

ReStore: In-Memory REplicated STORagE for Rapid Recovery in Fault-Tolerant Algorithms

Lukas Hübner^{*†‡}, Demian Hesse^{*†}, Peter Sanders[†], Alexandros Stamatakis^{†‡}

^{*} *primary authors*

[†] Karlsruhe Institute of Technology, Karlsruhe, Germany

Email: {hesse,huebner,sanders,stamatakis}@kit.edu

[‡] Heidelberg Institute of Theoretical Studies, Heidelberg, Germany

Abstract—Fault-tolerant distributed applications require mechanisms to recover data lost via a process failure. On modern cluster systems it is typically impractical to request replacement resources after such a failure. Therefore, applications have to continue working with the remaining resources. This requires redistributing the workload and that the non-failed processes reload data. We present an algorithmic framework and its C++ library implementation ReStore for MPI programs that enables recovery of data after process failures. By storing all required data in memory via an appropriate data distribution and replication, recovery is substantially faster than with standard checkpointing schemes that rely on a parallel file system. As the application developer can specify which data to load, we also support shrinking recovery instead of recovery using spare compute nodes. We evaluate ReStore in both controlled, isolated environments and real applications. Our experiments show loading times of lost input data in the range of milliseconds on up to 24 576 processors and a substantial speedup of the recovery time for the fault-tolerant version of a widely used bioinformatics application.

Index Terms—Fault-Tolerance, MPI, ULFM, HPC

I. INTRODUCTION

With the increasing number of processors in high performance computing clusters, the probability that some processors fail during a computation rises. Handling such failures constitutes a major challenge for future exascale systems [2]. For example ORNL’s Jaguar Titan Cray XK7 system had on average 2.33 failures/day between August 2008 and 2010 [3]. In upcoming systems, we expect a hardware failure to occur every 30 to 60 minutes [4]–[6]. To recover from such a failure, an important step is to restore the lost data that the failed processors were working on. To save the current state of a program’s data, applications write *checkpoints* which can be reloaded after a process failure. Checkpointing libraries usually write their checkpoints to a parallel file system (PFS) [7]–[10], implying slow recovery due to low disk access speeds and because many processors simultaneously access the same resources. Many checkpointing libraries also assume the nature of the failures to leave the machine in a state where the process can simply be started again, or they assume that enough spare resources are kept idle to start a new process for replacing the failed one [3], [7]–[15]. Under this assumption, a re-spawned process can simply read exactly the data of the failed process. In the case of the new process being located on the same

compute node as the failed one, the checkpoint can even be read from a local disk. To the best of our knowledge there exists no general purpose checkpointing solution that allows for in-memory recovery without requiring spare resources.

Contribution and Structure. We introduce ReStore, an in-memory checkpointing library that is optimized for recovery speed (in contrast to checkpoint creation speed). This is especially important for data which the program never or rarely changes but has to be redistributed after every failure. We do not assume that spare resources are available. Instead, ReStore enables recovery in an application that continues its execution only with the processes that are still alive. While this approach requires a more involved recovery mechanism and strategic data distribution, it saves resources because all available processors can participate in the application’s useful computations from the beginning. Keeping the checkpoints in-memory avoids the bottlenecks involved in a PFS and allows high scalability. The remainder of this paper is structured as follows: We introduce the concepts used in this paper in Section II and provide an overview of existing checkpointing libraries and other related work in Section III. We explain our general framework and data distribution in Section IV. In Section V and VI we present the implementation of our open source C++ library and the experimental results, respectively. We conclude in Section VII and outline future work.

II. PRELIMINARIES

In distributed memory parallel programs using the *Message Passing Interface* (MPI), p processes (or *processing elements* (PEs)) run on multiple machines (or *nodes*) and communicate via messages sent over the network. We consider two important factors for evaluating the running time of such parallel algorithms: The *bottleneck* number of messages sent and received, and the *bottleneck communication volume*. The bottleneck number of messages sent and received describes the maximum number of messages sent or received by a single PE. This influences performance, as there is a startup overhead (latency) for establishing a connection associated with each message. As a result, sending or receiving data from one PE to another in a single message is usually faster than splitting that data into many parts and sending each part to a different receiver. The bottleneck communication volume describes the

TABLE I

COMPARISON OF CHECKPOINTING LIBRARIES. See Section III for details. ¹The program needs to allocate spare nodes, which participate in the computation only in case of a failure. ²The program needs to allocate spare nodes and nodes used purely for checkpointing. ³The maintenance state unclear (Section III-A).

	ftRMA [13]	Fenix [3]	SCR [12]	Lu [14]	GPI_CP [15]	ReStore
Features						
in-memory checkpointing	✓	✓	✗	✓	✓	✓
substituting recovery	✓	✓	✓	✓	✓	✓
shrinking recovery	✗	✗	✗	✗	✗	✓
all nodes participate in computation	✗ ²	(✓) ¹	(✓) ¹	✗ ²	(✓) ¹	✓
programming model	MPI RDMA	MPI	MPI	MPI	PGAS/GPI	MPI
Reproducibility						
source-code available	✓	✓	✓	✗	✓	✓
still maintained (2022)	✗	? ³	✓	✗	✗	✓
other reproducibility issues	Cray-only	author-provided examples segfault			requires libverbs, GPI-2 and passwordless ssh-login on all compute nodes	

maximum amount of data sent or received by a single PE and represents a point on the critical path of the application.

As faults, we consider the case that one or multiple PEs suddenly stop working and do not contribute to the computation anymore (which we will refer to as *failed*). The upcoming MPI 4 standard will include fault-tolerance mechanisms and an implementation called “ULFM” is available for OpenMPI [16]. Following a fault, the application has to redistribute the work formerly performed by a failed PE using either the *shrink* or the *substitute* strategy [17]. Under the *substitute* strategy, a replacement PE takes over the work previously performed by the failed PE. This circumvents the need for re-balancing the workload and simplifies loading the required data. However, reserving idle processors for this purpose constitutes a waste of resources. In the *shrink* strategy, the program’s load balancer (re)distributes the work performed by the failed PE among the remaining (or *surviving*) PEs. This strategy does therefore not require spare PEs but requires reloading fractions of the data on many or even all PEs. While the number of failures an algorithm can tolerate using the *substitute* strategy is limited to the number of spare PEs; this limitation does not apply to the *shrink* strategy [17].

III. RELATED WORK

Scientific applications are increasingly implemented to tolerate faults. Examples include a numeric linear equation and partial equation solver [18], a plasma simulation [19], a molecular dynamics simulations [20], a Fast Fourier Transformation [21], and an algorithm for phylogenetic inference [22]. The three main techniques for implementing fault-tolerant algorithms are Algorithm-Based Fault-Tolerance [23], [24], restarting failed sub-jobs [25], and checkpointing/restart [22], [26]. Checkpointing/restart can be further subdivided into coordinated and uncoordinated checkpointing. In coordinated checkpointing, the program synchronizes before creating the checkpoint in a distributed-manner. This ensures that there are no messages in-flight and the program’s state is therefore well-defined. Gavaskar and Subbarao recommend coordinated

checkpointing for the high-bandwidth, low-latency interconnections of modern HPC systems [27]. Checkpointing libraries can save their checkpoint either to a (possibly network attached) disk or to the compute node’s main memory (“diskless”) [28]. Checkpointing libraries which save their checkpoints to disk include, for example, the algorithm presented by Agarwal et al. [7], FTI [8], CRAFT [9], SCR [12], and VeloC [10]. As the number of nodes per parallel program execution continues to grow, the congestion on the PFS increases – resulting in a bottleneck and reduced checkpointing performance [29], [30]. Examples for in-memory checkpointing libraries include ftRMA [13], Fenix [3], GPI_CP [15], and the algorithm described by Lu [14] (Table I). All of these employ the substitute strategy and therefore rely on the availability of replacement nodes, if we want to continue the computation in case of node failure. This implies that some nodes are allocated to the job but not available for computation. Some algorithms additionally designate some compute nodes as pure checkpointing nodes, which are neither participating in the computation nor available as spares. Ashraf et al. [17] describe an implementation of a fault-tolerance mechanism for a specific application which is able to checkpoint to memory and recover in a shrinking setting. This is, however, not a general-purpose checkpointing library but application-specific. Erasure codes are often used to reduce the file-size or memory footprint of checkpoints [8], [13], [14].

Other areas where replication approaches similar to the one presented in this paper are used are distributed fault-tolerant file systems like early versions of the Hadoop Distributed File System [31] or distributed processing frameworks like Apache Spark [32]. However, these target very different use cases and sometimes only support very basic replication like storing each PEs data on a single partner PE.

A. Reproducibility Study

In the following, we describe our attempts to replicate the results of competing tools. We provide a visual summary of these in Table I.

	PE 0				PE 1				PE 2				PE 3			
copy 1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
copy 2	8	9	10	11	12	13	14	15	0	1	2	3	4	5	6	7

Fig. 1. Example showing the data distribution of the copies stored with ReStore for $p = 4$ PEs, $n = 16$ data blocks, and $r = 2$ copies.

The ftRMA [13] tool has not been maintained since 2014 and relies on the Cray-only foMPI library which has also not been further maintained since 2014. The authors confirmed (pers. comm. 30. June 2022) that the current code exclusively works on Cray systems and is no longer being actively maintained. Although the authors suggested that ftRMA could – in principle – be ported to a non-Cray system, taking into account the unmaintained code base comprising 513 calls to foMPI functions, this would incur a prohibitive programming effort with uncertain outcomes. Further, as ULFM currently provides “little support for fault tolerance” with respect to RMA calls [33], deploying ftRMA would be bound to fail using a current fault-tolerant MPI implementation.

With respect to the Fenix tool, there has only been a single commit to its repository within the past six months. In addition, the author’s automated testing on GitHub failed for this commit. We thus denote Fenix’s maintenance status as being “unclear” (✓) in Table I. The author-provided Fenix examples [3] fail with a segmentation fault. As Fenix does currently not support restoring the data that was checkpointed on a different rank, setting up an experimental comparison is challenging. The authors did not respond to our e-mail requesting assistance.

SCR [12] has > 50 commits on 20 distinct days during the past six months. Hence, we consider that it is still being maintained. SCR supports *caching* checkpoints on a RAM-disk. These checkpoints, however, have to be transferred to the parallel file system such as to become available upon rank failure. We therefore do not consider SCR to be an in-memory checkpointing library in the context of node failures.

The source code of Lu [14] is not available, and the author can not be contacted, as they did not provide a contact e-mail address on their publications.

The git repository of GPI_CP [15] has only a single commit from six years ago; we thus also consider that it is no longer being maintained. As we do not have access to an HPC system where GPI-2 (a library for the Partitioned Global Address Space (PGAS) programming model) is supported, and its dependency `libverbs` has to be installed by a system administrator, we are unable to compare GPI_CP against ReStore.

IV. IN-MEMORY REPLICAS FOR FAST RECOVERY

ReStore allows application developers to store redundant copies of their data in-memory. In case of a failure, the surviving PEs can invoke a recovery routine to load all or parts of the data lost during the failure.

The remainder of this section is structured as follows: In Section IV-A we introduce our general framework for maintaining redundant copies of the user-supplied data in memory as well as the algorithm used for recovery. Section IV-B expands

on the distribution of copies by adding random permutations that accelerate the recovery algorithm. We analyze the memory usage of our proposed data distribution in Section IV-C and the probability of irrecoverable data loss in Section IV-D. In Section IV-E we describe a currently unimplemented approach to restore the level of redundancy after a failure.

A. General Framework

The main idea of ReStore is to store r copies of the data on different nodes. By storing them such that it is unlikely for all copies of one data element to fail at once, there will most likely (Section IV-D and Section VI-B1) be copies left to recover from. The application programmer can store data into ReStore using the *submit* and retrieve data from ReStore using the *load* function. To make the data addressable, we divide it into blocks where each block has a unique identifier.

Let n be the number of data blocks. In its most basic form (Figure 1), for $k \in [0, r)$ we store the block with ID x on PEs $L(x, k) = \lfloor \frac{xp}{n} \rfloor + k \cdot \frac{p}{r} \bmod p$. Under this distribution, we expect the copies of a block to not be stored on the same physical node/case/rack in most cluster setups. This decreases the probability of losing all copies of a block, as failures of PEs in the same node/case/rack are more likely to occur than a simultaneous failure of unrelated PEs [8]. In Section IV-B we explore a change to this basic distribution scheme that allows for faster recovery.

During recovery, when PE i requests to load block j , we choose one of the surviving PEs that hold block j at random to serve the request. If the user requests multiple successive blocks which are stored on the same set of PEs, we choose one PE to serve all requests. This strategy minimizes the bottleneck number of messages received. Next, we distribute the requested data using a custom sparse all-to-all communication.

B. Breaking Up Access Patterns for Faster Recovery

The goal of distributing data copies as described in Section IV-A is primarily to preserve the ability to recover from a fault. In the following, we explore how to adapt the data distribution such that it accelerates data recovery while preserving the level of failure resilience.

Assume a failed PE i which worked on data blocks $\left[i \frac{n}{p}, (i+1) \frac{n}{p} \right)$, where n is the total number of data blocks submitted to ReStore. If the application redistributes the lost data evenly to all surviving PEs, we would ideally want a dedicated sending PE for each receiver, resulting in a bottleneck communication volume of $\frac{n}{p^2}$ and a bottleneck number of messages received of 1. With the data distribution from Section IV-A, only the surviving subset of the $r \ll p$ PEs that hold copies of these blocks act as sources, resulting

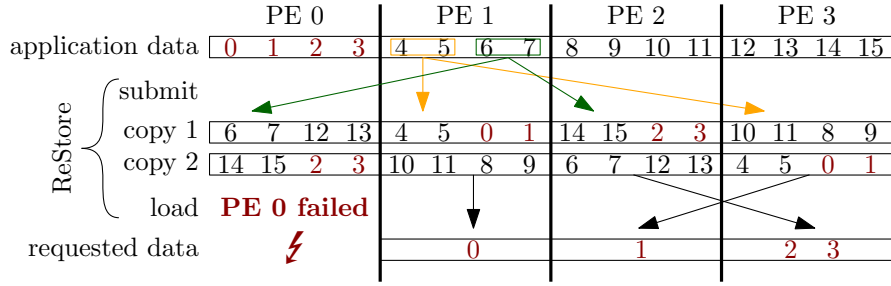


Fig. 2. Example showing the submit and load operations as well as the data distribution of the copies with a random permutation for $p = 4$ PEs, $n = 16$ data blocks, $r = 2$ copies, and $s_{pr} = 2$ blocks per permutation range. The first row shows the data submitted by the application. As an example, the orange and green arrows show the data ReStore sends from PE 1 to the target PEs which hold copies of the received data. After PE 0 fails, the application requests the data shown in the last row (dark red in all occurrences) which is served by ReStore as shown with the black arrows.

in a bottleneck communication volume of $\frac{n}{pr}$. We can alleviate this issue by evenly distributing the copies of data blocks $\left[i \frac{n}{p}, (i+1) \frac{n}{p}\right)$ among multiple PEs by randomly permuting the block identifiers. For a random permutation π and $k \in [0, r)$, the block x is stored on PEs

$$L(x, k) = \left\lfloor \frac{\pi(x) \cdot p}{n} \right\rfloor + k \cdot \frac{p}{r} \mod p$$

If the user requests blocks $\left[i \frac{n}{p}, (i+1) \frac{n}{p}\right)$, more PEs have parts of the data and can send them to the requesting PE. This approach, however, can lead to a large bottleneck number of messages being sent and received: If a PE requests $\frac{n}{p^2}$ data blocks, these blocks can reside on up to $\min(\frac{n}{p^2}, p)$ different PEs. To mitigate this, we group the data into *permutation ranges* of size s_{pr} . We then apply a random permutation to these permutation ranges instead of individual data blocks (Figure 2). If by a fault on PE i , data blocks $\left[i \frac{n}{p}, (i+1) \frac{n}{p}\right)$ need to be redistributed, these correspond to permutation ranges $\left[\frac{i \cdot n/p}{s_{pr}}, \frac{(i+1) \cdot n/p}{s_{pr}}\right)$. If we request the data to be evenly distributed among the $p-1$ surviving PEs such that PE j receives blocks $\left[i \frac{n}{p} + j \cdot \frac{n}{p \cdot (p-1)}, i \frac{n}{p} + (j+1) \cdot \frac{n}{p \cdot (p-1)}\right)$, only $\frac{n/(p \cdot (p-1))}{s_{pr}}$ PEs send to every receiving PE.

The best choice of s_{pr} and whether to use permutations at all depends on the data distribution, the expected amount of data lost by a fault, and the application's recovery strategy, but also on the frequency of checkpoint creation as submitting data with permutations enabled results in a more dense communication pattern. Section VI-B1 shows how we experimentally chose a good value for s_{pr} .

Note that with this data distribution, we always have sets of $\frac{n}{s_{pr}p}$ permutation ranges that are stored together for all r copies (e.g., blocks 6, 5, 12, 13 in Figure 2 are always stored together – once on PE 0 and once on PE 2). This means that for all copies of any permutation range whose first copy is stored on PE i to become lost, exactly the set of r PEs $i + k \cdot \frac{p}{r}$, $k \in [0, r)$ has to fail. One could also opt for a different approach – for example, using a distinct permutation for each copy. In this case, no sets of permutation ranges will always be stored together. So in order for any permutation range whose first copy resides on PE i to be lost, it is sufficient if *any* of the $\frac{n}{s_{pr}p}$ sets of PEs of

size r fail that hold the copies of one of the permutation ranges. In Section IV-D we analyze the probability of irrecoverable data loss under our proposed data distribution.

C. Memory Usage

Other fault-tolerance libraries [8], [13], [14] often use erasure coding – for example the Reed-Solomon code [34] – to reduce their memory footprint. This works for example by not storing the replicas A' and B' of two blocks A and B but rather the XOR of these blocks $A \oplus B$. We decide against using erasure coding as a means to reduce the memory footprint of our checkpoints, as this would incur additional messages upon checkpoint creation and recovery as well as a substantial computational overhead [35]. We therefore trade reduced communication overhead for increased memory consumption.

As in Section IV-A, let n be the number of data blocks, r be the number of replicas and p be the number of processes. On each PE ReStore requires main-memory to store $\frac{rn}{p}$ data blocks for the replicated storage. The memory requirement is doubled during submission as we require additional space for the send and receive buffers. During recovery, an additional copy of all data being sent and received is stored on each PE. We verified these formulas empirically (data not shown). A plethora of applications exist for which the amount of memory for the input data *and* the data that need to be checkpointed fit in memory r times. Examples include RAXML-NG [22], [36], k -means, and page-rank.¹ For example, RAXML-NG is memory bandwidth bound [37]. Hence, using additional cores with their associated larger cache memory capacity can even yield super-linear speedups due to increased cache-efficiency. Such applications can therefore substantially benefit from the reduced communication and computational overhead to create and restore a checkpoint without erasure codes.

D. Probability of Irrecoverable Data Loss.

Let r be the replication level and p be the number of PEs. In this analysis, we assume that $r|p$ (r divides p). This constitutes a reasonable assumption for current two socket systems that exhibit an even number of cores per socket and $r = 4$. If

¹We implemented fault-tolerant version for all three of these using ReStore and show running times for RAXML-NG and k -means in Section VI-C.

$r|p$, the PEs are divided into $g = \frac{p}{r}$ groups, with all PEs in a respective group storing the same data. Thus, if and only if all r PEs in a specific group fail, we will not be able to recover a part of the data. We denote such an event as Irrecoverable Data Loss (IDL). Let f be the number of failed PEs. There is exactly one possibility to draw r out of r PEs belonging to a single group. The number of possibilities to draw the remaining $f - r$ failed PEs among the remaining PEs such that they do not belong to the given group is $\binom{p-r}{f-r}$. The overall number of possibilities to draw f PEs from the p PEs that are still alive at program start is $\binom{p}{f}$. The probability that, given f failures, all processes of a *given* group fail is thus $1 \cdot \binom{p-r}{f-r} / \binom{p}{f}$. When generalizing this equation to the probability of all processors of *at least one* group failing, we have to apply the inclusion-exclusion principle to avoid counting the same combination multiple times. We thus obtain the following equation for the probability of an IDL at failure f or any failure before:

$$P_{\text{IDL}}^{\leq}(f) = \sum_{j=1}^g (-1)^{j+1} \binom{g}{j} \frac{\binom{p-jr}{f-jr}}{\binom{p}{f}}$$

The probability of an IDL at exactly failure f is thus:

$$P_{\text{IDL}}^{\equiv}(f) = P_{\text{IDL}}^{\leq}(f) - P_{\text{IDL}}^{\leq}(f-1)$$

The expected number of failures until an IDL occurs is:

$$E[\text{Failures until IDL}] = \sum_{f=r}^p P_{\text{IDL}}^{\equiv}(f) \cdot f$$

For small f , the *approximate* probability of all PEs of any group failing is given by $P_{\text{IDL}}^{\text{approx.}}(f) = g \cdot (f/p)^r$. Solving for the fraction of PEs that fail f/p such that $P_{\text{IDL}}^{\text{approx.}} = 1$ yields $f/p = (r/p)^{(1/r)} \in \mathcal{O}(p^{-1/r})$ for a fixed r .

In Section VI-B1 we simulate node failures using the actual data distribution to verify these formulas.

E. Recovering Lost Replicas After a Node Failure

To further increase the resilience of our framework, we introduce an approach to restore replicas that were lost upon a failure while keeping all other replicas in place. That is, we do not need to redistribute any replicas that reside on surviving nodes. As in the previous sections, let n be the number of data blocks, r be the number of replicas per block, and p be the number of nodes.

We draw a different random permutation ρ_x of $[0, p-1]$ (or long, non-repeating random sequences of nodes) for each block x and place the replicas of x on the first r alive nodes of that permutation. When a node dies, we copy all replicas that this node held to the next node in each replicas permutation. We can refine this approach to attain a perfectly balanced initial data distribution and reduce the probability of an IDL (Section IV-B): We initially place the first r replicas ($L(x, 0), L(x, 1), \dots, L(x, r-1)$) deterministically as described in Section IV-A. That is, the data distribution is given by

$$L(x, k) = \begin{cases} \text{as described in Section IV-A,} & \text{if } k < r \\ \rho_x(k), & \text{else} \end{cases}$$

See the Appendix for different options to draw such a sequence of nodes. Following this data distribution, we can compute the ranks on which we store a given block in $\mathcal{O}(r+f)$ time and $\mathcal{O}(1)$ space where r is the number of replicas per block and f is the number of node failures. In order to keep recovery fast, we can apply this technique on a permutation-range-level rather than on individual blocks as explained in Section IV-B.

V. IMPLEMENTATION

We implement ReStore as a C++ library² using the User Level Failure Mitigation (ULFM) proposal implementation [16]. Application programmers submit their data to ReStore by writing their serialized data blocks to a memory location supplied by the library or using ReStore's interface for already serialized data. After a failure, they can request data blocks by passing a list of ranges of block identifiers to ReStore. This can be done in two ways: Either by providing the full list of requested block IDs on all PEs or by providing exactly those ID ranges each individual PE needs on exactly that PE. By using the first approach, no communication is required to determine which PE serves which request. When using the second approach, the receiving PE will determine which PE should send each requested data block. Then, a sparse all-to-all communication is performed to issue the requests to the sending PEs. Preliminary experiments showed that the latter method performs substantially better because the full list of requests usually scales with the number of PEs in the application, slowing down the first approach. Thus, for all experiments in Section VI, we employ the second approach. As ReStore's implementation currently focuses on fast recovery, it provides only an interface for submitting data once and thus is currently not suitable for repeatedly checkpointing changing data. This is sufficient for many fault-tolerant applications – two of which are demonstrated in Section VI-C.

VI. EXPERIMENTAL EVALUATION

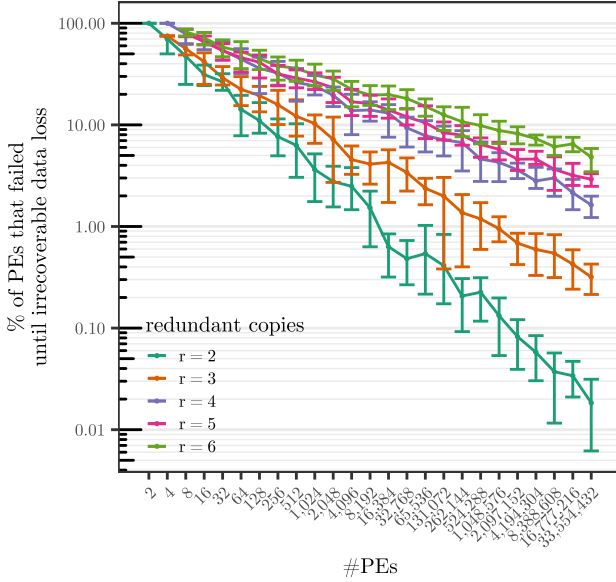
In this section we present the results of our experimental evaluation. We present the experimental environment in Section VI-A. In Section VI-B we evaluate ReStore's fault resilience and performance in isolation. In Section VI-C we show how ReStore performs when used in two fault-tolerant applications: A simple k -means algorithm and a complex bioinformatics application used by thousands of researchers. Finally, in Section VI-D we compare ReStore to reading from a parallel file system (PFS) – which represents a lower bound for checkpointing libraries using the PFS as storage – as well as the reported running times by other checkpointing libraries.

A. Environment and Experimental Setup

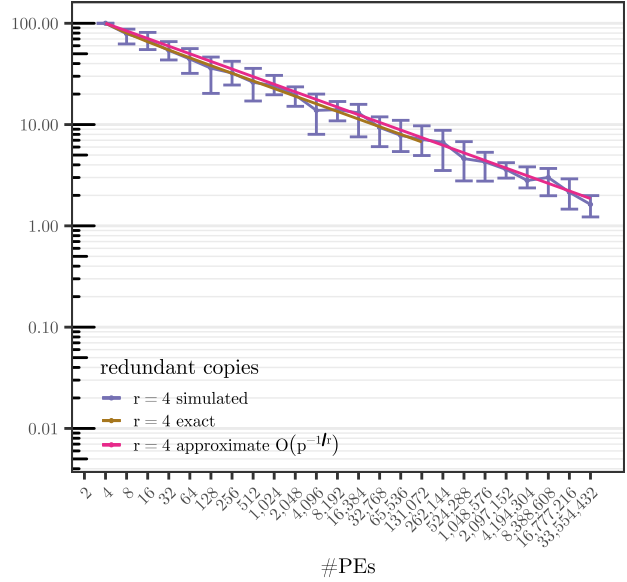
We run our experiments on the SuperMUC-NG super computer.³ Each node consists of two Intel Skylake Xeon Platinum 8174 processors with 24 cores and 96 GB of memory each, connected via an OmniPath network with a bandwidth of 100 Gbit/s. The operating system is SUSE Linux Enterprise

²<https://github.com/ReStoreCpp/ReStore>

³<https://doku.lrz.de/display/PUBLIC/SuperMUC-NG>



(a) Simulation (Section VI-B1)



(b) Theory (Section IV-D)

Fig. 3. Percentage of failed PEs until all redundant copies of one data block are lost. (a) Simulation of the data distribution described in Section IV-A. We continue simulating the failure of random PEs until there is at least one data block with no remaining copies on the surviving PEs. (b) Comparison of the probability given by the equations in Section IV-D and the simulated values from (a).

Server 15 SP1 running Linux Kernel version 4.12.14-197.78. We compile our benchmark applications using gcc version 10.2.0 with full optimizations enabled (`-O3`) and all assertions disabled. Unless otherwise stated, we communicate using OpenMPI version 4.0.4. We verify that our implementation *does work* if nodes actually fail and communication is recovered with ULFM as part of our fully automated unit tests. The current version of ULFM, however, is not stable enough to conduct reliable performance benchmark experiments. For example, processes may be reported incorrectly as failed or recovery may result in two separate groups of nodes that each assume that the other group has failed. We reported this behavior on the ULFM mailing list and the authors of ULFM reproduced and confirmed the bug.⁴ Additionally, most communication and fault tolerance mechanisms are currently slow (see Hübner et al. [22] for details). We expect these issues in ULFM to be fixed once fault-tolerance is part of the MPI standard and more resources are allocated to implementing these features. In our performance benchmarks, we thus use OpenMPI and simulate failures by removing processes from the calculation using `MPI_Comm_split` and replacing other required fault recovery steps by functionally similar ones (e.g., replacing `MPICH_Comm_agree` with `MPI_Barriers`). As ReStore currently only supports submitting data once, all experiments shown in this section submit only their input data. This is a restriction in the API, not the underlying algorithm, and will be removed in future work.

All plots show results for 10 repetitions per experiment.

⁴George Bosilca. Post `pbSToy94RhI/xUrFBx_1DAAJ` on the ULFM mailing list.

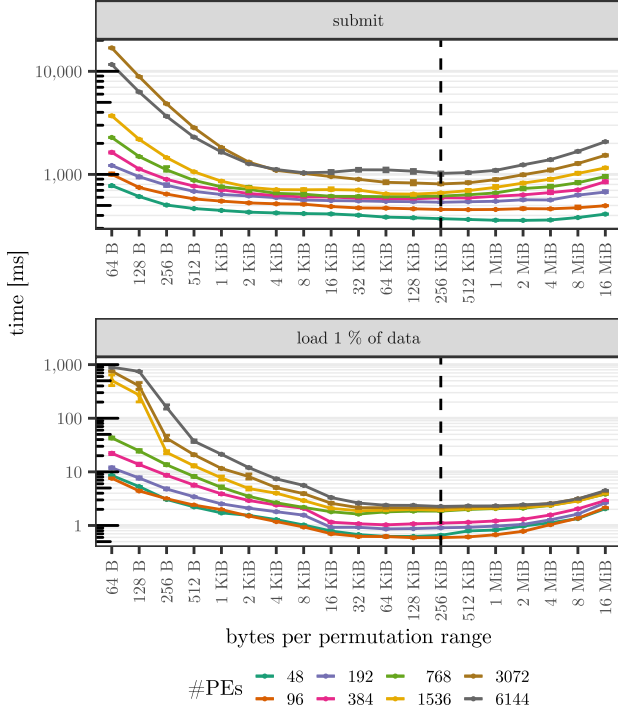
Plots depict the mean with error bars for the 10th and 90th percentile.

B. Isolated Evaluation

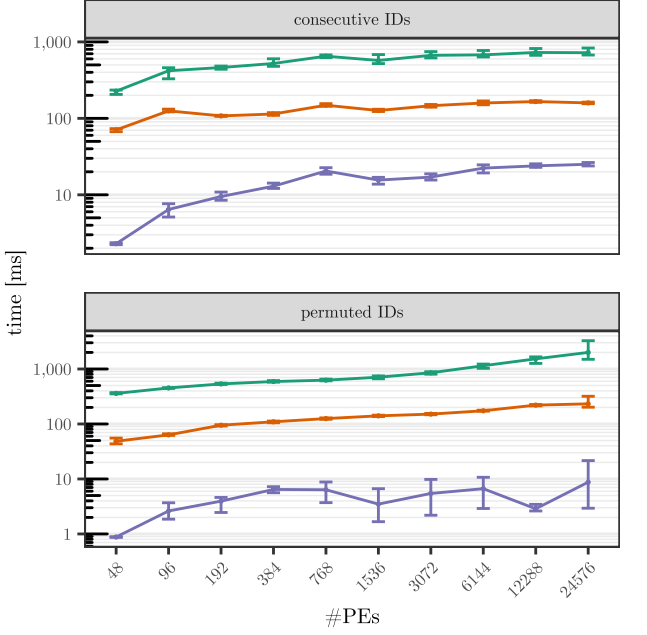
In this section we explore ReStore in isolation. We first choose the number of redundant copies (Section IV-A). Next, we analyze ReStore’s performance and experimentally optimize the size of permutation ranges (Section IV-B).

1) *Number of Redundant Copies.*: Figure 3a shows the result of a simulation of our data distribution: We continue simulating the failure of random PEs until at least for one data block no copies remain on the surviving PEs. We can see that even for 2^{25} PEs, more than 1% of all PEs have to fail until we can no longer recover all data when using $r = 4$ redundant copies. Even in the event of an irrecoverable data loss, the program will not crash, but will have to merely reload the input data from disk. For applications running on fewer PEs, an even smaller number r of redundant copies is sufficient to yield data loss unlikely. For all further experiments we therefore set the number of redundant copies to $r := 4$. We compare the values obtained by applying the formula in Section IV-D with the values obtained by simulation in Figure 3b showing that our theoretical formula matches the simulation very closely.

2) *Performance.*: For all experiments in this section we use data blocks of size 64 B and 16 MiB of data per PE. We show results for three different operations: In the `submit` operation we pass 16 MiB on all PEs to ReStore’s `submit` function. In `load 1%` data we load the data submitted by 1% of the PEs with contiguous data block IDs evenly across all PEs. So for n total data blocks, we pick a random starting PE i and request data blocks $i \cdot n/p$ to $(i + 0.01 \cdot p) \cdot n/p$. This



(a) Optimizing the number of bytes per permutation range



(b) Randomized vs consecutive IDs

Fig. 4. (a) Influence of the number of bytes per permutation range on the running time of submitting to and loading from ReStore. (b) Weak scaling experiment (16 MiB per PE) of our three benchmark operations with and without ID randomization. We copy all data over the network – i.e., no rank holds a copy of its requested data in its local part of the ReStore storage.

simulates the requests expected if 1% of PEs fail at once. In `load all data` we load all data stored in ReStore evenly distributed across all PEs in a way that no PE loads the same data it originally submitted.

By decreasing the size of permutation ranges (Section IV-B) we can control how many PEs can participate in sending requested data: When using smaller permutation ranges, more PEs are able to serve parts of the requested data to the requesting PEs. Small permutation ranges, on the other hand, lead to fragmentation of the data and therefore induce many small messages. In Figure 4a we show the number of bytes per permutation range on the x -axis and the running times of `submit` and `load 1% data` on the y -axis for different numbers of PEs. We do not show results for `load all data` because permutations even have a negative effect on performance here (Figure 4b). We therefore recommend turning them off when using a recovery mechanism which loads all data stored in ReStore. We observe that for few bytes per permutation range, both `submit` and `load 1% data` are slower by up to an order of magnitude than the fastest configuration because of a high bottleneck number of messages. Approaching 16 MiB of data per permutation range, fewer PEs can participate in sending data. Between these two extremes, there is a range of permutation range sizes which yield fast running times. For all further experiments, we thus fix the amount of data per permutation range to 256 KiB (0.65

to 2.27ms for `load 1% data` on 48 to 6144 PEs). For 16 MiB of data per PE, every PE receives approximately 164 KiB in `load 1% data` which results in an average of two PEs requesting the same permutation range and therefore induces a sparse communication pattern. On the sending side, this implies that the data submitted by a single PE is distributed among 64 permutations ranges. With $r = 4$ redundant copies this results in up to $64 \cdot 4 = 256$ PEs that participate in serving this part of the data.

As expected, enabling random permutations speeds up `load 1% data` and slows down `load all data`, especially for runs on many PEs (Figure 4b). This is because in `load all data`, even without permutations, every PE sends some part of the data. By enabling permutations, the data requested by a PE is distributed among more sending PEs, resulting in a denser communication pattern. We can tolerate an increase in running time of `submit` as it is called only once in the case of only submitting input data. In contrast, a load is issued after every failure.

C. Applications

To demonstrate realistic use cases of ReStore we use it to restore lost data in two different real-world applications. Figure 5 shows running times for a small example application

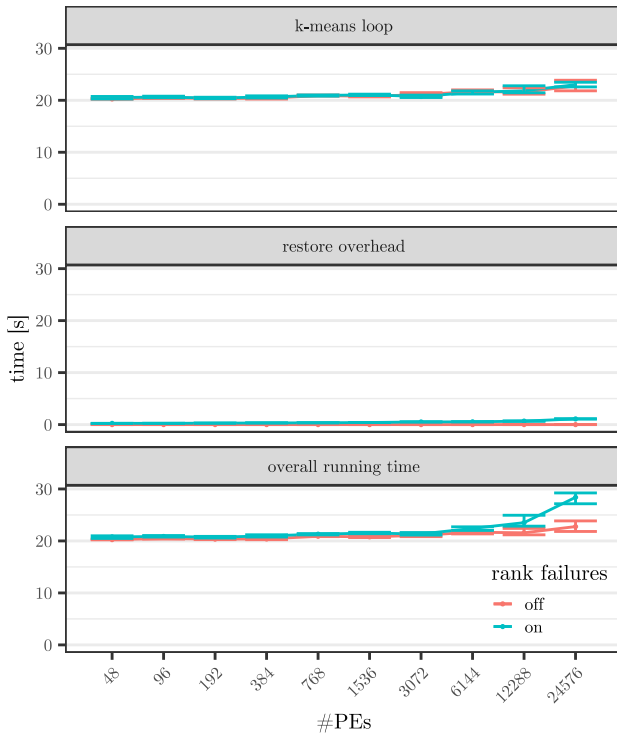


Fig. 5. Running time of the k -means clustering algorithm with and without failures on 16 MiB of data per PE. *k-means loop*: time spent for the core clustering algorithm. *Restore overhead*: time spent in ReStore’s functions. *Overall running time* also includes additional work required for attaining fault-tolerance, such as a load balancer to determine how to redistribute data and MPI functions for identifying the failed PEs.

that computes a k -means clustering [38].⁵ Each PE holds 65 536 points in a 32-dimensional space as input with 8 byte double precision floating point values per dimension resulting in 16 MiB of input data. All PEs start with the same 20 random starting centers. Iteratively, each PE assigns the nearest center to each of its local points and all PEs collaboratively calculate new centers positions using an all-reduce-operation over k elements. If a PE fails, the remaining PEs divide the dead PE’s data points evenly among them using ReStore and continue with the calculation. We perform 500 iterations of the algorithm and simulate an expected failure of 1 % of all nodes distributed uniformly at random during these iterations. This is done by determining a suitable probability for each PE to fail in each iteration of the algorithm.⁶ We find that ReStore accounts for only 1.6 % (median) of the overall running time on up to 24 576 PEs with up to 262 PEs failing. Note that the overall running time increases by more than ReStore’s overhead for large PE counts mainly due to MPI operations used to restore a functioning communicator after a PE failure.

Next, we demonstrate ReStore’s performance for the fault-tolerant version of the highly complex and widely used

phylogenetic tree inference software RAxML-NG [36] – called FT-RAxML-NG [22]⁷ – using the same empirical datasets as in [22] (Figure 6a). Additionally, we use a 19.1 GiB synthetic dataset [39] for scaling experiments (Figure 6b). FT-RAxML-NG redistributes its input data among all surviving PEs. We therefore deactivate permutation ranges for this application. We compare ReStore’s performance against FT-RAxML-NG’s currently implemented recovery mechanism: Loading the data from the PFS using RAxML-NG’s dedicated binary file format (RBA) which enables rapidly reading only the required subset of the input matrix. We distinguish between the input files being uncached by the file system (in the first read) and being cached by previous reads. Both, submitting data to ReStore and loading data after a failure, is faster than the original method of loading the data from files – often by more than an order of magnitude. On the synthetic data set, for low PE counts, submitting to ReStore is slower than reloading from a file. However, this is negligible because an actual phylogenetic inference on this dataset requires terabytes of memory for likelihood calculations and would therefore never run on this few nodes. We also want to emphasize that submitting to ReStore has to be done only once in FT-RAxML-NG, while loading has to be conducted after every failure.

D. Comparison with Other Approaches

We now evaluate ReStore’s performance in comparison to other checkpointing approaches. As shown in Section III, most checkpointing libraries store checkpoints on the PFS. Most on-disk and all in-memory checkpointing libraries support only substituting recovery, i.e., no shrinking recovery, yielding a comparison with ReStore challenging. Additionally, as detailed in Section III-A, to the best of our knowledge, there exists no in-memory checkpointing library which is still maintained and working to compare ReStore to.

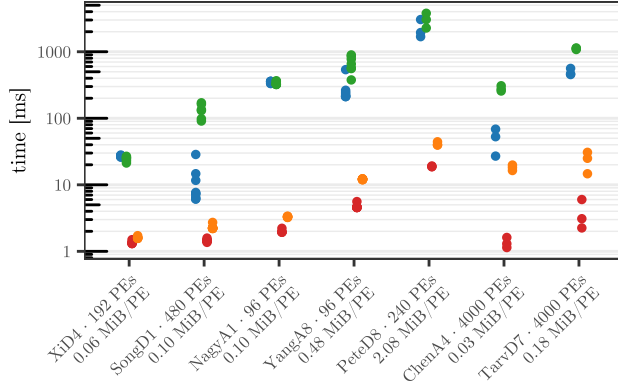
1) *Comparing to Disk-Based Approaches*: We compare ReStore against loading a copy of the data stored on the PFS (Figure 7). We create this file such that reading is a single consecutive read and therefore as fast as possible. We show running times for reading a separate file for each reading PE using C++’s `ifstream` and reading a single file for all PEs with `MPI_File_read_at_all` (MPI I/O in the plot). This is a lower bound for all checkpointing libraries that have to read their data from disk. ReStore outperforms disk access (`ifstream`) on 24 576 PEs by a factor of 206 (1 % of data; median) and 55 (all data) respectively.

2) *Comparing to Reported Measurements*: Gamell et al. [3] measure approximately 115 ms to write a checkpoint with 14.8 MB per rank on 1000 ranks using Fenix. Fenix implements a replication level of $r = 1$. This means, that there exists a single copy of the data in addition to the data the ranks are actively working on. According to our definition (Section IV-D), a single rank failure will cause irrecoverable data loss. This works in practice, as long as the data which resided on the failed rank(s) does not need to be restored. To serialize and store

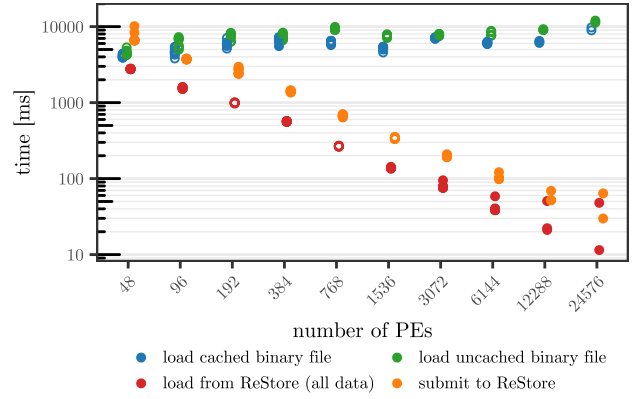
⁵We ran these experiments with IntelMPI, because its `Group_*`-functions – which we use to determine which PEs failed – perform better than OpenMPI’s.

⁶Using a discrete exponential decay with 1 % of failed PEs after 500 iterations.

⁷<https://github.com/lukashuebner/ft-raxml-ng/tree/restore-paper>



(a) Real world datasets



(b) 19.1 GiB synthetic dataset

Fig. 6. Performance of data loading after a fault in FT-RAXML-NG. In subplot (a), labels on the x -axis show the name of the data set, the number of PEs used for that data set, and the corresponding amount of input data per PE.

16 MiB per rank on 1536 ranks (32 nodes) with a replication level of $r = 1$ and using consecutive IDs, ReStore needs 126 ± 3 ms ($\mu \pm \sigma$, 10 repeats). Gamell et al. [3] expect Fenix’s recovery time to be the same as its checkpointing time but do not provide experimental results for that claim. ReStore restores the data of a single rank to another single rank in our experiments in 21 ± 2 ms. ReStore additionally offers to restore the data of a single rank scattered to all surviving ranks. This operation requires 20 ± 5 ms in our experiments. In the case that one expects more than one recovery per checkpoint ReStore offers ID permutations to speed up recovery at the cost of slower checkpoint creation (Section IV-B). This would for example be the case for static input data, of which multiple ranks need different but small fractions after getting assigned new work following a rank failure. With ID permutations enabled, saving the data to ReStore takes 215 ± 9 ms in our experiments. Restoring the data takes 15 ± 3 ms if restoring all the data to a single rank and 0.9 ± 0.2 ms if restoring the data scattered across the surviving ranks. As the latter evenly distributes the data across the surviving ranks, we expect it to become the more common scenario when working in a shrinking setting.

Bartsch et al. [15] report GPI_CP to require approximately 1 s to initialize, 200 ms to create a checkpoint and 15 ms to restore data from a checkpoint.

Fenix’s performance was measured on a Cray XK7 system with 16 cores per node and a 160 GB s^{-1} network [40]. GPI_CP’s performance was measured on an unnamed system with 16 cores per node and a QDR Infiniband network. We measured ReStore’s performance on the SuperMUC-NG, which has 48 cores per node and an OmniPath interconnection with 100 Gbit s^{-1} (Section VI-A). We choose our experiments such that data is always copied between different nodes and never between two processes running on the same node. Thus, all 48 processes on a single node have to share the same interconnect. Considering that we evaluate ReStore on a slower network

than Fenix, we expect an even more favourable comparison when having access to a similar HPC system.

Lu [14] reports checkpoint creation times of 8 to 20 s for 157 to 182 MB on 448 ranks. They report restoration times of 20 to 48 s. Thus, assuming linear scaling, we expect checkpoint creation times of approximately 1 s and restoration times of approximately 2 s for 16 MiB of data. Lu’s algorithm is thus an order of magnitude slower than ReStore and Fenix. We assume this is due to the fact, that Lu’s algorithm uses erasure codes (Section IV-C).

To summarize, ReStore can be configured to create and restore from checkpoints in the same manner and approximately the same time as existing checkpointing solutions. ReStore additionally has functionality to (a) increase the replication level (b) restore the data in a scattered manner to multiple ranks instead of to one rank and (c) enable ID permutations to decrease time to restore the data by an order of magnitude while doubling the time taken to create a checkpoint. The latter option is for example useful when creating a replicated storage for the input data of a program, which has to be partially reload after a failure.

VII. CONCLUSION AND FUTURE WORK

We show that by using a suitable data distribution strategy, recovery of lost data after a failure is possible in tens to hundreds of milliseconds, depending on the amount of data loaded. We achieve this by using a distribution scheme for redundant copies that ensures a low probability of data loss and a rapid recovery of the data. We also provide the – to the best of our knowledge – first in-memory checkpointing library which supports shrinking recovery, that is ReStore is able to restore the data of the failed PEs scattered to multiple or all surviving ranks instead of to a single respawned or spare PE. This alleviates the need for the application to allocate spare nodes which participate in the computation only in case of a node failure, thus increasing computation efficiency. We supply an analysis of the probability of irrevocable data loss and

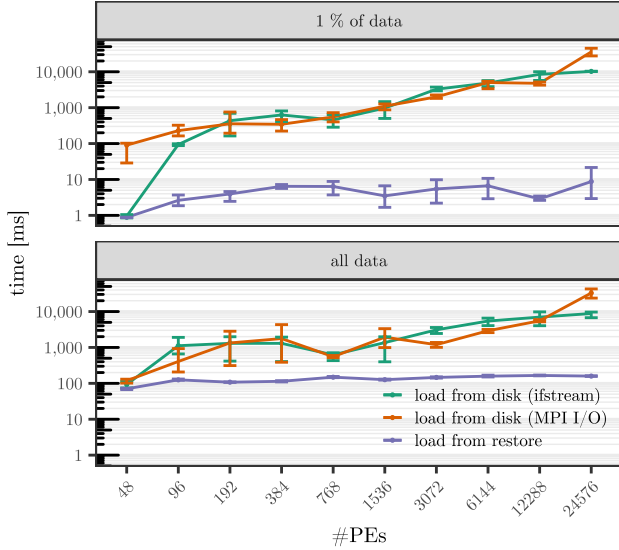


Fig. 7. Loading performance of ReStore vs. loading from files on the clusters parallel file system, representing the approach of most checkpointing libraries.

propose a data distribution to easily restore lost replicas after a failure. Experimental and theoretical evaluation of the proposed data redistribution after a node failure constitutes part of future work. This further decreases the probability to lose all copies of any data. With our C++ library, we were able to improve recovery performance of FT-RAXML-NG [22], [36] by up to two orders of magnitude. By using the proposal implementation of the fault-tolerance mechanisms included in the new MPI 4.0 standard, our library can be used by applications on HPC systems once the new standard is implemented. We also plan on evaluating ReStore for checkpointing of dynamic program state and extend its API for different data formats (e.g. 2D data).

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer SuperMUC-NG at Leibniz Supercomputing Centre (www.lrz.de). Part of this work was funded by the Klaus Tschira foundation. This work was supported by a grant from the Ministry of Science, Research and the Arts of Baden-Württemberg (Az: 33-7533.-9-10/20/2) to Peter Sanders and Alexandros Stamatakis. This project has received funding from the European Research Council (ERC)⁸ under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 882500). We thank an anonymous reviewer for pointing out that the simple approximation of the irrecoverable data loss given in Section IV-D is very accurate for small f .

Let $r = 3$ and $p = 500$ with prime factors 2 and 5.
The first hash function evaluates to $f(x) = f(y) = 1$.
Further $h_s(x) = 3$ ✓ (coprime to p)
and $h_s(y) = 20$ ✗ (not coprime to p), $h_{s'}(y) = 7$ ✓

Block x is stored on	Block y is stored on
PE $f(x) + 0 \cdot h_s(x) = 1$ ✓	PE $f(y) + 0 \cdot h'_s(y) = 1$ ✓
PE $f(x) + 1 \cdot h_s(x) = 4$ ✓	PE $f(y) + 1 \cdot h'_s(y) = 8$ ✗
PE $f(x) + 2 \cdot h_s(x) = 7$ ✓	PE $f(y) + 2 \cdot h'_s(y) = 15$ ✓
	PE $f(y) + 3 \cdot h'_s(y) = 22$ ✓

✓ PE is alive; ✗ PE failed

Fig. 8. A modified data distribution enabling easy recovery of replicas (Data Distribution A). r denotes the number of replicas, p the number of nodes, f and h are hash functions.

APPENDIX

CHOOSING A RANDOM PERMUTATION IN SECTION IV-E

1) *Data Distribution A*: Let f and h_s be fast-to-compute hash functions that avoid collisions, where s is a seed used to parametrize h_s . We compute the nodes where a block x is located by evaluating $\rho_x(k) = (f(x) + k \cdot h_s(x)) \bmod p$ for $k = 0, 1, \dots$ until we find r nodes that are still alive (Figure 8). This scheme is analogous to collision resolution in hash tables using open addressing and double hashing. The probing sequences ρ_x and ρ_y are likely to be different for two blocks x and y , even if $\rho_x(0) = \rho_y(0)$.

If $L(x, k) = L(x, j)$ for $k < j < p$, the entire sequence will repeat from that point on, yielding recovery of lost replicas impossible after more than j failures. We can avoid this by requiring the seeded hash function h_s to only yield integers that are coprime to p . We archive this by trying different seeds s for h_s until $h_s(x)$ is coprime to p . For this purpose, we draw a sequence of seeds at program startup. The probability of two random integers being coprime is $6/\pi^2$ [41]. For a random⁹ p , the expected number of differently seeded hash function we have to evaluate is thus

$$1 + \sum_{n=1}^{\infty} \left(1 - \frac{6}{\pi^2}\right)^n = \frac{7}{6} (\pi^2 - 6) \approx 1.65$$

To check if $h_s(x)$ is coprime to p , we divide $h_s(x)$ by every prime factor of p , which we factorize once during program startup. The Erdős-Kac theorem [42] states that the number of distinct prime factors m of a random number below \hat{p} approximately follows the normal distribution with a mean and variance of $\log \log \hat{p}$. For example, a node count of $p < 10^9$, has on average $m = 3 \pm 1.7$ prime factors. For node counts $p < 10^9$ ($m = 3$) we therefore expect $< m \cdot 1.65 = 5$ divisions to check for coprimality each time we need to compute on which nodes to store a block.

2) *Data Distribution B*: Alternatively, we can deploy a classical seeded pseudorandom permutation ρ_s on $[0, p-1]$ to generate an independent probing sequence for each element x . Let f be a collision avoiding hash function. We then use $f(x)$ as the seed for ρ . As ρ we could for example use a Feistel Network-based permutation with $f(x)$ as the seed and cycle walking for values of p which are not a power of two.

⁹Though realistically, we cannot expect p to be random.

REFERENCES

- [1] L. Hübner, D. Hespe, P. Sanders, and A. Stamatakis, "Restore: In-memory replicated storage for rapid recovery in fault-tolerant algorithms," in *2022 IEEE/ACM 12th Workshop on Fault Tolerance for HPC at eXtreme Scale (FTXS)*, 2022, pp. 24–35. [Online]. Available: <https://doi.org/10.1109/FTXS56515.2022.00008>
- [2] J. Shalf, S. S. Dosanjh, and J. Morrison, "Exascale computing technology challenges," in *High Performance Computing for Computational Science - VECPAR 2010 - 9th International conference, Berkeley, CA, USA, June 22-25, 2010, Revised Selected Papers*, ser. Lecture Notes in Computer Science, J. M. L. M. Palma, M. J. Daydé, O. Marques, and J. C. Lopes, Eds., vol. 6449. Springer, 2010, pp. 1–25. [Online]. Available: https://doi.org/10.1007/978-3-642-19328-6_1
- [3] M. Gamell, D. S. Katz, H. Kolla, J. Chen, S. Klasky, and M. Parashar, "Exploring automatic, online failure recovery for scientific applications at extreme scales," in *International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2014, New Orleans, LA, USA, November 16-21, 2014*, T. Damkroger and J. J. Dongarra, Eds. IEEE Computer Society, 2014, pp. 895–906. [Online]. Available: <https://doi.org/10.1109/SC.2014.78>
- [4] F. Cappello, A. Geist, W. Gropp, S. Kale, B. Kramer, and M. Snir, "Toward exascale resilience: 2014 update," *Supercomput. Front. Innov.*, vol. 1, no. 1, pp. 5–28, 2014. [Online]. Available: <https://doi.org/10.14529/jsfi140101>
- [5] J. Dongarra, T. Herault, and Y. Robert, *Fault-Tolerance Techniques for High-Performance Computing*, ser. Computer Communications and Networks. Springer, Cham, 2015, no. 1.
- [6] M. Snir, R. W. Wisniewski, J. A. Abraham, S. V. Adve, S. Bagchi, P. Balaji, J. F. Belak, P. Bose, F. Cappello, B. Carlson, A. A. Chien, P. Coteus, N. DeBardeleben, P. C. Diniz, C. Engelmann, M. Erez, S. Fazzari, A. Geist, R. Gupta, F. Johnson, S. Krishnamoorthy, S. Leyffer, D. Liberty, S. Mitra, T. S. Munson, R. Schreiber, J. Stearley, and E. V. Hensbergen, "Addressing failures in exascale computing," *Int. J. High Perform. Comput. Appl.*, vol. 28, no. 2, pp. 129–173, 2014. [Online]. Available: <https://doi.org/10.1177/1094342014522573>
- [7] S. Agarwal, R. Garg, M. S. Gupta, and J. E. Moreira, "Adaptive incremental checkpointing for massively parallel systems," in *Proceedings of the 18th Annual International Conference on Supercomputing, ICS 2004, Saint Malo, France, June 26 - July 01, 2004*, P. Feautrier, J. R. Goodman, and A. Sez nec, Eds. ACM, 2004, pp. 277–286. [Online]. Available: <https://doi.org/10.1145/1006209.1006248>
- [8] L. A. Bautista-Gomez, S. Tsuboi, D. Komatitsch, F. Cappello, N. Maruyama, and S. Matsuoka, "FTI: high performance fault tolerance interface for hybrid systems," in *Conference on High Performance Computing Networking, Storage and Analysis, SC 2011, Seattle, WA, USA, November 12-18, 2011*, S. A. Lathrop, J. Costa, and W. Kramer, Eds. ACM, 2011, pp. 32:1–32:32. [Online]. Available: <https://doi.org/10.1145/2063384.2063427>
- [9] F. Shahzad, J. Thies, M. Kreutzer, T. Zeiser, G. Hager, and G. Wellein, "CRAFT: A library for easier application-level checkpoint/restart and automatic fault tolerance," *IEEE Trans. Parallel Distributed Syst.*, vol. 30, no. 3, pp. 501–514, 2019. [Online]. Available: <https://doi.org/10.1109/TPDS.2018.2866794>
- [10] B. Nicolae, A. Moody, E. Gonsiorowski, K. Mohror, and F. Cappello, "Veloc: Towards high performance adaptive asynchronous checkpointing at large scale," in *2019 IEEE International Parallel and Distributed Processing Symposium, IPDPS 2019, Rio de Janeiro, Brazil, May 20-24, 2019*. IEEE, 2019, pp. 911–920. [Online]. Available: <https://doi.org/10.1109/IPDPS.2019.00099>
- [11] K. Teranishi and M. A. Heroux, "Toward local failure local recovery resilience model using MPI-ULFM," in *21st European MPI Users' Group Meeting, EuroMPI/ASIA '14, Kyoto, Japan - September 09 - 12, 2014*, J. J. Dongarra, Y. Ishikawa, and A. Hori, Eds. ACM, 2014, p. 51. [Online]. Available: <https://doi.org/10.1145/2642769.2642774>
- [12] A. Moody, G. Bronevetsky, K. Mohror, and B. R. de Supinski, "Design, modeling, and evaluation of a scalable multi-level checkpointing system," in *Conference on High Performance Computing Networking, Storage and Analysis, SC 2010, New Orleans, LA, USA, November 13-19, 2010*. IEEE, 2010, pp. 1–11. [Online]. Available: <https://doi.org/10.1109/SC.2010.18>
- [13] M. Besta and T. Hoefler, "Fault tolerance for remote memory access programming models," in *The 23rd International Symposium on High-Performance Parallel and Distributed Computing, HPDC'14, Vancouver, BC, Canada - June 23 - 27, 2014*, B. Plale, M. Ripeanu, F. Cappello, and D. Xu, Eds. ACM, 2014, pp. 37–48. [Online]. Available: <https://doi.org/10.1145/2600212.2600224>
- [14] C.-D. Lu, "Scalable diskless checkpointing for large parallel systems," phdthesis, University of Illinois at Urbana-Champaign, 2005.
- [15] V. Bartsch, R. Machado, D. Merten, M. Rahn, and F. Pfreundt, "GASPI/GPI in-memory checkpointing library," in *Euro-Par 2017: Parallel Processing - 23rd International Conference on Parallel and Distributed Computing, Santiago de Compostela, Spain, August 28 - September 1, 2017, Proceedings*, ser. Lecture Notes in Computer Science, F. F. Rivera, T. F. Pena, and J. C. Cabaleiro, Eds., vol. 10417. Springer, 2017, pp. 497–508. [Online]. Available: https://doi.org/10.1007/978-3-319-64203-1_36
- [16] W. Bland, A. Bouteiller, T. Hérault, G. Bosilca, and J. J. Dongarra, "Post-failure recovery of MPI communication capability: Design and rationale," *Int. J. High Perform. Comput. Appl.*, vol. 27, no. 3, pp. 244–254, 2013. [Online]. Available: <https://doi.org/10.1177/1094342013488238>
- [17] R. A. Ashraf, S. Hukerikar, and C. Engelmann, "Shrink or substitute: Handling process failures in HPC systems using in-situ recovery," in *26th Euromicro International Conference on Parallel, Distributed and Network-based Processing, PDP 2018, Cambridge, United Kingdom, March 21-23, 2018*, I. Merelli, P. Liò, and I. V. Kotenko, Eds. IEEE Computer Society, 2018, pp. 178–185. [Online]. Available: <https://doi.org/10.1109/PDP2018.2018.00032>
- [18] M. M. Ali, P. E. Strazdins, B. Harding, and M. Hegland, "Complex scientific applications made fault-tolerant with the sparse grid combination technique," *Int. J. High Perform. Comput. Appl.*, vol. 30, no. 3, pp. 335–359, 2016. [Online]. Available: <https://doi.org/10.1177/1094342015628056>
- [19] M. Obersteiner, A. Parra-Hinojosa, M. Heene, H. Bungartz, and D. Pflüger, "A highly scalable, algorithm-based fault-tolerant solver for gyrokinetic plasma simulations," in *Proceedings of the 8th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems, ScalA@SC 2017, Denver, CO, USA, November 13, 2017*, V. N. Alexandrov, A. Geist, and J. J. Dongarra, Eds. ACM, 2017, pp. 2:1–2:8. [Online]. Available: <https://doi.org/10.1145/3148226.3148229>
- [20] I. Laguna, D. F. Richards, T. Gambelin, M. Schulz, B. R. de Supinski, K. Mohror, and H. Pritchard, "Evaluating and extending user-level fault tolerance in MPI applications," *Int. J. High Perform. Comput. Appl.*, vol. 30, no. 3, pp. 305–319, 2016. [Online]. Available: <https://doi.org/10.1177/1094342015623623>
- [21] C. Engelmann and A. Geist, "A diskless checkpointing algorithm for super-scale architectures applied to the fast fourier transform," in *1st International Workshop on Challenges of Large Applications in Distributed Environments, CLADE@HPDC 2003, Seattle, WA, USA, June 21, 2003*. IEEE Computer Society, 2003, p. 47. [Online]. Available: <https://doi.org/10.1109/CLADE.2003.1209999>
- [22] L. Hübner, A. M. Kozlov, D. Hespe, P. Sanders, and A. Stamatakis, "Exploring parallel MPI fault tolerance mechanisms for phylogenetic inference with RaxML-NG," *Bioinformatics*, vol. 37, no. 22, pp. 4056–4063, 2021.
- [23] M. Vijay and R. Mittal, "Algorithm-based fault tolerance: a review," *Microprocess. Microsystems*, vol. 21, no. 3, pp. 151–161, 1997. [Online]. Available: [https://doi.org/10.1016/S0141-9331\(97\)00029-X](https://doi.org/10.1016/S0141-9331(97)00029-X)
- [24] G. Bosilca, R. Delmas, J. J. Dongarra, and J. Langou, "Algorithmic based fault tolerance applied to high performance computing," *CoRR*, vol. abs/0806.3121, 2008. [Online]. Available: <http://arxiv.org/abs/0806.3121>
- [25] B. Memishi, S. Ibrahim, M. S. Pérez, and G. Antoniu, "Fault tolerance in MapReduce: A survey," in *Resource Management for Big Data Platforms - Algorithms, Modelling, and High-Performance Computing Techniques*, ser. Computer Communications and Networks, F. Pop, J. Kolodziej, and B. D. Martino, Eds. Springer, 2016, pp. 205–240. [Online]. Available: https://doi.org/10.1007/978-3-319-44881-7_11
- [26] N. Kohl, J. Hötzer, F. Schornbaum, M. Bauer, C. Godenschwager, H. Köstler, B. Nestler, and U. Rüde, "A scalable and extensible checkpointing scheme for massively parallel simulations," *Int. J. High Perform. Comput. Appl.*, vol. 33, no. 4, 2019. [Online]. Available: <https://doi.org/10.1177/1094342018767736>
- [27] S. P. Gavaskar and C. D. V. Subbarao, "A survey of distributed fault tolerance strategies," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 11, Nov. 2013.
- [28] J. S. Plank, K. Li, and M. A. Puening, "Diskless checkpointing," *IEEE Trans. Parallel Distributed Syst.*, vol. 9, no. 10, pp. 972–986, 1998. [Online]. Available: <https://doi.org/10.1109/71.730527>

- [29] M. J. Gossman, B. Nicolae, J. C. Calhoun, F. Cappello, and M. C. Smith, "Towards aggregated asynchronous checkpointing," *CoRR*, vol. abs/2112.02289, 2021. [Online]. Available: <https://arxiv.org/abs/2112.02289>
- [30] T. Hérault, Y. Robert, A. Bouteiller, D. C. Arnold, K. B. Ferreira, G. Bosilca, and J. J. Dongarra, "Checkpointing strategies for shared high-performance computing platforms," *Int. J. Netw. Comput.*, vol. 9, no. 1, pp. 28–52, 2019. [Online]. Available: <http://www.ijnc.org/index.php/ijnc/article/view/195>
- [31] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *2010 IEEE 26th symposium on mass storage systems and technologies (MSST)*. Ieee, 2010, pp. 1–10.
- [32] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauly, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2012, San Jose, CA, USA, April 25-27, 2012*, S. D. Gribble and D. Katabi, Eds. USENIX Association, 2012, pp. 15–28. [Online]. Available: <https://www.usenix.org/conference/nsdi12/technical-sessions/presentation/zaharia>
- [33] A. Bouteiller, "ULFM 4.0.2u1 release notes," online, Nov. 2019. [Online]. Available: <https://fault-tolerance.org/2019/11/18/ulfm-4-0-2u1/>
- [34] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, jun 1960.
- [35] T. Chiueh and P. Deng, "Evaluation of checkpoint mechanisms for massively parallel machines," in *Digest of Papers: FTCS-26, The Twenty-Sixth Annual International Symposium on Fault-Tolerant Computing, Sendai, Japan, June 25-27, 1996*. IEEE Computer Society, 1996, pp. 370–379. [Online]. Available: <https://doi.org/10.1109/FTCS.1996.534622>
- [36] A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, and A. Stamatakis, "Raxml-ng: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference," *Bioinform.*, vol. 35, no. 21, pp. 4453–4455, 2019. [Online]. Available: <https://doi.org/10.1093/bioinformatics/btz305>
- [37] A. Kozlov, "Models, optimizations, and tools for large-scale phylogenetic inference, handling sequence uncertainty, and taxonomic validation." Ph.D. dissertation, Karlsruher Institut für Technologie (KIT), Jan. 2018.
- [38] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1. Oakland, CA, USA, 1967, pp. 281–297.
- [39] A. J. Aberer, K. Kobert, and A. Stamatakis, "ExaBayes: massively parallel Bayesian tree inference for the whole-genome era," *Molecular biology and evolution*, vol. 31, no. 10, pp. 2553–2556, 2014.
- [40] Cray, "Cray xk7 specifications," online, Jan. 2013. [Online]. Available: <https://web.archive.org/web/20130106091417/http://www.cray.com/Products/Computing/XK7/Specifications.aspx>
- [41] G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*. At the Clarendon Press, 1960.
- [42] P. Erdos and M. Kac, "The gaussian law of errors in the theory of additive number theoretic functions," *American Journal of Mathematics*, vol. 62, no. 1/4, p. 738, 1940.