and do not call further MPI operations (in particular, they do not yet undergo ABFT recovery). Instead, they checkpoint their current state independently and abort (2, 3). When all processes exited, the job is usually terminated, but the user (or a managing script, batch scheduler, runtime support system, etc.) can launch a new MPI application (4), which reloads processes from checkpoint (5). In the new application, the MPI library is functional and communications possible; the ABFT recovery procedure is called to restore the data of the process(es) that could not be restarted from checkpoint (6). When the global state has been repaired by the ABFT procedure, the application is ready to resume normal execution.

Compared to periodic checkpointing, in CoF, a process pays the cost of creating a checkpoint only when a failure, or multiple simultaneous failures have happened, hence an optimal number of checkpoints during the run (and no checkpoint overhead on failure-free executions). Moreover, in periodic checkpointing, a process is protected only when its checkpoint is stored on safe, remote storage, while in CoF, local checkpoints are sufficient: the forward recovery algorithm reconstructs datasets of processes which cannot restart from checkpoint. Of course, CoF also exhibits the same overhead as the standard ABFT approach: the application might need to do extra computation, even in the absence of failures, to maintain internal redundancy (whose degree varies with the maximum number of simultaneous failures) used to recover data damaged by failures. However,

[§]`http://www.mpich.org/documentation/guides/files/mpich2-1.5-README.txt`

DOI: 10.1002/cpe

ABFT techniques often demonstrate excellent scalability; for example, the overhead on failure-free execution of the ABFT QR operation (used as an example in Section 4) is inversely proportional to the number of processes [13].

3.2. MPI Requirements for Checkpoint-on-Failure

Returning Control over Failures: In most MPI implementations, MPI_ERRORS_ABORT is the default (and often, only functional) error handler. However, the MPI Standard also defines the MPI_ERRORS_RETURN handler. To support CoF, the MPI library should never deadlock because of failures, but invoke the error handler, at least on processes doing direct communications with the failed process. The handler takes care of cleaning up at the library level and returns control to the application.

Termination After Checkpoint: A process that detects a failure ceases to use MPI. It only checkpoints on some storage and exits without calling MPI_Finalize. Exiting without calling MPI_Finalize is an error from the MPI perspective, hence the failure cascades and MPI eventually returns with a failure notification on every process, which triggers their own checkpoint procedure and termination.

3.3. Open MPI Implementation

Open MPI is an MPI 2.2 implementation architected such that it contains two main levels, the runtime (ORTE) and the MPI library (OMPI). As with most MPI library implementations, the default behavior of Open MPI is to abort after a process failure. This policy was implemented in the runtime system, preventing any kind of decision from the MPI layer or the user-level. The major change requested by the CoF protocol was to make the runtime system resilient, and leave the decision in case of failure to the MPI library policy, and ultimately to the user application.

Failure Resilient Runtime: The ORTE runtime layer provides an out-of-band communication mechanism (OOB) that relays messages based on a routing policy. Node failures not only impact the MPI communications, but also disrupt routing at the OOB level. The default routing policy in the Open MPI runtime has been amended to allow for self-healing behaviors; this effort is not entirely necessary, but it avoids the significant downtime imposed by a complete redeployment of the parallel job with resubmission in queues. The underlying OOB topology is automatically updated to route around failed processes. In some routing topologies, such as a star, this is a trivial operation and only requires excluding the failed process from the routing tables. For more elaborate topologies, such as a binomial tree, the healing operation involves computing the closest neighbors in the direction of the failed process and reconnecting the topology through them. The repaired topology is not rebalanced, resulting in degraded performance but complete functionality after failures. Although in-flight messages that were currently "hopping" through the failed processes are lost, other in-flight messages are safely routed on the repaired topology. Thanks to self-healing topologies, the runtime remains responsive, even when MPI processes leave.

Failure Notification: The runtime has been augmented with a failure detection service. To track the status of the failures, an incarnation number has been included in the process names. Following a failure, the name of the failed process (including the incarnation number) is broadcasted over the OOB topology. By including this incarnation number, we can identify transient process failures, prevent duplicate detections, and track message status. ORTE processes monitor the health of their neighbors in the OOB routing topology. Detection of other processes rely on a failure resilient broadcast that overlays on the OOB topology. This algorithm has a low probability of creating a bi-partition of the routing topology, hence ensuring a high accuracy of the failure detector. However, the underlying OOB routing algorithm has a significant influence on failure detection and propagation time, as the experiments will show. On each node, the ORTE runtime layer forwards

DOI: 10.1002/cpe

failure notifications to the MPI layer, which has been modified to invoke the appropriate MPI error handler.

4. EXAMPLE: THE QR FACTORIZATION

In this section, we propose to illustrate the applicability of CoF by considering a representative routine of a widely used class of algorithms: dense linear factorizations. The QR factorization is a cornerstone building block in many applications, including solving $Ax = b$ when matrices are ill-conditioned, computing eigenvalues, least square problems, or solving sparse systems through the GMRES iterative method. For an $M \times N$ matrix A, the QR factorization produces Q and R, such that $A = QR$ and Q is an $M \times M$ orthogonal matrix and R is an $M \times N$ upper triangular matrix. The most commonly used implementation of the QR algorithm on a distributed memory machine comes from the ScaLAPACK linear algebra library [18], based on the block QR algorithm. It uses a 2D block-cyclic distribution for load balance, and is rich in level 3 BLAS operations, thereby achieving high performance.

4.1. ABFT QR Factorization

In the context of FT-MPI, the ScaLAPACK QR algorithm has been rendered fault tolerant through an ABFT method in previous works [13]. This ABFT algorithm protects both the left (Q) and right (R) factors from fail-stop failures at any time during the execution. At the time of failure, every surviving process is notified by FT-MPI. FT-MPI then spawns a replacement process that takes the same grid coordinates in the $P \times Q$ block-cyclic distribution. Missing checksums are recovered from duplicates, a reduction collective communication recovers missing data blocks in the right factor from checksums. The left factor is protected by the Q-parallel panel checksum, it is either directly recovered from checksum, or by recomputing the panels in the current Q-wide section (see [13]). Although this algorithm is fault tolerant, it requires continued service from the MPI library after failures – which is a stringent requirement that can be waived with CoF.

4.2. Checkpoint-on-Failure QR

Checkpoint Procedure: Compared to a regular ABFT algorithm, CoF requires a different checkpoint procedure. System-level checkpointing is not applicable, as it would result in restoring the state of the broken MPI library upon restart. Instead, a custom MPI error handler invokes an algorithm specific checkpoint procedure, which simply dumps the matrices and the value of important loop indices into a file.

State Restoration: A ScaLAPACK program has a deep call stack, layering functions from multiple software packages, such as PBLAS, BLACS, LAPACK and BLAS. In the FT-MPI version of the algorithm, regardless of when the failure is detected, the current iteration of the algorithm must be completed before entering the recovery procedure. This ensures an identical call stack on every process and a complete update of the checksums. In the case of the CoF protocol, failures interrupt the algorithm immediately, the current iteration cannot be completed due to lack of communication capabilities. This results in potentially diverging call stacks and incomplete updates of checksums. However, because failure notification happens only in MPI, lower level, local procedures (BLAS, LAPACK) are never interrupted.

To resolve the call stack issue, every process restarted from checkpoint undergoes a “dry run” phase. This operation mimics the loop nests of the QR algorithm down to the PBLAS level, without actually applying modifications to or exchanging data. When the same loop indices as before the failure are reached, the matrix content is loaded from the checkpoint; the state is then similar to that of the FT-MPI based ABFT QR after a failure. The regular recovery procedure can be applied: the current iteration of the factorization is completed to update all checksums and the dataset is rebuilt using the ABFT reduction.

DOI: 10.1002/cpe

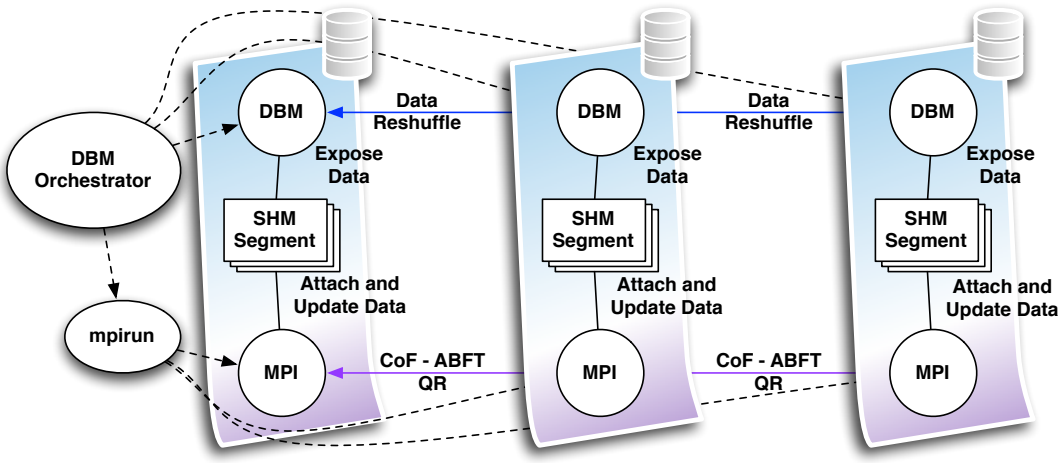


Figure 1. SciDB / CoF ABFT ScaLAPACK Integration

5. APPLICATION OF THE COF PROTOCOL TO A BROADER APPLICATION HORIZON

The CoF protocol circumvents one of the major limitations of current MPI implementations: the lack of confidence that the MPI library is capable of successfully completing communications once a failure happened. As illustrated above, algorithms based on the class of naturally fault tolerant algorithms are capable of taking advantage of this technique and provide efficient fault tolerance support. In this section, we explain how the CoF protocol can be efficiently integrated with other kinds of algorithms and applications, increasing the scope of such methods. We will illustrate the approach with a fault-tolerant database management system, SciDB.

Fault tolerant database management exposes a set of requirements that is best addressed today using replication and transactional operations. SciDB [19] combines database operations and many scientific specific operations (including linear algebra routines) to create a highly expressive request query language suitable for scientists to solve their data analysis problems. The SciDB system is not implemented on top of MPI for its communication, mainly because of the lack of fault tolerance capabilities from the MPI Standard. It makes use, however, of the MPI implementation of the distributed linear algebra operations in ScaLAPACK, to provide, among other things, various factorization routines. Because most MPI implementations are not usable after a process failure, and high availability is a necessity in database management systems, the SciDB implementation cannot integrate the MPI library in its main process. As a result, its linear algebra operations are implemented as separate processes: a query coordinator will order the distributed database managers to locate the data on which the factorization operation must be applied and to expose this data in the expected ScaLAPACK layout using one shared memory segment per node; it will then launch a ScaLAPACK/MPI application that will attach to this memory segment and apply the operation on it. If a failure hits a node, the MPI application will abort, and the *mpirun* child process reports the error to the data query coordinator. The original data is recovered from the database management system (using database-specific fault tolerant techniques), and the linear algebra operation relaunches from scratch on the original data.

This approach can be improved using the CoF protocol and an ABFT implementation of the factorization operation. The idea is described in Figure 1. The same general approach to combine SciDB and ScaLAPACK is used; however, the DB managers will compute the initial checksum of the original data, and expose both the data and checksum to the ABFT-ScaLAPACK process. The ABFT operation is applied, and if no failure happened, the result of the factorization is accessible in the shared memory segments as it was before (the checksum data can then be discarded by the DB managers). If a failure occurs, the MPI process updates the shared memory segments with the

DOI: 10.1002/cpe

meta information of the checkpoint (values of the loop counters, etc...); the current statuses of the shared memory segments represent the rest of the checkpoints that were done in the normal CoF protocol. Then, the MPI processes quit and the *mpirun* child process reports the error to the database coordinator. Instead of fixing the data issue at the DB level, the coordinator relaunches a new ABFT-ScaLAPACK operation on the same set of nodes plus a spare node with an empty shared memory segment, and it lets the ABFT algorithm recover the data and continue the original operation. Once this is successfully completed, the *mpirun* child process reports success to the database coordinator whic will find the data in the shared memory segments of the living nodes and can then discard the checksum data.

This approach has two advantages compared to the original and the checkpoint-based CoF approaches:

- First, in the case of a failure, instead of restarting from scratch, the factorization incurs only the small recovery overhead of ABFT, ensuring a faster time-to-solution for the linear algebra operation. In exchange, a small overhead, for creating and maintaining the checksum data during the operation, is imposed on the failure-free case.
- Second, this approach removes the cost of writing the checkpoint to a file: the shared memory segment that survives the exit of the MPI processes where the node was not subject to a failure and the checksum information maintained by the ABFT algorithm are sufficient to recover the missing data. The segment of memory on which the operation is computed is made remanent, creating the bulk of the checkpoint data and reducing to an insignificant value the cost of checkpointing when a failure occurs. This will be demonstrated in the experimental section, below.

6. PERFORMANCE DISCUSSION

In this section, we use our Open MPI and ABFT QR implementations to evaluate the performance of the CoF protocol. We use two test platforms. The first machine, “Dancer”, is a 16-node cluster. All nodes are equipped with two 2.27GHz quad-core Intel E5520 CPUs with a 20GB/s Infiniband interconnect. Solid State Drive (SSD) disks are used as the checkpoint storage media. The second system is the “Kraken” supercomputer. Kraken is a Cray XT5 machine with 9,408 compute nodes. Each node has two Istanbul 2.6 GHz six-core AMD Opteron processors, 16 GB of memory, and is connected to other nodes through the SeaStar2+ interconnect. The scalable cluster file system “Lustre” is used to store checkpoints.

6.1. MPI Library Overhead

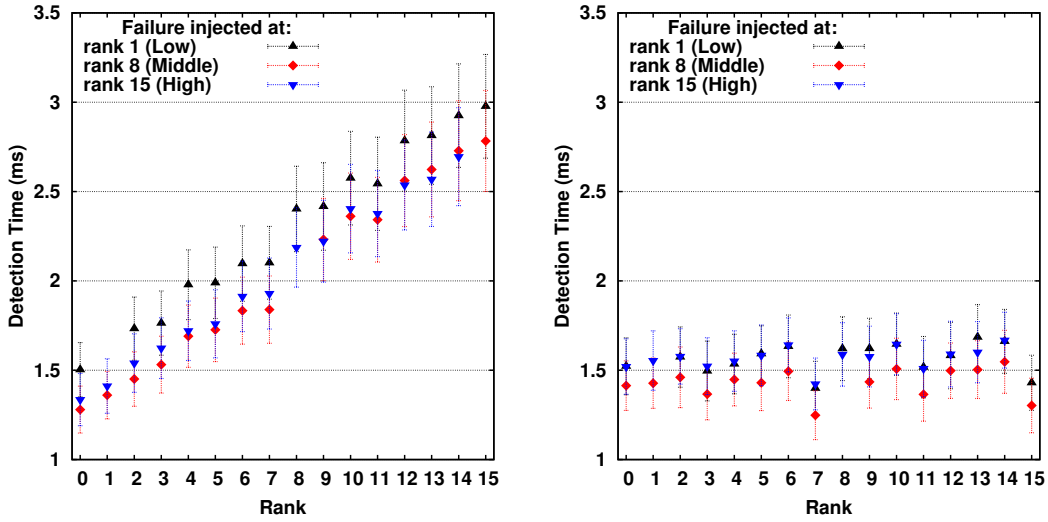
One of the concerns when evaluating the performance of fault tolerance techniques is the amount of overhead introduced by the fault tolerance management additions. Our implementation of fault detection and notification is mostly implemented in the non-critical ORTE runtime. Typical HPC systems feature a separated service network (usually Ethernet based) and a performance interconnect, hence health monitoring traffic, which happens on the OOB service network, is physically separated from the MPI communications, leaving no opportunity for network jitter. Changes to MPI functions are minimal: the same condition that used to trigger unconditional abort has been repurposed to trigger error handlers. As expected, no impact on MPI bandwidth or latency was measured. The memory usage of the MPI library is slightly increased, as the incarnation number doubles the size of process names; however, this is negligible in typical deployments.

6.2. Failure Detection

According to the requirement specified in Section 3.2, only in-band failure detection is required to enable CoF. Processes detecting a failure checkpoint then exit, cascading the failure to processes communicating with them. However, no recovery action (in particular checkpointing) can take place before a failure has been notified. Thanks to asynchronous failure propagation in the runtime,

Prepared using cpeauth.cls
DOI: 10.1002/cpe

responsiveness can be greatly improved, with a high probability for the next MPI call to detect the failures, regardless of communication pattern or checkpoint duration.



(a) Linear OOB Routing

(b) Binomial OOB Routing

Figure 2. Failure detection time, sorted by process rank, depending on the OOB overlay network used for failure propagation.

We designed a micro-benchmark to measure failure detection time as experienced by MPI processes. The benchmark code synchronizes with an `MPI_BARRIER`, stores the reference date, injects a failure at a specific rank, and enters a ring algorithm until the MPI error handler stores the detection date. The OOB routing topology used by the ORTE runtime introduces a non-uniform distance to the failed process, hence failure detection time experienced by a process may vary with the position of the failed process in the topology, and the OOB topology. Figure 2(a) and 2(b) present the case of the linear and binomial OOB topologies, respectively. The curves "Low, Middle, High" present the behavior for failures happening at different positions in the OOB topology. On the horizontal axis is the rank of the detecting process, on the vertical axis is the detection time it experienced. The experiment uses 16 nodes, with one process per node, MPI over Infiniband, OOB over Ethernet, an average of 20 runs, and the MPI barrier latency is four orders of magnitude lower than measured values.

In the linear topology (Figure 2(a)) every runtime process is connected to the *mpirun* process. For a higher rank, failure detection time increases linearly because it is notified by the *mpirun* process only after the notification has been sent to all lower ranks. This issue is bound to increase with scale. The binomial tree topology (Figure 2(b)) exhibits a similar best failure detection time. However, this more scalable topology has a low output degree and eliminates most contentions on outgoing messages, resulting in a more stable, lower average detection time, regardless of the failure position. Overall, failure detection time is on the order of milliseconds, a much smaller figure than typical checkpoint time.

6.3. Checkpoint-on-Failure QR Performance

Supercomputer Performance: Figure 3 presents the performance on the Kraken supercomputer. The process grid is 24×24 and the block size is 100. ABFT-QR (no failure) presents the performance of the CoF QR implementation, in a fault-free execution; it is noteworthy, that when there are no failures, the performance is exactly identical to the performance of the unmodified ABFT-QR implementation. The ABFT-QR (with CoF recovery, latter called CoF-QR for brevity) curves present the performance when a failure is injected after the first step of the PDLARFB kernel. The performance of the non-fault tolerant ScaLAPACK QR is also presented for reference.

DOI: 10.1002/cpe

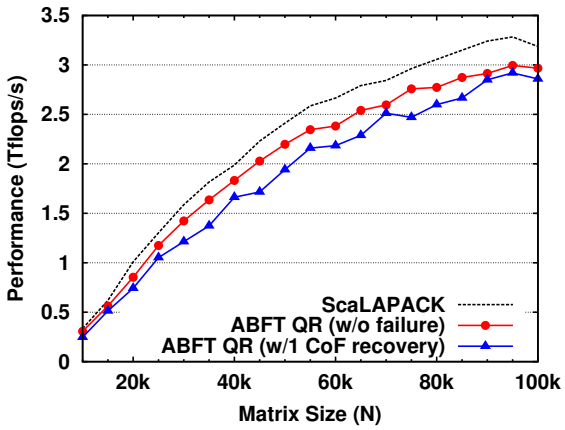
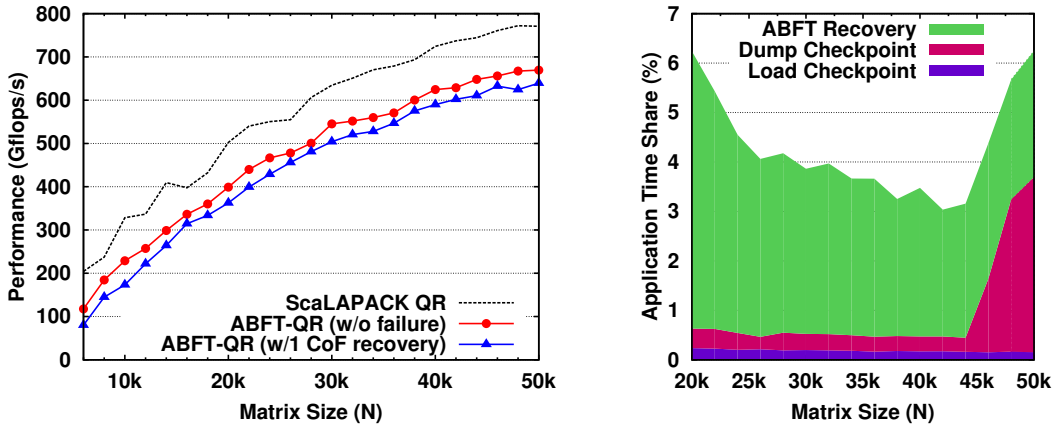


Figure 3. ABFT QR and one CoF recovery on Kraken (Lustre).

Without failures, the performance overhead compared to the regular ScaLAPACK is caused by the extra computation to maintain the checksums inherent to the ABFT algorithm [13]; this extra computation is unchanged when applying the CoF method to the ABFT-QR. Only on runs where failures occur does the CoF protocol undergoe the supplementary overhead of storing and reloading checkpoints. However, the performance of CoF-QR remains very close to the no-failure case. For instance, at matrix size N=100,000, CoF-QR still achieves 2.86 Tflop/s after recovering from a failure, which is 90% of the performance of the non-fault tolerant ScaLAPACK QR. This demonstrates that the CoF protocol enables efficient, practical recovery schemes on supercomputers.



(a) Performance

(b) Time breakdown of one CoF recovery

Figure 4. ABFT QR and one CoF recovery on Dancer (local SSD).

Impact of Local Checkpoint Storage: Figure 4(a) presents the performance of the CoF-QR implementation on the Dancer cluster with a 8×16 process grid. Although a smaller test platform, the Dancer cluster features local storage on nodes and a variety of performance analysis tools unavailable on Kraken. As expected (see [13]), the ABFT method has a higher relative cost on this smaller machine (with a smaller number of processors and a smaller problem size, the cost in supplementary operations to update checksums is relatively larger). Compared to the Kraken platform, the relative cost of CoF failure recovery is smaller on Dancer. The CoF protocol incurs disk accesses to store and load checkpoints when a failure hits, hence the recovery overhead

DOI: 10.1002/cpe

depends on I/O performance. By breaking down the relative cost of each recovery step in CoF, Figure 4(b) shows that checkpoint saving and loading only takes a small percentage of the total run-time, thanks to the availability of solid state disks on every node. Since checkpoint reloading immediately follows checkpointing, the OS cache satisfies most disk accesses, resulting in high I/O performance. For matrices larger than N=44,000, the memory usage on each node is high and decrease the available space for disk cache, explaining the decline in I/O performance and the higher cost of checkpoint management. Overall, the presence of fast local storage can be leveraged by the CoF protocol to speedup recovery (unlike periodic checkpointing, which depends on remote storage by construction). Nonetheless, as demonstrated by the efficiency on Kraken, while this is a valuable optimization, it is not a mandatory requirement for satisfactory performance.

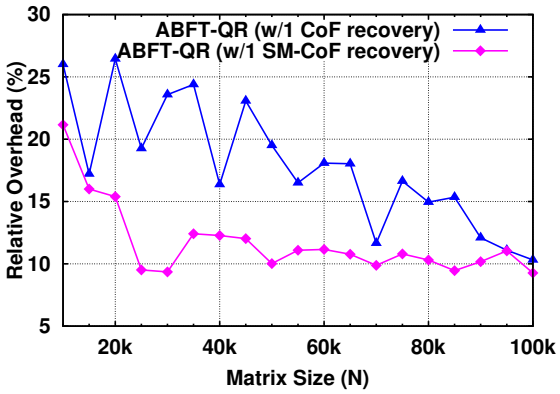


Figure 5. ABFT QR and one recovery on Kraken: comparing CoF and SM-CoF overheads.

Checkpoint-on-Failure, without the Checkpoints: An interesting optimization to CoF is to avoid the checkpointing cost by using the SM-CoF approach described in Section 5. In this paragraph, we present the performance of the QR factorization, when applied by a fragile helper MPI application, onto a dataset exported through a shared memory segment from a resilient, non-MPI application. Figure 5 compares the overhead incurred by introducing a failure with checkpoint-based CoF recovery versus a shared-memory-CoF recovery where a master application maintains the dataset resident in memory.

The cost of the ABFT recovery is unchanged by the use of SM-CoF; the obvious consequence is that, for very small matrix sizes, when the relative cost of ABFT checksum inversion represents a large portion of the overall compute time, the difference between the shared-memory optimization and the checkpoint based CoF is small. A similar result is observed for very large matrices: for a matrix of size N, checkpointing time is $O(N^2)$ while compute time is $O(N^3)$, thus the cost of storing and reloading checkpoints is dwarfed by the total execution time of the application and achieve similar asymptotic performance. For intermediate matrix sizes, however, the cost of checkpointing represents a significant share of the overhead experienced by the application during the recovery procedure. In that case, which is the most relevant in production deployments, the SM-CoF optimization successfully suppresses the checkpoint overhead and performs similarly to ABFT-QR on a fully fault tolerant MPI implementation, although at the expense of more complexity in the application code.

7. CONCLUDING REMARKS

In this paper, we presented an original scheme to enable forward recovery using only features of the current MPI Standard. Rollback recovery, which relies on periodic checkpointing, has a variety of issues. The ideal period between checkpoints, a critical parameter, is particularly hard to assess. Too

DOI: 10.1002/cpe

short a period wastes time and resources on unnecessary Input/Output. Overestimating the period results in dramatically increasing the lost computation when returning to the distant last successful checkpoint. Although Checkpoint-on-Failure involves checkpointing, it takes checkpoint images at optimal times by design: only after a failure has been detected. This small modification enables the deployment of ABFT techniques, without requiring a complex, unlikely to be available MPI implementation that itself survives failures. The MPI library needs only to provide the feature set of a high quality implementation of the MPI Standard: the MPI communications may be dysfunctional after a failure, but the library must return control to the application instead of aborting brutally.

We demonstrated, by providing such an implementation in Open MPI, that this feature set can be easily integrated without noticeable impact on communication performance. We then converted an existing ABFT QR algorithm to the CoF protocol. Beyond this example, the CoF protocol is applicable on a large range of applications that already feature an ABFT version (LLT, LU [20], CG [21], etc.). Many master-slave and iterative methods enjoy an extremely inexpensive forward recovery strategy where the damaged domains are simply discarded, and therefore can also benefit from the CoF protocol.

The performance on the Kraken supercomputer reaches 90% of the non-fault tolerant algorithm, even when including the cost of recovering from a failure (a figure similar to regular, non-compliant MPI ABFT). In addition, on a platform featuring node local storage, the CoF protocol can leverage low overhead checkpoints (unlike rollback recovery that requires remote storage). To the extreme, the cost of checkpointing can be completely avoided when the application uses a master process to actively retain the dataset in memory during the MPI restart.

The MPI standardization body, the MPI Forum, is currently considering the addition of new MPI constructs, functions and semantics to support fault-tolerant applications[¶]. While these additions may decrease the cost of recovery, they are likely to increase the failure-free overhead on fault tolerant application performance. It is therefore paramount to compare the cost of the CoF protocol with prospective candidates to standardization on a wide, realistic range of applications, especially those that feature a low computational intensity.

[¶]https://svn.mpi-forum.org/trac/mpi-forum-web/wiki/FaultToleranceWikiPage

Prepared using cpeauth.cls
DOI: 10.1002/cpe

REFERENCES

1. Dongarra J, Beckman P, Moore T, Aerts P, Aloisio G, Andre JC, Barkai D, Berthou JY, Boku T, Braunschweig B, *et al.*. The international exascale software project roadmap. *Int. J. High Perform. Comput. Appl.* Feb 2011; **25**(1):3–60, doi:10.1177/1094342010391989. URL http://dx.doi.org/10.1177/1094342010391989.
2. Schroeder B, Gibson GA. Understanding Failures in Petascale Computers. *SciDAC, Journal of Physics: Conference Series* 2007; **78**.
3. Luk F, Park H. An analysis of algorithm-based fault tolerance techniques. *Journal of Parallel and Distributed Computing* 1988; **5**(2):172–184.
4. The MPI Forum. MPI: A Message-Passing Interface Standard, Version 3.0. *Technical Report* 2012.
5. Bland W, Du P, Bouteiller A, Herault T, Bosilca G, Dongarra J. A Checkpoint-on-Failure protocol for Algorithm-Based Recovery in standard MPI. *Proceedings of Euro-Par 2012: 18th International Conference on Parallel Processing*, *Lecture Notes in Computer Science*, vol. 7484, Kaklamanis C, Papatheodorou TS, Spirakis PG (eds.), Springer: Rhodes Island, Greece, 2012; 477–488.
6. Cappello F, Geist A, Gropp B, Kalé LV, Kramer B, Snir M. Toward exascale resilience. *International Journal on High Performance Computing and Applications* 2009; **23**(4):374–388.
7. Young JW. A first order approximation to the optimum checkpoint interval. *Communication of the ACM* September 1974; **17**:530–531.
8. Gelenbe E. On the optimum checkpoint interval. *Journal of the ACM* 1979; **26**:259–270.
9. Plank JS, Thomason MG. Processor allocation and checkpoint interval selection in cluster computing systems. *Journal of Parallel and Distributed Computing* 2001; **61**:1590.
10. Daly JT. A higher order estimate of the optimum checkpoint interval for restart dumps. *Future Generation Computer Systems* February 2006; **22**:303–312.
11. Cappello F, Casanova H, Robert Y. Preventive migration vs. preventive checkpointing for extreme scale supercomputers. *Parallel Processing Letters* 2011; :111–132.
12. Huang K, Abraham J. Algorithm-based fault tolerance for matrix operations. *IEEE Transactions on Computers* 1984; **100**(6):518–528.
13. Du P, Bouteiller A, Bosilca G, Herault T, Dongarra J. Algorithm-based Fault Tolerance for Dense Matrix Factorizations. *17th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, ACM, 2012.
14. Fagg G, Dongarra J. FT-MPI: Fault tolerant MPI, supporting dynamic applications in a dynamic world. *Proceedings of the 7th European PVM/MPI Users' Group Meeting on Recent Advances in Parallel Virtual Machine and Message Passing Interface (EuroPVM/MPI)* 2000; .
15. Hursey J, Graham RL, Bronevetsky G, Buntinas D, Pritchard H, Solt DG. Run-through stabilization: An MPI proposal for process fault tolerance. *EuroMPI 2011: Proceedings of the 18th EuroMPI Conference*, Santorini, Greece, 2011.
16. Bland W, Bouteiller A, Hérault T, Hursey J, Bosilca G, Dongarra JJ. An evaluation of user-level failure mitigation support in mpi. *EuroMPI*, *Lecture Notes in Computer Science*, vol. 7490, Träff JL, Benkner S, Dongarra JJ (eds.), Springer, 2012; 193–203.
17. Gropp W, Lusk E. Fault tolerance in message passing interface programs. *International Journal of High Performance Computing and Applications* August 2004; **18**:363–372, doi:10.1177/1094342004046045.
18. Dongarra J, Blackford L, Choi J, *et al.*. ScaLAPACK user's guide. *Society for Industrial and Applied Mathematics, Philadelphia, PA* 1997; .
19. The SciDB Development Team. Overview of SciDB: large scale array storage, processing and analysis. *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, SIGMOD'10, ACM: New York, NY, USA, 2010; 963–968, doi:10.1145/1807167.1807271. URL http://doi.acm.org/10.1145/1807167.1807271.
20. Davies T, Karlsson C, Liu H, Ding C, , Chen Z. High Performance Linpack Benchmark: A Fault Tolerant Implementation without Checkpointing. *Proceedings of the 25th ACM International Conference on Supercomputing (ICS 2011)*, ACM.
21. Chen Z, Fagg GE, Gabriel E, Langou J, Angskun T, Bosilca G, Dongarra J. Fault tolerant high performance computing by a coding approach. *Proceedings of the tenth ACM SIGPLAN symposium on Principles and practice of parallel programming*, PPoPP '05, ACM: New York, NY, USA, 2005; 213–223, doi:http://doi.acm.org/10.1145/1065944.1065973. URL http://doi.acm.org/10.1145/1065944.1065973.