Fault Tolerance for Remote Memory Access Programming Models

Maciej Besta
Dept. of Computer Science
ETH Zurich
Universitätstr. 6, 8092 Zurich, Switzerland
maciej.besta@inf.ethz.ch

Torsten Hoefler
Dept. of Computer Science
ETH Zurich
Universitätstr. 6, 8092 Zurich, Switzerland
htor@inf.ethz.ch

ABSTRACT

Remote Memory Access (RMA) is an emerging mechanism for programming high-performance computers and datacenters. However, little work exists on resilience schemes for RMA-based applications and systems. In this paper we analyze fault tolerance for RMA and show that it is fundamentally different from resilience mechanisms targeting the message passing (MP) model. We design a model for reasoning about fault tolerance for RMA, addressing both flat and hierarchical hardware. We use this model to construct several highly-scalable mechanisms that provide efficient low-overhead in-memory checkpointing, transparent logging of remote memory accesses, and a scheme for transparent recovery of failed processes. Our protocols take into account diminishing amounts of memory per core, one of major features of future exascale machines. The implementation of our fault-tolerance scheme entails negligible additional overheads. Our reliability model shows that in-memory checkpointing and logging provide high resilience. This study enables highly-scalable resilience mechanisms for RMA and fills a research gap between fault tolerance and emerging RMA programming models.

Categories and Subject Descriptors

C.4 [**Computer Systems Organization**]: Performance of systems—*Fault tolerance*

General Terms

Reliability, Performance, Algorithms

1. INTRODUCTION

Partitioned Global Address Space (PGAS), and the wider class of Remote Memory Access (RMA) programming models enable high-performance communications that often outperform Message Passing [35, 63]. RMA utilizes remote direct memory access (RDMA) hardware features to access memories at remote processes without involvement of the OS or the remote CPU.

RDMA is offered by most modern HPC networks (InfiniBand, Myrinet, Cray's Gemini and Aries, IBM's Blue Gene, and PERCS) and many Ethernet interconnects that use the RoCE or iWARP protocols. RMA languages and libraries include Unified Parallel C (UPC), Fortran 2008 (formerly known as CAF), MPI-3 One Sided, Cray's SHMEM interface, or Open Fabrics (OFED). Thus, we observe that RMA is quickly emerging to be the programming model of choice for cluster systems, HPC computers, and large datacenters.

Fault tolerance of such systems is important because hardware and software faults are ubiquitous [76]. Two popular resilience schemes used in today's computing environments are coordinated checkpointing (CC) and uncoordinated checkpointing augmented with message logging (UC) [32]. In CC applications regularly synchronize to save their state to memory, local disks, or parallel file system (PFS) [76]; this data is used to restart after a crash. In UC processes take checkpoints independently and use message logging to avoid rollbacks caused by the *domino effect* [71]. There has been considerable research on CC and UC for the message passing (MP) model [7, 32]. Still, no work addresses the exact design of these schemes for RMA-based systems.

In this work we develop a generic model for reasoning about resilience in RMA. Then, using this model, we show that CC and UC for RMA fundamentally differ from analogous schemes for MP. We also construct protocols that enable simple checkpointing and logging of remote memory accesses. We *only* use *in-memory* mechanisms to avoid costly I/O flushes and frequent disk and PFS failures [44, 76]. We then extend our model to cover two features of today's petascale and future exascale machines: (1) the growing complexity of hardware components and (2) decreasing amounts of memory per core. *With this, our study fills an important knowledge gap between fault-tolerance and emerging RMA programming in large-scale computing systems.*

In detail, we provide the following major contributions:

- We design a model for reasoning about the reliability of RMA systems running on flat and hierarchical hardware with limited memory per core. To our knowledge, this is the first work that addresses these issues.
- We construct schemes for in-memory checkpointing, logging, and recovering RMA-based applications.
- We unify these concepts in a topology-aware diskless protocol and we use real data and an analytic model to show that the protocol can endure concurrent hardware failures.
- We present the implementation of our protocol, analyze

	MPI-3 one sided operation	UPC operation	Fortran 2008 operation	Cat.
comm.	MPI_Put, MPI_Accumulate, MPI_Get_accumulate, MPI_Fetch_and_op, MPI_Compare_and_swap	upc_memput, upc_memcpy, upc_memset, assignment (=), all UPC collectives	assignment (=)	PUT
	MPI_Get, MPI_Compare_and_swap, MPI_Get_accumulate, MPI_Fetch_and_op	upc_memget, upc_memcpy, upc_memset, assignment (=), all UPC collectives	assignment (=)	GET
sync.	MPI_Win_lock, MPI_Win_lock_all	upc_lock	lock	LOCK
	MPI_Win_unlock, MPI_Win_unlock_all	upc_unlock	unlock	UNLOCK
	MPI_Win_fence	upc_barrier	sync_all, sync_team, sync_images	GSYNC
	MPI_Win_flush, MPI_Win_flush_all, MPI_Win_sync	upc_fence	sync_memory	FLUSH

Table 1: Categorization of MPI One Sided/UPC/Fortran 2008 operations in our model. Some atomic functions are considered as both PUTs and GETs. In UPC, the collectives, assignments and upc_memset/upc_memcpy behave similarly depending on the values of pointers to shared objects; the same applies to Fortran 2008. We omit MPI's post-start-complete-wait synchronization and request-based RMA operations for simplicity.

its performance, show it entails negligible overheads, and compare it to other schemes.

2. RMA PROGRAMMING

We now discuss concepts of RMA programming and present a formalization that covers generalizes existing RMA/PGAS models with strict or relaxed memory consistency (e.g., UPC or MPI-3 One Sided). In RMA, each process explicitly exposes an area of its local memory as shared. Memory can be shared in different ways (e.g., MPI windows, UPC shared arrays or Co-Arrays in Fortran 2008); details are outside the scope of this work. Once shared, memory can be accessed with various language-specific operations.

2.1 RMA Operations

We identify two fundamental types of RMA operations: *communication* actions (often called *accesses*; they transfer data between processes), and *synchronization* actions (synchronize processes and guarantee memory consistency). A process p that issues an RMA action targeted at q is called the *active source*, and q is called the *passive target*. In the following we assume p is active and q is passive (unless stated otherwise).

2.1.1 Communication Actions

We denote an action that transfers data from p to q and from q to p as $\text{PUT}(p \rightrightarrows q)$ and $\text{GET}(p \leftrightarrows q)$, respectively. We use double-arrows to emphasize the asymmetry of the two operations: the upper arrow indicates the direction of data flow and the lower arrow indicates the direction of control flow. The upper part of Table 1 categorizes communication operations in various RMA languages. Some actions (e.g., atomic compare and swap) transfer data in *both* directions and thus fall into the family of PUTs *and* GETs.

We also distinguish between PUTs that "blindly" replace a targeted memory region at q with a new value (e.g., UPC assignment), and PUTs that combine the data moved to q with the data that already resides at q (e.g., MPI_Accumulate). When necessary, we refer to the former type as the *replacing* PUT, and to the latter as the *combining* PUT.

2.1.2 Memory Synchronization Actions

We identify four major categories of memory synchronization actions: $\text{LOCK}(p \to q, str)$ (locks a structure str in q's memory to provide exclusive access), $\text{UNLOCK}(p \to q, str)$ (unlocks str in q's memory and enforces consistency of str), $\text{FLUSH}(p \to q, str)$ (enforces consistency of str in p's and q's memories), and $\text{GSYNC}(p \to \diamond, str)$ (enforces consistency str); $\diamond$ indicates that a call targets all processes. Arrows indicate the flow of control (synchronization). When we refer to the whole process memory (and not a single structure), we omit str (e.g., $\text{LOCK}(p \to q)$). The lower part of Table 1 categorizes synchronization calls in various RMA languages.

2.2 Epochs and Consistency Order

RMA's relaxed memory consistency enables non-blocking PUTs and GETs. Issued operations are completed by memory consistency actions (FLUSH, UNLOCK, GSYNC). The period between any two such actions issued by p and targeting the same process q is called an *epoch*. Every $\text{UNLOCK}(p \to q)$ or $\text{FLUSH}(p \to q)$ *closes* p's current epoch and *opens* a new one (i.e., increments p's epoch number denoted as $E(p \to q)$). p can be in several independent epochs related to each process that it communicates with. As GSYNC is a collective call, it increases epochs at every process.

An important concept related to epochs is the *consistency order* (denoted as $\xrightarrow{co}$). $\xrightarrow{co}$ orders the visibility of actions: $x \xrightarrow{co} y$ means that memory effects of action x are globally visible before action y. Actions issued in different epochs by process p targeting the same process q are always ordered with $\xrightarrow{co}$. Epochs and $\xrightarrow{co}$ are illustrated in Figure 1. $x \parallel_{co} y$ means that actions x and y are *not* ordered with $\xrightarrow{co}$.

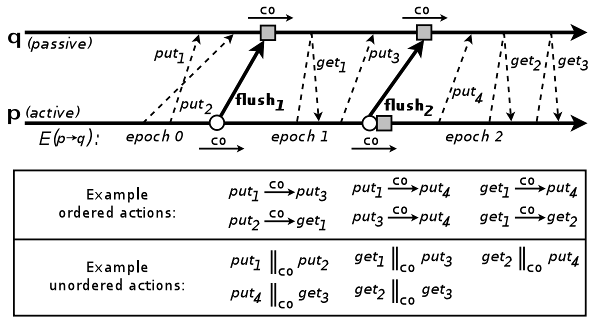


Figure 1: Epochs and consistency order $\xrightarrow{co}$ (§ 2.2). White circles symbolize synchronization calls (in this case FLUSH). Grey squares show when a calls' results become globally visible in q's or p's memory.

2.3 Program, Synchronization, and Happened Before Orders

In addition to $\xrightarrow{co}$ we require three more orders to specify an RMA execution [41]: The *program order* ($\xrightarrow{po}$) specifies the order of actions of a single thread, similarly to the program order in Java [58] ($x \xrightarrow{po} y$ means that x is called before y by some thread). The *synchronization order* ($\xrightarrow{so}$)

orders LOCK and UNLOCK and other synchronizing operations. *Happened-before* (HB, $\xrightarrow{hb}$), a relation well-known in message passing [51], is the transitive closure of the union of $\xrightarrow{po}$ and $\xrightarrow{so}$. We abbreviate a *consistent happen-before* as $\xrightarrow{cohb}$: $a \xrightarrow{cohb} \equiv a \xrightarrow{co} b \wedge a \xrightarrow{hb} b$. To state that actions are *parallel* in some order, we use the symbols $\|_{po}$, $\|_{so}$, $\|_{hb}$. We illustrate the orders in Figure 2; more details can be found in [41].

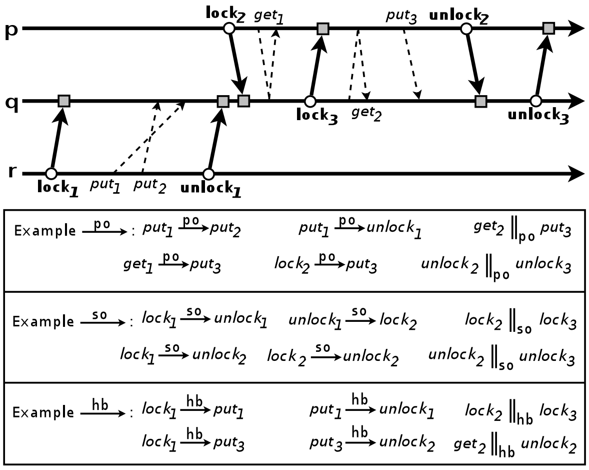


Figure 2: Example RMA orderings $\xrightarrow{po}$, $\xrightarrow{so}$, $\xrightarrow{hb}$ (§ 2.3).

2.4 Formal Model

We now combine the various RMA concepts and fault tolerance into a single formal model. We assume fail-stop faults (processes can disappear nondeterministically but behave correctly while being part of the program). The data communication may happen out of order as specified for most RMA models. That communication channels between non-failed processes are asynchronous, reliable, and error-free. The user code can only communicate and synchronize using RMA functions specified in Section 2.1. Finally, checkpoints and logs are stored in *volatile* memories.

We define a communication action a as a tuple

$$a = \langle type, src, trg, combine, EC, GC, SC, GNC, data\rangle \quad (1)$$

where *type* is either a put or a get, *src* and *trg* specify the source and the target, and *data* is the data carried by a. *Combine* determines if a is a replacing PUT ($combine = false$) or a combining PUT ($combine = true$). EC (*Epoch Counter*) is the epoch number in which a was issued. GC, SC, and GNC are counters required for correct recovery; we discuss them in more detail in Section 4. We combine the notation from Section 2.1 with this definition and write $\text{PUT}(p \rightrightarrows q).EC$ to refer to the epoch in which the put happens. We also define a *determinant* of a (denoted as $\#a$, cf. [7]) to be tuple a without *data*:

$$\#a = \langle type, src, trg, combine, EC, GC, SC, GNC\rangle. \quad (2)$$

Similarly, a synchronization action b is defined as

$$b = \langle type, src, trg, EC, GC, SC, GNC, str\rangle. \quad (3)$$

Finally, a trace of an RMA program running on a distributed system can be written as the tuple

$$\mathcal{D} = \langle \mathcal{P}, \mathcal{E}, \mathcal{S}, \xrightarrow{po}, \xrightarrow{so}, \xrightarrow{hb}, \xrightarrow{co}\rangle, \quad (4)$$

where

$\mathcal{P}$ is the set of all $\mathcal{P}$rocesses in $\mathcal{D}$ ($|\mathcal{P}| = N$),

$\mathcal{E} = \mathcal{A} \cup \mathcal{I}$ is the set of all $\mathcal{E}$vents:

$\mathcal{A}$ is the set of RMA $\mathcal{A}$ctions,

$\mathcal{I}$ is the set of three types of $\mathcal{I}$nternal actions at process p: (1) $\text{READ}(x, p)$ loads local variable x, (2) $\text{WRITE}(x := val, p)$ assignes val to x, and (3) the set of all Checkpoints: C_p^i is the ith checkpoint action. Internal events are partially ordered with actions using $\xrightarrow{po}$, $\xrightarrow{co}$, and $\xrightarrow{hb}$.

$\mathcal{S}$ is the set of all data $\mathcal{S}$tructures used by the program.

3. FAULT-TOLERANCE FOR RMA

We now present schemes that make RMA codes fault tolerant. We start with with the simpler CC and then present RMA protocols for UC.

3.1 Coordinated Checkpointing (CC)

In many CC schemes, the user explicitly calls a function to take a checkpoint. Such protocols may leverage RMA's features (e.g., direct memory access) to improve the performance. However, these schemes have several drawbacks: they complicate the code because they can only be called when the network is quiet [39] and they do not always fit the optimality criteria such as Daly's checkpointing interval [28]. In this section, we first identify how CC in RMA differs from CC in MP and then describe a scheme for RMA codes that performs CC *transparently* to the application. We model a coordinated checkpoint as a set $C = \{C_{p_1}^{i_1}, C_{p_2}^{i_2}, ..., C_{p_N}^{i_N}\} \subseteq \mathcal{I}, p_m \neq p_n$ for any m, n.

3.1.1 RMA vs. MP: Coordinated Checkpointing

In MP, every C has to satisfy a *consistency condition* [39]: $\forall C_p^i, C_q^j \in C: C_p^i \|_{hb} C_q^i$. This condition ensures that C does not reflect a system state in which one process received a message that was *not* sent by any other process. We adopt this condition and extend it to cover all RMA semantics:

DEFINITION 1. *C is RMA-consistent iff* $\forall C_p^i, C_q^j \in C: C_p^i \|_{cohb} C_q^i$.

We extend $\|_{hb}$ to $\|_{cohb}$ to guarantee that the system state saved in C does not contain a process affected by a memory access that was *not* issued by any other process. In RMA, unlike in MP, this condition can be easily satisfied because each process can drain the network with a local FLUSH (enforcing consistency at any point is legal [41])

3.1.2 Taking a Coordinated Checkpoint

We now propose two diskless schemes that obey the RMA-consistency condition and target MPI-3 RMA codes. The first ("Gsync") scheme can be used in programs that *only* synchronize with GSYNCs. The other ("Locks") scheme targets codes that *only* synchronize with LOCKs and UNLOCKs. Note that in correct MPI-3 RMA programs GSYNCs and

LOCKs/UNLOCKs cannot be mixed [60]. All our schemes assume that a GSYNC may also introduce an additional $\xrightarrow{hb}$ order, which is true in some implementations [60].

The "Gsync" Scheme Every process may take a coordinated checkpoint right after the user calls a GSYNC and before any further RMA calls by: (1) optionally enforcing the global $\xrightarrow{hb}$ order with an operation such as MPI_Barrier (denoted as BAR), and taking the checkpoint. Depending on the application needs, not every GSYNC has to be followed by a checkpoint. We use Daly's formula [28] to compute the best interval between such checkpoints and inset checkpoints after the right GSYNC calls.

THEOREM 3.1. *The Gsync scheme satisfies the RMA-consistency condition and does not deadlock.*

PROOF. We assume correct MPI-3 RMA programs represented by their trace $\mathcal{D}$ [41, 60]. For all $p, q \in \mathcal{P}$, each $\text{GSYNC}(p \to \diamond)$ has a matching $\text{GSYNC}(q \to \diamond)$ such that $[\text{GSYNC}(p \to \diamond) \parallel_{hb} \text{GSYNC}(q \to \diamond)]$. Thus, if every process calls BAR right after GSYNC then BAR matching is guaranteed and the program cannot deadlock. In addition, the GSYNC calls introduce a global consistency order $\xrightarrow{co}$ such that the checkpoint is coordinated and consistent. □

The "Locks" Scheme Every process p maintains a local *Lock Counter* LC_p that starts with zero and is incremented after each LOCK and decremented after each UNLOCK. When $LC_p = 0$, process p can perform a checkpoint in three phases: (1) enforce consistency with a $\text{FLUSH}(p \to \diamond)$, (2) call a BAR to provides the global $\xrightarrow{hb}$ order, and (3) take the coordinated checkpoint C_p^i. The last phase, the actual checkpoint stage, is performed collectively thus all processes can take the checkpoint C^i in coordination.

THEOREM 3.2. *The Locks scheme satisfies the RMA-consistency condition and does not deadlock.*

PROOF. The call to $\text{FLUSH}(p \to \diamond)$ in phase 1 guarantees global consistency at each process. The BAR in phase 2 guarantees that all processes are globally consistent before the checkpoint inphase (3).

It remains to proof deadlock-freedom: We assume correct MPI-3 RMA programs [41, 60]. A $\text{LOCK}(p \to q)$ can only block waiting for an active lock $\text{LOCK}(z \to q)$ and no BAR can be started at z while the lock is held. In addition, for every $\text{LOCK}(z \to q)$, there is a matching $\text{UNLOCK}(z \to q)$ in the execution such that $\text{LOCK}(z \to q) \xrightarrow{po} \text{UNLOCK}(z \to q)$ (for any $z, p, q \in \mathcal{P}$). Thus, all locks must be released eventually, i.e., $\exists a \in \mathcal{E} : a \xrightarrow{po} \text{WRITE}(LC_p := 0, p)$ for any $p \in \mathcal{P}$. □

The above schemes show that the transparent CC can be achieved much simpler in RMA than in MP. In MP, such protocols usually have to analyze inter-process dependencies due to sent/received messages, and add protocol-specific data to messages [21, 32], which reduces the bandwidth. In RMA this is not necessary.

3.2 Uncoordinated Checkpointing (UC)

Uncoordinated checkpointing augmented with message logging reduces energy consumption and synchronization costs because a single process crash does not force all other processes to revert to the previous checkpoint and recompute [32, 71]. Instead, a failed process fetches its last checkpoint and replays messages logged beyond this checkpoint. However, UC schemes are usually more complex than CC [32]. We now analyze how UC in RMA differs from UC in MP, followed by a discussion of our UC protocols.

3.2.1 RMA vs. MP: Uncoordinated Checkpointing

The first and obvious difference is that we now log not *messages* but *accesses*. Other differences are as follows:

Storing Access Logs In MP, processes exchange messages that *always* flow *from* the sender (process p) *to* the receiver (process q). Messages can be recorded at the sender's side [32,71]. During a recovery, the restored process interacts with other processes to get and reply the logged messages (see Figure 3 (part (1)).

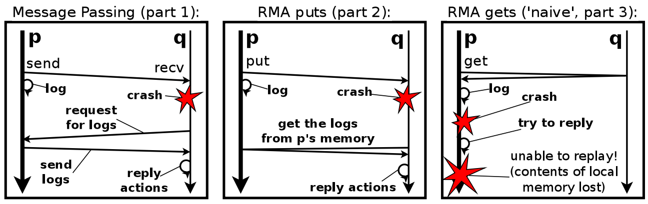


Figure 3: The logging of messages vs. RMA puts and gets (§ 3.1.1).

In RMA, a $\text{PUT}(p \rightrightarrows q)$ changes the state of q, but a $\text{GET}(p \leftrightarrows q)$ modifies the state of p. Thus, $\text{PUT}(p \rightrightarrows q)$ can be logged in p's memory, but $\text{GET}(p \leftrightarrows q)$ cannot because a failure of p would prevent a successful recovery (see Figure 3, part 2 and 3).

Transparency of Schemes In MP, both p and q actively participate in communication. In RMA, q is oblivious to accesses to its memory and thus any recovery or logging performed by p can be *transparent* to (i.e., does not obstruct) q (which is usually *not* the case in MP, cf. [71]).

No Piggybacking Adding some protocol-specific data to messages (e.g., *piggybacking*) is a popular concept in MP [32]. Still, it cannot be used in RMA because PUTs and GETs are accesses, not messages. Yet, issuing additional accesses is cheap in RMA.

Access Determinism Recent works in MP (e.g., [36]) explore *send determinism*: the output of an application run is oblivious to the order of received messages. In our work we identify a similar concept in RMA that we call *access determinism*. For example, in race-free MPI-3 programs the application output does not depend on the order in which two accesses a and b committed to memory if $a \parallel_{co} b$.

Orphan Processes In some MP schemes (called *optimistic*), senders postpone logging messages for performance reasons [32]. Assume q received message m from p and then sent a message m' to r. If q crashes and m is not logged by p at that time, then q may follow a run in that it *does not* send m'. Thus, r becomes an *orphan*: its state depends on a message m' that was *not* sent [32] (see Figure 4, part 1).

In RMA a process may also become an orphan. Consider Figure 4 (part 2). First, p modifies variable x at q. Then, q reads x and conditionally issues a $\text{PUT}(q \rightrightarrows r)$. If q crashes and p postponed logging $\text{PUT}(p \rightrightarrows q)$, then q (while recovering) may follow a run in which it does not issue $\text{PUT}(q \rightrightarrows r)$; thus r becomes an orphan.

3.2.2 Taking an Uncoordinated Checkpoint

We denote the ith uncoordinated checkpoint taken by process p as C_p^i. Taking C_p^i is simple and entails: (1) locking

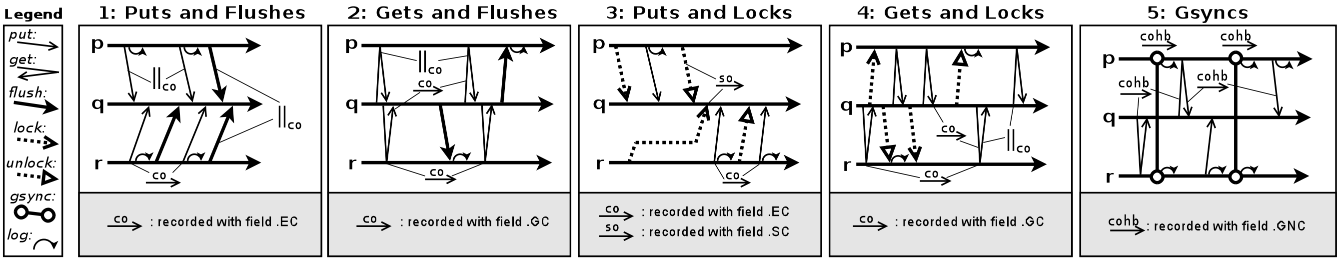


Figure 5: Logging orders $\xrightarrow{so}$, $\xrightarrow{co}$, and $\xrightarrow{hb}$ (§ 4.1). In each figure we illustrate example orderings.

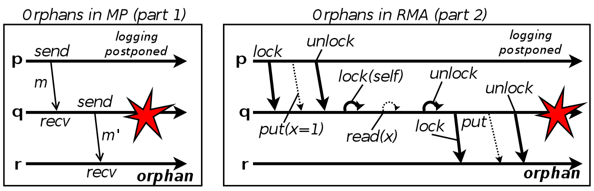


Figure 4: Illustration of orphans in MP and RMA (§ 3.1.1).

Structure	Description
$LP_p[q] \in \mathcal{S}$	Logs of PUTs issued by p and targeted at q.
$LG_q[p] \in \mathcal{S}$	Logs of GETs targeted at q and issued by p.
$LP_p \in \mathcal{S}$	Logs of PUTs issued and stored by p and targeted at any other process; $LP_p \equiv \bigcup_{r \in \mathcal{P} \wedge r \neq p} LP_p[r]$.
$LG_q \in \mathcal{S}$	Logs of *gets* targeted and stored at q, issued by any other process; $LG_q \equiv \bigcup_{r \in \mathcal{P} \wedge r \neq q} LG_p[r]$.
$Q_p \in \mathcal{S}$	A helper container stored at p, used to temporarily log #GETs issued by p.
$N_q[p] \in \mathcal{S}$	A structure (stored at q) that determines whether or not p issued a non-blocking GET$(p \rightrightarrows q)$ ($N_q[p] = true$ or $false$, respectively)

Table 2: Data structures used in RMA logging (§ 3.2.3). $LP_p[q]$ and LP_p are stored at p. $LG_q[p]$ and LG_q are stored at q.

local application data, (2) sending the copy of the data to some remote volatile storage, and (3) unlocking the application data (we defer the discussion on the implementation details until Section 6). After p takes C_p^i, any process q can delete the logs of every PUT$(q \rightrightarrows p)$ (from $LP_q[p]$) and GET$(p \leftrightarrows q)$ (from $LG_q[p]$) that committed in p's memory before C_p^i (i.e., PUT$(q \rightrightarrows p) \xrightarrow{co} C_p^i$ and GET$(p \leftrightarrows q) \xrightarrow{co} C_p^i$).

We enforce that every C_p^i is taken *immediately after* closing/opening an epoch and *before* issuing any new communication operations (we call this the *epoch condition*). This condition is required because, if p issues a GET$(p \leftrightarrows q)$, the application data is guaranteed to be consistent only after closing the epoch.

3.2.3 Transparent Logging of RMA Accesses

We now describe the logging of PUTs and GETs; all necessary data structures are shown in Table 2.

Logging Puts To log a PUT$(p \rightrightarrows q)$, p first calls LOCK$(p \to p, LP_p)$. Self-locking is necessary because there may be other processes being recovered that may try to read LP_p. Then, the PUT is logged ($LP_p[q] := LP_p[q] \cup \{PUT(p \rightrightarrows q)\}$; ":=" denotes the assignment of a new value to a variable or a structure). Finally, p unlocks LP_p. Atomicity between logging and putting is not required because, in the weak consistency memory model, the source memory of the put operation may not be modified until the current epoch ends. If the program modifies it nevertheless, RMA implementations are allowed to return any value, thus the logged value is irrelevant. We log PUT$(p \rightrightarrows q)$ before closing the epoch PUT$(p \rightrightarrows q).EC$. If the PUT is blocking then we log it before issuing, analogously to the *pessimistic* message logging [32].

Logging Gets We log a GET$(p \leftrightarrows q)$ in two phases to retain its asynchronous behavior (see Algorithm 1). First, we record the determinant #GET$(p \leftrightarrows q)$ in Q_p (lines 2-3). We cannot access GET$(p \leftrightarrows q).data$ as the local memory will only be valid after the epoch ends. We avoid issuing an additional blocking FLUSH$(p \to q)$, instead we rely on the user's call to end the epoch. Second, when the user ends the epoch, we lock the remote log LG_q, record GET$(p \leftrightarrows q)$, and unlock LG_q (lines 4-7).

Note that if p fails between issuing GET$(p \leftrightarrows q)$ and closing the epoch, it will not be able to replay it consistently. To address this problem, p sets $N_q[p]$ at process q to *true* right before issuing the first GET$(p \rightrightarrows q)$ (line 1), and to *false* after closing the epoch GET$(p \rightrightarrows q).EC$ (line 8). During the recovery, if p notices that any $N_q[p] = true$, it falls back to another resilience mechanism (i.e., the last coordinated checkpoint). If the GET is blocking then we set $N_q[p] = false$ after returning from the call.

Algorithm 1: Logging *gets* (§ 3.2.3)

```
  Input: get := GET(p ⇆ q)
  /* Phase 1: starts right before issuing the get      */
1 N_q[p] := true
  /* Now we issue the get and log the #get             */
2 issue GET(p ⇆ q)
3 Q_p ← Q_p ∪ #get
  /* Phase 2: begins after ending the epoch get.EC     */
4 LOCK(p → q, LG_q)
5 LG_q[p] := LG_q[p] ∪ get
6 Q_p := Q_p \ #get
7 UNLOCK(p → q, LG_q)
8 N_q[p] := false
```

4. CAUSAL RECOVERY FOR UC

We now show how to causally recover a failed process (*causally* means preserving $\xrightarrow{co}$, $\xrightarrow{so}$, and $\xrightarrow{hb}$). This section describes technical details on how to guarantee all orders to ensure a correct access replay. If the reader is not interested in all details, she may proceed to Section 5 without disrupting the flow. A causal process recovery has three phases: (1) fetching uncoordinated checkpoint data, (2) replaying accesses from remote logs, and (3) in case of a problem during the replay, falling back to the last coordinated checkpoint. We first show how we log the respective orderings between accesses (Section 4.1) and how we prevent replaying some accesses twice (Section 4.2). We finish with our recovery scheme (Section 4.3) and a discussion (Section 4.4).

4.1 Logging Order Information

We now show how to record $\xrightarrow{so}$, $\xrightarrow{hb}$, and $\xrightarrow{co}$. For clarity, but without loss of generality, we separately present several scenarios that exhaust possible communication/synchronization patterns in our model. We consider three processes (p, q, r) and we analyze what data is required to replay q. We show each pattern in Figure 5.

A. Puts and Flushes First, p and r issue PUTs and FLUSHes at q. At both p and r, PUTs separated by FLUSHes are ordered with $\xrightarrow{co}$. This order is preserved by recording epoch counters ($.EC$) with every logged $\text{PUT}(p \rightrightarrows q)$. Note that, however, RMA semantics *do not* order calls issued by p and r: $[\text{PUT}(p \rightrightarrows q) \parallel_{co} \text{PUT}(r \rightrightarrows q)]$ without additional process synchronization. Here, we assume *access determinism*: the recovery output does not depend on the order in which such PUTs committed in q's memory.

B. Gets and Flushes Next, q issues GETs and FLUSHes targeted at p and r. Again, $\xrightarrow{co}$ has to be logged. However, this time GETs targeted at *different* processes *are* ordered (because they are issued by the same process). To log this ordering, q maintains a local *Get Counter* GC_q that is incremented each time q issues a $\text{FLUSH}(q \to \diamond)$ to any other process. The value of this counter is logged with each GET using the field $.GC$ (cf. Section 2.4).

C. Puts and Locks In this scenario p and r issue PUTs at q and synchronize their accesses with LOCKs and UNLOCKs. This pattern requires logging the $\xrightarrow{so}$ order. We achieve this with a *Synchronization Counter* SC_q stored at q. After issuing a $\text{LOCK}(p \to q)$, p (the same refers to r) fetches the value of SC_q, increments it, updates remote SC_q, and records it with every PUT using the field $.SC$ (cf. Section 2.4). In addition, this scenario requires recording $\xrightarrow{co}$ that we solve with $.EC$, analogously as in the "Puts and Flushes" pattern.

D. Gets and Locks Next, q issues GETs and uses LOCKs targeted at p and r. This pattern is solved analogously to the "Gets and Flushes" pattern.

E. Gsyncs The final pattern are GSYNCs (that may again introduce $\xrightarrow{hb}$) combined with any communication action. Upon a GSYNC, each process q increments its *GsyNc Counter* GNC_q that is logged in an actions' $.GNC$ field (cf. Section 2.4).

4.2 Preventing Replaying Accesses Twice

Assume that process p issues a $\text{PUT}(p \rightrightarrows q)$ (immediately logged by p in $LP_p[q]$) such that $\text{PUT}(p \rightrightarrows q) \xrightarrow{co} C_q^j$. It means that the state of q recorded in checkpoint C_q^j is affected by $\text{PUT}(p \rightrightarrows q)$. Now assume that q fails and begins to replay the logs. If p did not delete the log of $\text{PUT}(p \rightrightarrows q)$ from $LP_p[q]$ (it was allowed to do it after q took C_q^j), then q replays $\text{PUT}(p \rightrightarrows q)$ and this PUT affects its memory *for the second time*. This is not a problem if $\text{PUT}(p \rightrightarrows q).combine = false$, because such a PUT always overwrites the memory region with the same value. However, if $\text{PUT}(p \rightrightarrows q).combine = true$, then q ends up in an inconsistent state (e.g., if this PUT increments a memory cell, this cell will be incremented twice).

Algorithm 2: The causal recovery scheme for codes that synchronize with GSYNCs (§ 4.3, § 4.4).

```
1  Function recovery()
2      fetch_checkpoint_data()
3      put_logs := {}; get_logs := {}
4      forall the q ∈ P : q ≠ p_new do
5          LOCK(p_new → q)
6          if N_q[p_f] = 1 ∨ M_q[p_f] = true then
               /* Stop the recovery and fall back to the
                  last coordinated checkpoint                */
7          end
8          put_logs := put_logs ∪LP_q[p_f]
9          get_logs := get_logs ∪LG_q[p_f]
10         UNLOCK(p_new → q)
11     end
12     while |put_logs| > 0 ∨ |get_logs| > 0 do
13         gnc_logs := logsWithMinCnt(GNC, put_logs ∪ get_logs)
14         gnc_logs_temp := gnc_logs
15         while |gnc_logs| > 0 do
16             gnc_put_logs := gnc_logs ∩ put_logs
17             gnc_get_logs := gnc_logs ∩ get_logs
18             ec_logs := logsWithMinCnt(EC, gnc_put_logs)
19             gc_logs := logsWithMinCnt(GC, gnc_get_logs)
20             replayEachAction(ec_logs)
21             replayEachAction(gc_logs)
22             gnc_logs := gnc_logs \ (ec_logs ∪ gc_logs)
23         end
24         put_logs := put_logs \ gnc_logs_temp
25         get_logs := get_logs \ gnc_logs_temp
26     end
27     return
28 Function logsWithMinCnt(Counter, Logs)
       /* Return a set with logs from Logs that have the
          smallest value of the specified counter (one
          of:GNC, EC, GC, SC).                               */
29 Function replayEachAction(Logs)
       /* Reply each log from set Logs in any order.         */
30 Function fetchCheckpointData()
       /* Fetch the last checkpoint and load into the
          memory.                                            */
```

To solve this problem, every process p maintains a local structure $M_p[q] \in \mathcal{S}$. When p issues and logs a $\text{PUT}(p \rightrightarrows q)$ such that $\text{PUT}(p \rightrightarrows q).combine = true$, it sets $M_p[q] := true$. When p deletes $\text{PUT}(p \rightrightarrows q)$ from its logs, it sets $M_p[q] := false$. If q fails, starts to recover, and sees that any $M_p[q] = true$, it stops the recovery and falls back to the coordinated checkpoint. This scheme is valid if access determinism is assumed. Otherwise we set $M_p[q] := true$ regardless of the value of $\text{PUT}(p \rightrightarrows q).combine$; we use the same approach if q can issue WRITEs to the memory regions accessed with remote PUTs parallel in $\parallel_{co}$ to these WRITEs.

4.3 Recovering a Failed Process

We now describe a protocol for codes that synchronize with GSYNCs. Let us denote the failed process as p_f. We assume an underlying batch system that provides a new process p_{new} in the place of p_f, and that other processes resume their communication with p_{new} after it fully recovers. We illustrate the scheme in Algorithm 2. First, p_{new} fetches the checkpointed data. Second, p_{new} gets the logs of PUTs (put_logs) and GETs (get_logs) related to p_f (lines 3-11). It also checks if any $N_q[p_f] = true$ (see § 3.2.3) or $M_q[p_f] = true$ (see § 4.2), if yes it instructs all processes to roll back to the last coordinated checkpoint. The protocol uses LOCKs (lines 5,10) to prevent data races due to, e.g., concurrent recoveries and log cleanups by q.

The main part (lines 12-27) replays accesses causally. The recovery ends when there are no logs left (line 12; $|logs|$ is the size of the set "logs"). We first get the logs with the smallest $.GNC$ (line 13) to maintain $\xrightarrow{cohb}$ introduced by GSYNCs (see § 4.1 E). Then, within this step, we find the logs with minimum $.EC$ and $.GC$ to preserve $\xrightarrow{co}$ in issued PUTs and GETs, respectively (lines 18-19, see § 4.1 A, B). We replay them in lines 20-21.

THEOREM 4.1. *The recovery scheme presented in Algorithm 2 replays each fetched action exactly once.*

PROOF. Consider the gnc_logs set obtained in line 13. The definition of function logsWithMinCnt ensures that, after executing the action in line 13 and before entering the loop that starts in line 15, every $a \in$ gnc_logs has identical $a.GNC$. Then, the condition in line 15 together with the actions in line 22 and the definition of logsWithMinCnt ensure that gnc_logs is empty when the loop in lines 15-23 exits (all logs in gnc_logs are replayed). This result, together with the actions in lines 24-25, guarantee that each action a obtained in the lines 4-11 is extracted from log_puts $\cup$ log_gets in line 13 and replayed exactly once. ☐

THEOREM 4.2. *The recovery scheme presented in Algorithm 2 preserves the $\xrightarrow{cohb}$ order introduced by* GSYNCs *(referred to as the* gsync order*).*

PROOF. Let us denote the action that replays communication action a at process p_{new} as $\mathcal{R}(a, p_{new}) \in \mathcal{I}$ (as $\mathcal{R}(a, p_{new})$ affects only the memory of the calling process p_{new}, it is an internal action). Assume by contradiction that the gsync order is not preserved while recovering. Thus, $\exists a_1, a_2 \in$ log_puts $\cup$ log_gets : $(a_1.GNC > a_2.GNC) \wedge (\mathcal{R}(a_1, p_{new}) \xrightarrow{po} \mathcal{R}(a_2, p_{new}))$. It means that the action of including a_1 into gnc_logs (line 13) took place before the analogous action for a_2 (in the $\xrightarrow{po}$ order). But this contradicts the definition of function logsWithMinCnt(GNC, set) that returns all the actions from set that have the minimum value of the GNC counter. ☐

We now present a recovery scheme for codes that synchronize with LOCKs and communicate with PUTs. The first part of the scheme is identical to the one that targets GSYNCs; the difference is that we do not have to check the values of $N_q[p_f]$.

In the main part (lines 11-20) actions are replayed causally. We first get the logs with the smallest $.SC$ (line 12) to maintain $\xrightarrow{so}$ introduced by LOCKs (see § 4.1 C). Then, within this step, we find the logs with minimum $.EC$ to preserve the $\xrightarrow{co}$ order (line 4, see § 4.1 A). The PUTs are replayed in line 15.

4.4 Discussion

Our recovery schemes present a trade-off between memory efficiency and time to recover. Process p_{new} fetches all related logs and only then begins to replay accesses. Thus, we assume that its memory has capacity to contain put_logs and get_logs; a reasonable assumption if the user program has regular communication patterns (true for most of today's RMA applications [35]). A more memory-efficient scheme fetches logs while recovering. This incurs performance issues as p_{new} has to access remote logs multiple times.

Algorithm 3: The causal recovery scheme for codes that synchronize with LOCKs and communicate with PUTs (§ 4.3, § 4.4).

```
1  Function recovery()
2      fetch_checkpoint_data()
3      put_logs := {}
4      forall the q ∈ P : q ≠ p_new do
5          LOCK(p_new → q)
6          if M_q[p_f] = true then
               /* Stop the recovery and fall back to the
                  last coordinated checkpoint                 */
7          end
8          put_logs := put_logs ∪LP_q[p_f]
9          UNLOCK(p_new → q)
10     end
11     while |put_logs| > 0 do
12         sc_put_logs := logsWithMinCnt(SC, put_logs)
13         while |sc_put_logs| > 0 do
14             ec_logs := logsWithMinCnt(EC, sc_put_logs)
15             replayEachAction(ec_logs)
16             sc_put_logs := sc_put_logs \ ec_logs
17         end
18         put_logs := put_logs \ sc_put_logs
19     end
20     return
21 Function logsWithMinCnt(Counter, Logs)
       /* Return a set with logs from Logs that have the
          smallest value of the specified counter (one
          of:GNC, EC, GC, SC).                                */
22 Function replayEachAction(Logs)
       /* Reply each log from set Logs in any order.      */
23 Function fetchCheckpointData()
       /* Fetch the last checkpoint and load into the
          memory.                                             */
```

5. EXTENDING THE MODEL FOR MORE RESILIENCE

Our model and in-memory resilience schemes are oblivious to the underlying hardware. However, virtually all of today's systems have a hierarchical hardware layout (e.g., cores reside on a single chip, chips reside in a single node, nodes form a rack, and racks form a cabinet). Multiple elements may be affected by a single failure at a higher level, jeopardizing the safety of our protocols. We now extend our model to cover arbitrary hierarchies and propose *topology-aware* mechanisms to make our schemes handle concurrent hardware failures. Specifically, we propose three following extensions:

The Hierarchy of Failure Domains A *failure domain* (FD) is an element of a hardware hierarchy that can fail (e.g., a node or a cabinet). FDs constitute an FD hierarchy (FDH) with h levels. An example FDH is shown in Figure 6, $h = 4$. We skip the level of single cores because in practice the smallest FD is a node (e.g., in the TSUBAME2.0 system failure history, there are no core failures [3]). Then, we define $\mathcal{H} = \bigcup_{1 \le j \le h} \left(\bigcup_{1 \le i \le H_j} H_{i,j} \right)$ to be the set of all the FD elements in an FDH. $H_{i,j}$ and H_j are element i of hierarchy level j and the number of such elements at level j, respectively. For example, in Figure 6 $H_{3,2}$ is the third blade (level 2) and $H_2 = 96$.

Groups of Processes To improve resilience, we split the process set $\mathcal{P}$ into g equally-sized groups G_i and add m *checksum* processes to each group to store checksums of checkpoints taken in each group (using, e.g., the Reed-Solomon [70] coding scheme). Thus, every group can resist m concurrent process crashes. The group size is $|G| = \frac{|\mathcal{P}|}{g} + m$.

New System Definition We now extend the definition

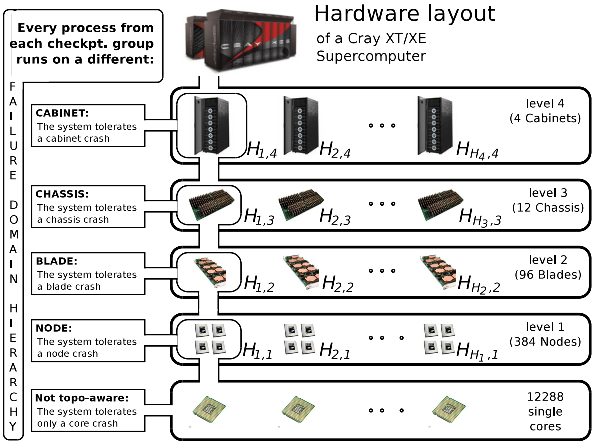


Figure 6: An example hardware layout (Cray XT/XE) and the corresponding FDH (§ 5). In this example, $h = 4$.

of a distributed system $\mathcal{D}$ to cover the additional concepts

$$\langle \mathcal{P}, \mathcal{E}, \mathcal{S}, \mathcal{H}, \mathcal{G}, \xrightarrow{po}, \xrightarrow{so}, \xrightarrow{hb}, \xrightarrow{co}, \mathscr{M} \rangle. \tag{5}$$

$\mathcal{G} = \{G_1, ..., G_g\}$ is a set of $\mathcal{G}$roups of processes and $\mathscr{M} : \mathcal{P} \times \mathbb{N} \to \mathcal{H}$ is a function that $\mathscr{M}$aps process p to an FD at hierarchy level k where p runs: $\mathscr{M}(p,k) = H_{j,k}$. $\mathscr{M}$ defines how processes are distributed over FDH. For example, if p runs on blade $H_{1,2}$ from Figure 6, then $\mathscr{M}(p,2) = H_{1,2}$.

5.1 Handling Multiple Hardware Failures

More than m process crashes in any group G_i result in a *catastrophic failure* (CF; we use the name from [14]) that incurs restarting the whole computation. Depending on how $\mathscr{M}$ distributes processes, such a CF may be caused by several (or even one) crashed FDs. To minimize the risk of CFs, $\mathscr{M}$ has to be *topology-aware* (t-aware): for a given level n (called a *t-awareness level*), no more than m processes from the same group can run on the same $H_{i,k}$ at any level $k, k \le n$:

$$\forall p_1, p_2, ..., p_m \in \mathcal{P} \quad \forall G \in \mathcal{G} \quad \forall 1 \le k \le n :$$
$$(p_1 \in G \wedge ... \wedge p_m \in G) \Rightarrow (\mathscr{M}(p_1,k) \ne ... \ne \mathscr{M}(p_m,k)) \tag{6}$$

Figure 7 shows an example t-aware process distribution.

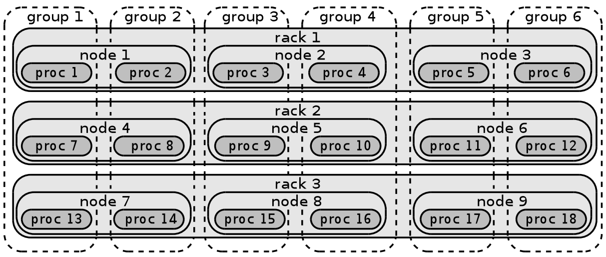


Figure 7: T-aware distribution at the node *and* rack level (§ 5.1).

5.2 Calculating Probability of a CF

We now calculate the probability of a catastrophic failure (P_{cf}) in our model. We later (§ 7.1) use P_{cf} to show that our protocols are resilient on a concrete machine (the TSUMABE2.0 supercomputer [3]). If a reader is not interested in the derivation details, she may proceed to Section 6 where we present the results. We set $m = 1$ and thus use the XOR erasure code, similar to an additional disk in a RAID5 [22]. We assume that failures at different hierarchy levels are independent and that any number x_j of elements from any hierarchy level j ($1 \le x_j \le H_j$, $1 \le j \le h$) can fail. Thus,

$$P_{cf} = \sum_{j=1}^{h} \sum_{x_j=1}^{H_j} P(x_j \cap x_{j,cf}) = \sum_{j=1}^{h} \sum_{x_j=1}^{H_j} P_j(x_j) P_j(x_{j,cf}|x_j). \tag{7}$$

$P(x_j \cap x_{j,cf})$ is the probability that x_j elements of the j hierarchy level will fail *and* result in a catastrophic failure. $P_j(x_j)$ is the probability of the failure of x_j elements from level j of the hierarchy. $P_j(x_{j,cf}|x_j)$ is the probability that x_j given concurrent failures at hierarchy level j are catastrophic to the system. It is difficult to analytically derive $P_j(x_j)$ as it is specific for every machine. For our example study (see Section 7.1) we use the failure rates from the TSUBAME2 failure history [3].

In contrast, $P_j(x_{j,cf}|x_j)$ can be calculated using combinatorial theory. Assume that $\mathscr{M}$ distributes processes in a t-aware way at levels 1 to n of the FDH ($1 \le n \le h$). First, we derive $P_j(x_{j,cf}|x_j)$ for any level j such that $1 \le j \le n$:

$$P_j(x_{j,cf}|x_j) = \frac{D_j \cdot \binom{|G|}{2} \cdot \binom{H_j-2}{x_j-2}}{\binom{H_j}{x_j}}. \tag{8}$$

$\binom{|G|}{2}$ is the number of the possible catastrophic failure scenarios *in a single group* ($m = 1$ thus any two process crashes in one group are catastrophic). D_j is the number of such single-group scenarios *at the whole level* j and is equal to $\left\lceil \frac{H_j}{|G|} \right\rceil$ (see Figure 8 for intuitive explanation). $\binom{H_j-2}{x_j-2}$ is the number of the remaining possible failure scenarios and $\binom{H_j}{x_j}$ is the total number of the possible failure scenarios. Second, for remaining levels j ($n+1 \le j \le h$) $\mathscr{M}$ is *not* t-aware and thus in the worst-case scenario any element crash is catastrophic: $P_j(x_{j,cf}|x_j) = 1$. The final formula for P_{cf} is thus

$$P_{cf} = \sum_{j=1}^{n} \sum_{x_j=1}^{H_j} P_j(x_j) \frac{D \cdot \binom{|G|}{2} \cdot \binom{H_j-2}{x_j-2}}{\binom{H_j}{x_j}} + \sum_{j=n+1}^{h} \sum_{x_j=1}^{N_j} P_j(x_j). \tag{9}$$

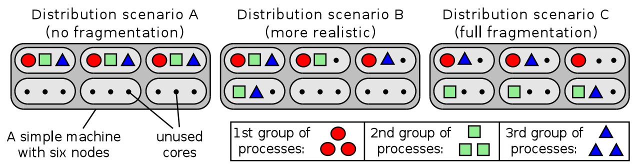


Figure 8: (§ 5.2) Consider three process distribution scenarios by $\mathscr{M}$ (*each* is t-aware). Optimistically, processes can be distributed contiguously (scenario A) or partially fragmented (scenario B). To get the upper bound for P_{cf} we use the worst-case pattern (scenario C). Now, to get the number of single-group CF scenarios at the whole level j (D_j), we need to obtain the number of the groups of *hardware elements* at j that hold process groups: $\lceil H_j / |G| \rceil$.

6. HOLISTIC RESILIENCE PROTOCOL

We now describe an example conceptual implementation of holistic fault tolerance for RMA that we developed to understand the tradeoffs between the resilience and performance in RMA-based systems. We implement it as a portable library (based on C and MPI) called FTRMA. We utilize MPI-3's one sided interface, but any other RMA model enabling relaxed memory consistency could be used instead (e.g., UPC or Fortran 2008). We use the publicly available FOMPI implementation of MPI-3 one sided as MPI library [1] but any other MPI-3 compliant library would be suitable. For simplicity we assume that the user application uses one contiguous region of shared memory of the same size at each process. Still, all the conclusions drawn are valid for any other application pattern based on RMA. Following the MPI-3 specification, we call this shared region of memory at every process a *window*. Finally, we divide user processes (referred to as CoMputing processes, CMs) into groups (as described in Section 5) and add one CHecksum process (denoted as CH) per group ($m = 1$). For any computing process p, we denote the CH in its group as $CH(p)$. CHs store and update XOR checksums of their CMs.

6.1 Protocol Overview

In this section we provide a general overview of the layered protocol implementation (see Figure 9). The first part (layer 1) logs accesses. The second layer takes uncoordinated checkpoints (called *demand* checkpoints) to trim the logs. Layer 3 performs regular coordinated checkpoints. All layers are diskless. Causal recovery replays memory accesses. Finally, our FDH increases resilience of the whole protocol.

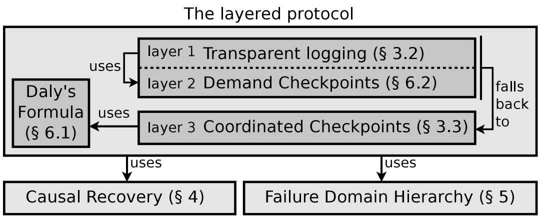


Figure 9: The overview of the protocol (§ 6.1). Layer 1 and 2 constitute the uncoordinated part of the protocol that falls back to the coordinated checkpointing if logging fails or if its overhead is too high.

Daly's Interval Layer 3 uses Daly's formula [28] as the optimum interval between coordinated checkpoints: $\sqrt{2\delta M} \cdot [1 + 1/3\sqrt{\delta/(2M)} + (1/9)(\delta/(2M))] - \delta$ (for $\delta < 2M$), or M (for $\delta \geq 2M$). M is the MTBF (mean time between failure) for the target machine and δ is the time to take a checkpoint. The user provides M while δ is estimated by our protocol.

Interfacing with User Programs and Runtime FTRMA routines are called after each RMA action. This would entail runtime system calls in compiled languages and we use the PMPI profiling interface [60] in our implementation. During window creation the user can specify (1) the number of CHs, (2) MTBF, (3) whether to use topology-awareness. After window creation, the protocol divides processes into CMs and CHs. If the user enables t-awareness, groups of processes running on the same FDs are also created. In the current version FTRMA takes into account computing nodes when applying t-awareness.

6.2 Demand Checkpointing

Demand checkpoints address the problem of diminishing amounts of memory per core in today's and future computing centers. If free memory at CM process p is scarce, p selects the process q with the largest $LP_p[q]$ or $LG_p[q]$ and requests a demand checkpoint. First, p sends a *checkpoint request* to $CH(q)$ which, in turn, forces q to checkpoint. This can be done by: closing all the epochs, locking all the relevant data structures, calculating the XOR checksum, and: (1) streaming the result to $CH(q)$ piece by piece or (2) sending the result in one bulk. $CH(q)$ integrates the received checkpoint data into the existing XOR checksum. Variant (1) is memory-efficient, and (2) is less time-consuming. Next, q unlocks all the data structures. Finally, $CH(q)$ sends a confirmation with the epoch number $E(p \to q)$ and respective counters (GNC_q, GC_q, SC_q) to p. Process p can delete logs of actions a where $a.EC < E(p \to q)$, $a.GNC < GNC_q$, $a.GC < GC_q$, $a.SC < SC_q$.

7. TESTING AND EVALUATION

In this section we first analyze the resilience of our protocol using real data from TSUBAME2.0 [3] failure history. Then, we test the performance of FTRMA with a NAS benchmark [9] that computes 3D Fast Fourier Transformation and a distributed key-value store. We denote the number of CHs and CMs as $|CH|$ and $|CM|$, respectively.

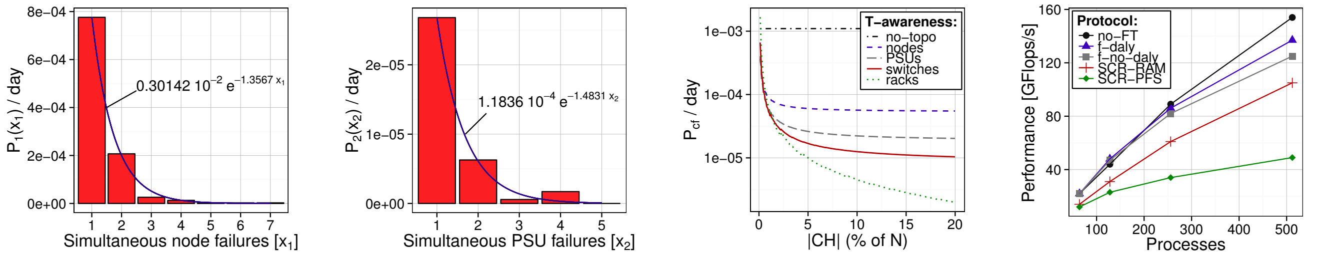
7.1 Analysis of Protocol Resilience

Our protocol stores all data in volatile memories to avoid I/O performance penalties and frequent disk and parallel file system failures [44, 76]. This brings several questions on whether the scheme is resilient in practical environments. To answer this question, we calculate the probability of a catastrophic failure P_{cf} (using Equations (7) and (9)) of our protocol, applying t-awareness at different levels of FDH.

We first fix model parameters (H_j, h) to reflect the hierarchy of TSUBAME2.0. TSUBAME2.0 FDH has 4 levels [76]: nodes, power supply units (PSUs), edge switches, and racks ($h = 4$) [76]. Then, to get P_{cf}, we calculate distributions $P_j(x_j)$ that determine the probability of x_j concurrent crashes at level j of TSUBAME FDH. To obtain $P_j(x_j)$ we analyzed 1962 crashes in the history of TSUBAME2.0 failures [3]. Following [14] we decided to use exponential probability distributions, where the argument is the number of concurrent failures x_j. We derived four probability density functions (PDFs) that approximate the failure distributions of, respectively: nodes ($0.30142 \cdot 10^{-2} e^{-1.3567 x_1}$), PSUs ($1.1836 \cdot 10^{-4} e^{-1.4831 x_2}$), switches ($3.9249 \cdot 10^{-5} e^{-1.5902 x_3}$), and racks ($3.2257 \cdot 10^{-5} e^{-1.5488 x_4}$). The unit is failures per day. Figures 10a and 10b illustrate two PDF plots with histograms. The distributions for PSUs, switches, and racks are based on real data only. For nodes it was not always possible to determine the exact correlation of failures. Thus, we pessimistically assumed (basing on [14]) that single crashes constitute 75% of all node failures, two concurrent crashes constitute 20%, and the remaining values decrease exponentially.

7.1.1 Comparison of Resilience

Figure 10c shows the resilience of our protocol when using five t-awareness strategies. The number of processes N is 4,000. P_{cf} is normalized to one day period. Without t-awareness (`no-topo`) a single crash of any FD of TSUBAME2.0 is catastrophic, thus P_{cf} does not depend on $|CH|$. In other scenarios every process from every group runs on a different node (`nodes`), PSU (`PSUs`), switch enclosure (`switches`) and rack (`racks`). In all cases P_{cf} decreases proportionally to the increasing $|CH|$, however at some point the exponential distributions ($P_j(x_j)$) begin to dominate the results. Topology-awareness at higher hierarchy levels significantly improves the resilience of our protocol. For example, if $CH = 5\%N$, P_{cf} in the `switches` scenario is ≈4 times lower than in `nodes`. Furthermore, all t-aware schemes are 1-3 orders of magnitude more resilient than `no-topo`.



(a) Distribution of node crashes (samples and the fit) (§ 7.1).
(b) Distribution of PSU crashes (samples and the fit) (§ 7.1).
(c) Probability of a catastrophic failure (§ 7.1.1).
(d) NAS FFT (class C) fault-free runs: checkpointing (§ 7.2.1).

Figure 10: Distribution of PSU & node failures, P_{cf} in TSUBAME2.0 running 4,000 processes, and the performance of NAS 3D FFT.

The results show that even a simple scheme (`nodes`) significantly improves the resilience of our protocol that performs only in-memory checkpointing and logging. We conclude that costly I/O flushes to the parallel file system (PFS) are not required for obtaining a high level of resilience. On the contrary, such flushes may even *increase* the risk of failures. They usually entail stressing the I/O system for significant amounts of time [76], and stable storage is often the element most susceptible to crashes. For example, a Blue Gene/P supercomputer had 4,164 disk fail events in 2011 (for 10,400 total disks) [44], and its PFS failed 77 times, almost two times more often than other hardware [44].

7.2 Analysis of Protocol Performance

We now discuss the performance of our fault tolerance protocol after the integration with two applications: NAS 3D FFT and a distributed key-value store. Both of these applications are characterized by intensive communication patterns, thus they demonstrate worst-case scenarios for our protocol. Integrating FTRMA with the application code was trivial and required minimal code changes resulting in the same code complexity.

Comparison to Scalable Checkpoint/Restart We compare FTRMA to Scalable Checkpoint-Restart (SCR) [2] a popular open-source message passing library that provides checkpoint and restart capability for MPI codes but does not enable logging. We turn on the XOR scheme in SCR and we fix the size of SCR groups [2] so that they match the analogous parameter in FTRMA ($|G|$). To make the comparison fair, we configure SCR to save checkpoints to both in-memory tmpfs (`SCR-RAM`) and to the PFS (`SCR-PFS`).

Comparison to Message Logging To compare the logging overheads in MP and RMA we also developed a simple MP scheme (basing on the protocol from [71]) that records accesses. Similarly to [71] we use additional processes to store protocol-specific logs; the data is stored at the sender's or receiver's side depending on the type of operation.

We execute all benchmarks on the Monte Rosa system and we use Cray XE6 computing nodes. Each node contains four 8-core 2.3 GHz AMD Opterons 6276 (Interlagos) and is connected to a 3D-Torus Gemini network. We use the Cray Programming Environment 4.1.46 to compile the code.

7.2.1 NAS 3D Fast Fourier Transformation

Our version of the NAS 3D FFT [9] benchmark is based on MPI-3 nonblocking PUTs (we exploit the overlap of computation and communication). The benchmark calculates 3D FFT using a 2D decomposition.

Performance of Coordinated Checkpointing We begin with evaluating our checkpointing "Gsync" scheme. Figure 10d illustrates the performance of NAS FFT fault-free runs. We compare: the original application code without any fault-tolerance (`no-FT`), FTRMA, `SCR-RAM`, and `SCR-PFS`. We fix $|CH| = 12.5\%|CM|$. We include two FTRMA scenarios: `f-daly` (use Daly's formula for coordinated checkpoints), and `f-no-daly` (fixed frequency of checkpoints without Daly's formula). We use the same t-awareness policy in all codes (`nodes`). The tested schemes have the respective fault-tolerance overheads over the baseline `no-FT`: 1-5% (`f-daly`), 1-15% (`f-no-daly`), 21-37% (`SCR-RAM`) and 46-67% (`SCR-PFS`). The performance of SCR-RAM is lower in comparison to `f-daly` and `f-no-daly` because the RMA-based checkpoints are faster due to higher performance of RDMA and less synchronization. `SCR-PFS` entails the highest overheads due to costly I/O flushes.

Performance of Demand Checkpointing We now analyze how the size of the log impacts the number of demand checkpoints and the performance of fault-free runs (see Figure 11a). Dedicating less than 44 MiB of memory for storing logs (per process) triggers demand checkpoint requests to clear the log. This results in performance penalties but leaves more memory available to the the user.

Second, we illustrate how $|CH|$ impacts the performance of recovering a process from its last demand checkpoint (see Figure 12). We run the NAS benchmark 10 times and after every such iteration we communicate the checksum necessary to recover the process. We use the `nodes` t-awareness and compare `no-FT`, `f-12.5-nodes` ($|CH| = 12.5\%|CM|$), and `f-6.25-nodes` ($|CH| = 6.25\%|CM|$). RMA's direct memory accesses ensure transparency and relatively small overheads: when $|CH| = 12.5\%|CM|$ 10 checksum transfers during 10 iterations make the run only 60% slower than `no-FT`.

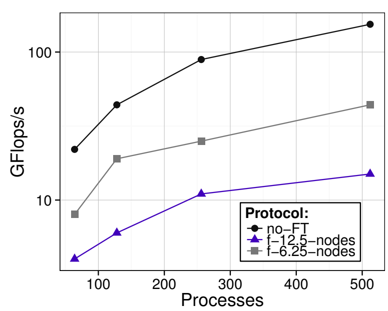
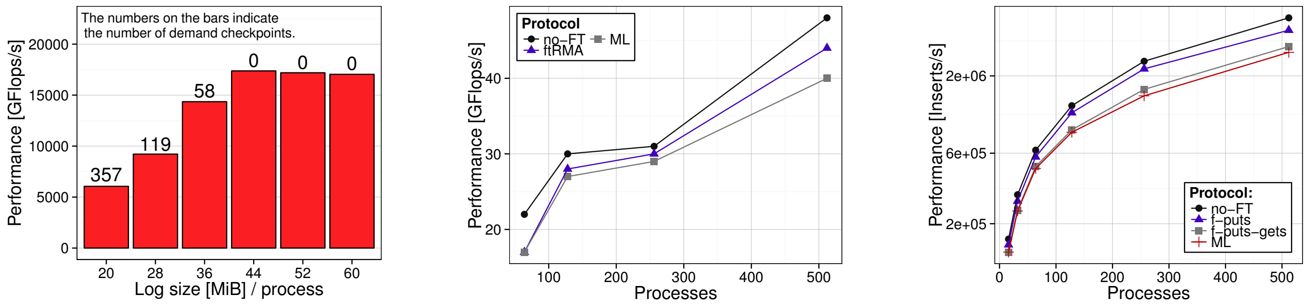


Figure 12: NAS FFT (class C) recovery from a demand checkpoint. (§ 7.2.1).



(a) NAS FFT (class A) fault-free runs: demand checkpointing.

(b) NAS FFT (class A) fault-free runs: logging.

(c) Key-value store fault-free runs.

Figure 11: Performance of the NAS FFT code (§ 7.2.1) and the key-value store (§ 7.2.2).

Performance of Access Logging As the next step we evaluate our logging scheme. Figure 11b illustrates the performance of fault-free runs. We compare `no-FT`, FTRMA, and our ML protocol (ML). FTRMA adds only ≈8% of overhead to the baseline (`no-FT`) and consistently outperforms ML by ≈9% due to the smaller amount of protocol-specific interaction between processes.

Varying |CH| and T-Awareness Policies Here, we analyze how $|CH|$ and t-awareness impact the performance of NAS FFT fault-free runs. We set $|CH| = 12.5\%|CM|$ and $|CH| = 6.25\%|CM|$, and we use the `no-topo` and `nodes` t-awareness policies. The results show that all these schemes differ negligibly from `no-FT` by 1-5%.

7.2.2 Key-Value Store

Our key-value store is based on a simple distributed hashtable (DHT) that stores 8–Byte integers. The DHT consists of parts called *local volumes* constructed with fixed-sized arrays. Every local volume is managed by a different process. Inserts are based on MPI-3 atomic Compare-And-Swap and Fetch-And-Op functions. Elements after hash collisions are inserted in the overflow heap that is the part of each local volume. To insert an element, a thread atomically updates the pointers to the next free cell and the last element in the local volume. Memory consistency is ensured with flushes. One GET and one PUT are logged if there is no hash collision, otherwise 6 PUTs and 4 GETs are recorded.

Performance of Access Logging We now measure the relative performance penalty of logging PUTs and GETs. During the benchmark, processes insert random elements with random keys. We focus on inserts only as they are perfectly representative for the logging evaluation. To simulate realistic requests, every process waits for a random time after every insert. The function that we use to calculate this interval is based on the exponential probability distribution: $f\delta e^{-\delta x}$, where f is a scaling factor, δ is a rate parameter and $x \in [0; b)$ is a random number. The selected parameter values ensure that every process spends ≈10% of the total runtime on inserting elements. For many computation-intense applications this is already a high amount of communication. We again compare `no-FT`, ML, and two FTRMA scenarios: `f-puts` (logging only PUTs) and `f-puts-gets` (logging PUTs and GETs). We fix $|CH| = 12.5\%|CM|$ and use the `nodes` t-awareness. We skip SCR as it does not enable logging.

We present the results in Figure 11c. For $N = 256$, the logging overhead over the baseline (`no-FT`) is: ≈12% (`f-puts`), 33% (`f-gets`), and 40% (ML). The overhead of logging PUTs in is due to the fact that every operation is recorded directly after issuing. Traditional message passing protocols suffer from a similar effect [32]. The overhead generated by logging GETs in `f-puts-gets` and ML is more significant because, due to RMA's one-sided semantics, every GET has to be recorded remotely. In addition, `f-puts-gets` suffers from synchronization overheads (caused by concurrent accesses to LG), while ML from inter-process protocol-specific communication. Discussed overheads heavily depend on the application type. Our key-value store constitutes a worst-case scenario because it does not allow for long epochs that could enable, e.g., sending the logs of multiple GETs in a bulk. The performance penalties would be smaller in applications that overlap computation with communication and use non blocking GETs.

8. RELATED WORK

In this section we discuss existing checkpointing and logging schemes (see Figure 13). For excellent surveys, see [7, 32, 81]. Existing work on fault tolerance in RMA/PGAS is scarce, an example scheme that uses PGAS for data replication can be found in [6].

8.1 Checkpointing Protocols

These schemes are traditionally divided into *uncoordinated*, *coordinated*, and *communication induced*, depending on process coordination scale [32]. There are also *complete* and *incremental* protocols that differ in checkpoint sizes [81].

Uncoordinated Schemes Uncoordinated schemes do not synchronize while checkpointing, but may suffer from *domino effect* or complex recoveries [32]. Example protocols are based on *dependency* [16] or *checkpoint graphs* [32]. A recent scheme targeting large-scale systems is Ken [85].

Coordinated Schemes Here, processes synchronize to produce consistent global checkpoints. There is no domino effect and recovery is simple but synchronization may incur severe overheads. Coordinated schemes can be *blocking* [32] or *non-blocking* [21]. There are also schemes based on *loosely synchronized clocks* [80] and *minimal coordination* [48].

Communication Induced Schemes Here, senders add scheme-specific data to application messages that receivers use to, e.g., avoid taking useless checkpoints. These schemes can be *index-based* [39] or *model-based* [32, 61].

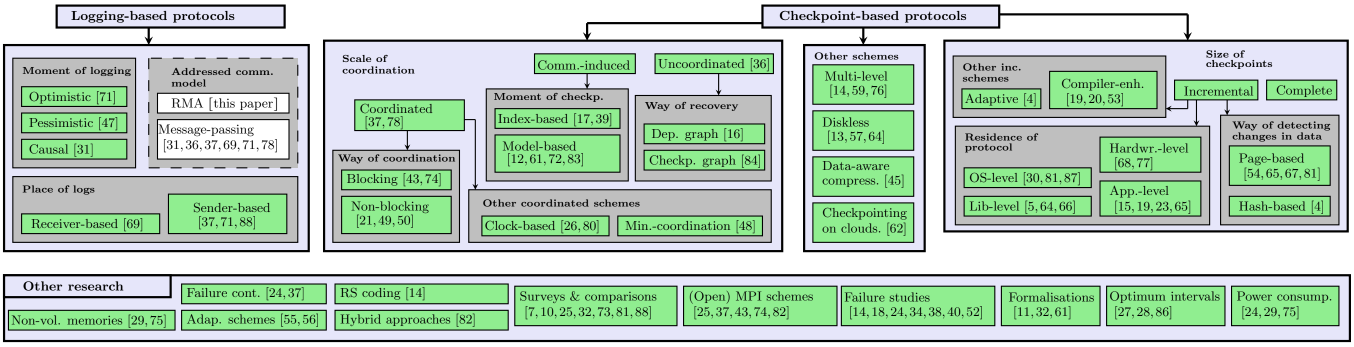


Figure 13: An overview of existing checkpointing and logging schemes (§ 8). A dashed rectangle illustrates a new sub-hierarchy introduced in the paper: dividing the logging protocols with respect to the *communication model* that they address.

Incremental Checkpointing An incremental checkpoint updates only the data that changed since the previous checkpoint. These protocols are divided into page-based [81] and hash-based [4]. They can reside at the level of an *application*, a *library*, an *OS*, or *hardware* [81]. Other schemes can be *compiler-enhanced* [20] or *adaptive* [4].

Others Recently, *multi-level* checkpointing was introduced [14, 59, 76]. *Adaptive* checkpointing based on failure prediction is discussed in [56]. A study on checkpointing targeted specifically at GPU-based computations can be found in [79]. [64] presents diskless checkpointing. Other interesting schemes are based on: Reed-Solomon coding [14], cutoff and compression to reduce checkpoint sizes [42], checkpointing on clouds [62], reducing I/O bottlenecks [46], and performant checkpoints to PFS [8].

8.2 Logging Protocols

Logging enables restored processes to replay their execution beyond the most recent checkpoint. Log-based protocols are traditionally categorized into: *pessimistic*, *optimistic*, *causal* [32]; they can also be *sender-based* [36, 71] and *receiver-based* [32] depending on which side logs messages.

Pessimistic Schemes Such protocols log events before they influence the system. This ensures no orphan processes and simpler recovery, but may incur severe overheads during fault-free runs. An example protocol is V-MPICH [33].

Optimistic Schemes Here, processes postpone logging messages to achieve, e.g., better computation-communication overlap. However, the algorithms for recovery are usually more complicated and crashed processes may become orphans [32]. A recent scheme can be found in [71].

Causal Schemes In such schemes processes log and exchange (by piggybacking to messages) dependencies needed for recovery. This ensures no orphans but may reduce bandwidth [32]. An example protocol is discussed in [31].

8.3 Other Important Studies & Discussion

Deriving an optimum checkpointing interval is presented in [28]. Formalizations targeting resilience can be found in [32, 61]. Power consumption was addressed in [24, 75]. *Containment domains* for encapsulating failures within a hierarchical scope are discussed in [24]. Modeling and prediction of failures is addressed in [14, 24]. Work on send determinism in MP can be found in [36].

Our study goes beyond the existing research scope presented in this section. First, we develop a fault tolerance model that covers virtually whole rich RMA semantics. Other existing formalizations (e.g., [7, 32, 61]) target MP only. We then use the model to formally analyze why resilience for RMA differs from MP and to design checkpointing, logging, and recovery protocols for RMA. We identify and propose solutions to several challenges in resilience for RMA that *do not* exist in MP, e.g.: consistency problems caused by the relaxed RMA memory model (§ 3.1, § 3.2.2, § 3.2.3), access non-determinism (§ 4.2), issues due to one-sided RMA communication (§ 3.2.1), logging multiple RMA-specific orders (§ 4.1), etc. Our model enables proving correctness of proposed schemes; all proofs omitted due to space constraints can be found in the techreport[1]. Extending our model for arbitrary hardware hierarchies generalizes the approach from [14] and enables formal reasoning about crashes of hardware elements and process distribution. Readers interested in a more extended analysis of this part are again advised to consult[1]. Finally, our protocol leverages and combines several important concepts and mechanisms (Daly's interval [28], multi-level design [59], etc.) to improve the resilience of RMA systems even further and is the first implementation of holistic fault tolerance for RMA.

9. CONCLUSION

RMA programming models are growing in popularity and importance as they allow for the best utilization of hardware features such as OS-bypass or zero-copy data transfer. Still, little work addresses fault tolerance for RMA.

We established, described, and explored a complete formal model of fault tolerance for RMA and illustrated how to use it to design and reason about resilience protocols running on flat and hierarchical machines. It will play an important role in making emerging RMA programming fault tolerant and can be easily extended to cover, e.g., stable storage.

Our study does not resort to traditional less scalable mechanisms that often rely on costly I/O flushes. The implementation of our holistic protocol adds negligible overheads to the applications runtime, for example 1-5% for in-memory checkpointing and 8% for fully transparent logging of remote memory accesses in the NAS 3D FFT code. Our probability study shows that the protocol offers high resilience. The idea of demand checkpoints will help alleviate the problem of limited memory amounts in today's petascale and future exascale computing centers.

Finally, our work provides the basis for further reasoning about fault-tolerance not only for RMA, but also for all the other models that can be constructed upon it, such as taskbased programming models. This will play an important role in complex heterogeneous large-scale systems.

10. REFERENCES

[1] foMPI, 2013. http://spcl.inf.ethz.ch/Research/Parallel_Programming/foMPI.

[2] Scalable Checkpoint / Restart, 2013. http://sourceforge.net/projects/scalablecr/.

[3] TSUBAME2.0: Failure History, April 2013. http://mon.g.gsic.titech.ac.jp/trouble-list/index.htm.

[4] S. Agarwal, R. Garg, M. S. Gupta, and J. E. Moreira. Adaptive incremental checkpointing for massively parallel systems. In *Proc. of Ann. Intl. Conf. on Supercomp.*, ICS '04, pages 277–286, 2004.

[5] A. Agbaria and J. S. Plank. Design, Implementation, and Performance of Checkpointing in NetSolve. In *Proc. of the 2000 International Conference on Dependable Systems and Networks (formerly FTCS-30 and DCCA-8)*, DSN '00, pages 49–.

[6] N. Ali, S. Krishnamoorthy, N. Govind, and B. Palmer. A Redundant Communication Approach to Scalable Fault Tolerance in PGAS Programming Models. In *Par., Dist. and Net. Proc. (PDP), the Eur. Intl. Conf. on*, pages 24–31, 2011.

[7] L. Alvisi and K. Marzullo. Message Logging: Pessimistic, Optimistic, Causal, and Optimal. *IEEE Trans. Softw. Eng.*, 24(2):149–159, Feb. 1998.

[8] D. Arteaga and M. Zhao. Towards Scalable Application Checkpointing with Parallel File System Delegation. In *Proc. of the IEEE Intl. Conf. on Net., Arch., and Stor.*, NAS '11, pages 130–139, 2011.

[9] D. H. Bailey, E. Barszcz, J. T. Barton, D. S. Browning, R. L. Carter, R. A. Fatoohi, P. O. Frederickson, T. A. Lasinski, H. D. Simon, V. Venkatakrishnan, and S. K. Weeratunga. The NAS parallel benchmarks. Technical report, The Intl. J. of Supercomp. App., 1991.

[10] P. Balaji, D. Buntinas, and D. Kimpe. Fault Tolerance Techniques for Scalable Computing. In L. W. Samee U. Khan, Albert Y. Zomaya, editor, *Scalable Computing and Communications: Theory and Practice*. 2013.

[11] R. Baldoni, J. Brzezinski, J. Helary, A. Mostefaoui, and M. Raynal. Characterization of consistent global checkpoints in large-scale distributed systems. In *Distributed Computing Systems, 1995., Proc. of the Fifth IEEE Computer Society Workshop on Future Trends of*, pages 314 –323, aug.

[12] R. Baldoni, F. Quaglia, and B. Ciciani. A VP-accordant checkpointing protocol preventing useless checkpoints. In *Reliable Distributed Systems, 1998. Proc.. Seventeenth IEEE Symposium on*, pages 61 –67, oct.

[13] L. Bautista Gomez, N. Maruyama, F. Cappello, and S. Matsuoka. Distributed Diskless Checkpoint for Large Scale Systems. In *Cluster, Cloud and Grid Computing (CCGrid), 2010 10th IEEE/ACM International Conference on*, pages 63–72.

[14] L. Bautista-Gomez, S. Tsuboi, D. Komatitsch, F. Cappello, N. Maruyama, and S. Matsuoka. FTI: high performance fault tolerance interface for hybrid systems. In *Proc. of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '11, pages 32:1–32:32.

[15] A. Beguelin, E. Seligman, and P. Stephan. Application level fault tolerance in heterogeneous networks of workstations. *J. Parallel Distrib. Comput.*, 43(2):147–155, June 1997.

[16] B. Bhargava and S.-R. Lian. Independent checkpointing and concurrent rollback for recovery in distributed systems-an optimistic approach. In *Reliable Distributed Systems, 1988. Proc.., Seventh Symposium on*, pages 3 –12, oct.

[17] D. Briatico, A. Ciuffoletti, and L. Simoncini. A distributed domino-effect free recovery algorithm. In *Symposium on Reliability in Distributed Software and Database Systems'84*, pages 207–215.

[18] G. Bronevetsky, I. Laguna, B. De Supinski, and S. Bagchi. Automatic fault characterization via abnormality-enhanced classification. In *Dependable Systems and Networks (DSN), 2012 42nd Annual IEEE/IFIP International Conference on*, pages 1–12.

[19] G. Bronevetsky, D. Marques, K. Pingali, and P. Stodghill. Automated application-level checkpointing of MPI programs. *SIGPLAN Not.*, 38(10):84–94, June.

[20] G. Bronevetsky, D. J. Marques, K. K. Pingali, R. Rugina, and S. A. McKee. Compiler-enhanced incremental checkpointing for OpenMP applications. In *Proc. of the 13th ACM SIGPLAN Symposium on Principles and practice of parallel programming*, PPoPP '08, pages 275–276.

[21] K. M. Chandy and L. Lamport. Distributed snapshots: determining global states of distributed systems. *ACM Trans. Comput. Syst.*, 3(1):63–75, Feb. 1985.

[22] P. M. Chen, E. K. Lee, G. A. Gibson, R. H. Katz, and D. A. Patterson. RAID: high-performance, reliable secondary storage. *ACM Comput. Surv.*, 26(2):145–185, June 1994.

[23] Y. Chen, J. S. Plank, and K. Li. CLIP: a checkpointing tool for message-passing parallel programs. In *Proc. of the 1997 ACM/IEEE conference on Supercomputing (CDROM)*, Supercomputing '97, pages 1–11.

[24] J. Chung, I. Lee, M. Sullivan, J. H. Ryoo, D. W. Kim, D. H. Yoon, L. Kaplan, and M. Erez. Containment domains: a scalable, efficient, and flexible resilience scheme for exascale systems. In *Proc. of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 58:1–58:11.

[25] C. Coti, T. Herault, P. Lemarinier, L. Pilard, A. Rezmerita, E. Rodriguezb, and F. Cappello. Blocking vs. Non-Blocking Coordinated Checkpointing for Large-Scale Fault Tolerant MPI. In *SC 2006 Conference, Proc. of the ACM/IEEE*, page 18, nov.

[26] F. Cristian and F. Jahanian. A timestamp-based checkpointing protocol for long-lived distributed computations. In *Reliable Distributed Systems, 1991. Proc.., Tenth Symposium on*, pages 12 –20, sep-2 oct.

[27] J. Daly. A model for predicting the optimum checkpoint interval for restart dumps. In *Proc. of the 2003 international conference on Computational*

science, ICCS'03, pages 3–12.

[28] J. T. Daly. A higher order estimate of the optimum checkpoint interval for restart dumps. *Future Gener. Comput. Syst.*, 22(3):303–312, Feb. 2006.

[29] X. Dong, N. Muralimanohar, N. Jouppi, R. Kaufmann, and Y. Xie. Leveraging 3D PCRAM technologies to reduce checkpoint overhead for future exascale systems. In *Proc. of the Conference on High Performance Computing Networking, Storage and Analysis*, SC '09, pages 57:1–57:12.

[30] J. Duell. The design and implementation of Berkeley Lab's linux Checkpoint/Restart. Technical report, 2003.

[31] E. Elnozahy and W. Zwaenepoel. Manetho: transparent roll back-recovery with low overhead, limited rollback, and fast output commit. *Computers, IEEE Transactions on*, 41(5):526 –531, may 1992.

[32] E. N. M. Elnozahy, L. Alvisi, Y.-M. Wang, and D. B. Johnson. A survey of rollback-recovery protocols in message-passing systems. *ACM Comput. Surv.*, 34(3):375–408, Sept. 2002.

[33] G. Bosilca et al. MPICH-V: Toward a Scalable Fault Tolerant MPI for Volatile Nodes. In *Supercomputing, the ACM/IEEE Conf.*, pages 29–29, 2002.

[34] A. Gainaru, F. Cappello, M. Snir, and W. Kramer. Fault prediction under the microscope: a closer look into HPC systems. In *Proc. of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 77:1–77:11.

[35] R. Gerstenberger, M. Besta, and T. Hoefler. Enabling Highly-scalable Remote Memory Access Programming with MPI-3 One Sided. In *Proc. of ACM/IEEE Supercomputing*, SC '13, pages 53:1–53:12, 2013.

[36] A. Guermouche, T. Ropars, E. Brunet, M. Snir, and F. Cappello. Uncoordinated Checkpointing Without Domino Effect for Send-Deterministic MPI Applications. In *Parallel Distributed Processing Symposium (IPDPS), 2011 IEEE International*, pages 989–1000.

[37] A. Guermouche, T. Ropars, M. Snir, and F. Cappello. HydEE: Failure Containment without Event Logging for Large Scale Send-Deterministic MPI Applications. In *Parallel Distributed Processing Symposium (IPDPS), 2012 IEEE 26th International*, pages 1216–1227.

[38] E. Heien, D. Kondo, A. Gainaru, D. LaPine, B. Kramer, and F. Cappello. Modeling and tolerating heterogeneous failures in large parallel systems. In *Proc. of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '11, pages 45:1–45:11.

[39] J.-M. Helary, A. Mostefaoui, R. Netzer, and M. Raynal. Preventing useless checkpoints in distributed computations. In *Reliable Distributed Systems, 1997. Proc.., The Sixteenth Symposium on*, pages 183 –190, oct.

[40] C.-H. Ho, M. de Kruijf, K. Sankaralingam, B. Rountree, M. Schulz, and B. De Supinski. Mechanisms and Evaluation of Cross-Layer Fault-Tolerance for Supercomputing. In *Parallel Processing (ICPP), 2012 41st International Conference on*, pages 510–519.

[41] T. Hoefler, J. Dinan, R. Thakur, B. Barrett, P. Balaji, W. Gropp, and K. Underwood. Remote Memory Access Programming in MPI-3. *Argonne National Laboratory, Tech. Rep*, 2013.

[42] S. Hogan, J. Hammond, and A. Chien. An evaluation of difference and threshold techniques for efficient checkpoints. In *Dep. Sys. and Net. Work. (DSN-W), IEEE/IFIP Intl. Conf.*, pages 1–6, 2012.

[43] J. Hursey, T. I. Mattox, and A. Lumsdaine. Interconnect agnostic checkpoint/restart in Open MPI. In *Proc. of the 18th ACM international symposium on High Performance Distributed Computing HPDC*, pages 49–58, 2009.

[44] F. Isaila, J. Garcia, J. Carretero, R. Ross, and D. Kimpe. Making the case for reforming the I/O software stack of extreme-scale systems. In *Ex. App. and Soft. Conf. (EASC'13)*, 2013.

[45] T. Z. Islam, K. Mohror, S. Bagchi, A. Moody, B. R. de Supinski, and R. Eigenmann. McrEngine: a scalable checkpointing system using data-aware aggregation and compression. In *Proc. of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 17:1–17:11.

[46] H. Jin and K. Hwang. Distributed Checkpointing on Clusters with Dynamic Striping and Staggering. In *Advances in Computing Science - ASIAN 2002*, volume 2550, pages 19–33. 2002.

[47] T.-Y. Juang and S. Venkatesan. Crash recovery with little overhead. In *Distributed Computing Systems, 1991., 11th International Conference on*, pages 454–461.

[48] R. Koo and S. Toueg. Checkpointing and Rollback-Recovery for Distributed Systems. *Software Engineering, IEEE Transactions on*, SE-13(1):23 – 31, jan. 1987.

[49] A. Kshemkalyani. Fast and Message-Efficient Global Snapshot Algorithms for Large-Scale Distributed Systems. *Parallel and Distributed Systems, IEEE Transactions on*, 21(9):1281 –1289, sept. 2010.

[50] T. H. Lai and T. H. Yang. On distributed snapshots. *Information Processing Letters*, 25(3):153–158, 1987.

[51] L. Lamport. Time, Clocks, and the Ordering of Events in a Distributed System. *Commun. ACM*, 21(7):558–565, July 1978.

[52] I. Lee, M. Basoglu, M. Sullivan, D. H. Yoon, L. Kaplan, and M. Erez. Survey of error and fault detection mechanisms. *University of Texas at Austin, Tech. Rep*, 2011.

[53] C.-C. J. Li and W. Fuchs. CATCH-compiler-assisted techniques for checkpointing. In *Fault-Tolerant Computing, 1990. FTCS-20. Digest of Papers., 20th International Symposium*, pages 74–81.

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

[55] Y. Li and Z. Lan. Exploit Failure Prediction for Adaptive Fault-Tolerance in Cluster Computing. In *Proc. of the Sixth IEEE International Symposium on Cluster Computing and the Grid*, CCGRID '06, pages

531–538.

[56] Y. Li and Z. Lan. Using adaptive fault tolerance to improve application robustness on the Teragrid. *Proc. of TeraGrid*, 322, 2007.

[57] C.-D. Lu. *Scalable diskless checkpointing for large parallel systems*. PhD thesis, University of Illinois, 2005.

[58] J. Manson, W. Pugh, and S. V. Adve. The Java Memory Model. In *Proc. of ACM Symp. on Princ. of Prog. Lang.*, POPL '05, pages 378–391, 2005.

[59] A. Moody, G. Bronevetsky, K. Mohror, and B. R. d. Supinski. Design, Modeling, and Evaluation of a Scalable Multi-level Checkpointing System. In *Proc. of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '10, pages 1–11.

[60] MPI Forum. MPI: A Message-Passing Interface Standard. Version 3, September 2012. available at: http://www.mpi-forum.org (Sep. 2012).

[61] R. Netzer and J. Xu. Necessary and sufficient conditions for consistent global snapshots . *Parallel and Distributed Systems, IEEE Transactions on*, 6(2):165 –169, feb 1995.

[62] B. Nicolae and F. Cappello. BlobCR: efficient checkpoint-restart for HPC applications on IaaS clouds using virtual disk image snapshots. In *Proc. of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '11, pages 34:1–34:12.

[63] D. Petrović, O. Shahmirzadi, T. Ropars, and A. Schiper. High-performance RMA-based broadcast on the Intel SCC. In *Proc. of ACM Symp. Par. Alg. Arch.*, SPAA '12, pages 121–130, 2012.

[64] J. Plank, K. Li, and M. Puening. Diskless checkpointing. *Parallel and Distributed Systems, IEEE Transactions on*, 9(10):972–986, 1998.

[65] J. S. Plank, M. Beck, G. Kingsley, and K. Li. Libckpt: transparent checkpointing under Unix. In *Proc. of the USENIX 1995 Technical Conference Proc.*, TCON'95, pages 18–18.

[66] J. S. Plank and K. Li. ickp: A Consistent Checkpointer for Multicomputers. *IEEE Parallel Distrib. Technol.*, 2(2):62–67, June 1994.

[67] J. S. Plank, J. Xu, and R. H. B. Netzer. Compressed differences: An algorithm for fast incremental checkpointing. Technical Report CS-95-302, University of Tennessee, August 1995.

[68] M. Prvulovic, Z. Zhang, and J. Torrellas. ReVive: cost-effective architectural support for rollback recovery in shared-memory multiprocessors. In *Proc. of the 29th annual international symposium on Computer architecture*, ISCA '02, pages 111–122.

[69] S. Rao, L. Alvisi, and H. M. Vin. Hybrid Message Logging Protocols for Fast Recovery. In *In Digest of FastAbstracts of Fault-Tolerant Computing Symposium (FTCS-28)*, pages 41–42. IEEE Computer Society, 1998.

[70] I. S. Reed and G. Solomon. Polynomial codes over certain finite fields. *Journal of the Society for Industrial & Applied Mathematics*, 8(2):300–304, 1960.

[71] R. Riesen, K. Ferreira, D. Da Silva, P. Lemarinier, D. Arnold, and P. G. Bridges. Alleviating scalability issues of checkpointing protocols. In *Proc. of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 18:1–18:11.

[72] D. Russell. State Restoration in Systems of Communicating Processes. *Software Engineering, IEEE Transactions on*, SE-6(2):183 – 194, march 1980.

[73] J. Sancho, F. Petrini, G. Johnson, and E. Frachtenberg. On the feasibility of incremental checkpointing for scientific computing. In *Parallel and Distributed Processing Symposium, 2004. Proc.. 18th International*, pages 58–.

[74] S. Sankaran, J. M. Squyres, B. Barrett, V. Sahay, A. Lumsdaine, J. Duell, P. Hargrove, and E. Roman. The LAM/MPI Checkpoint/Restart Framework: System-Initiated Checkpointing. *International Journal of High Performance Computing Applications*, 19(4):479–493, Winter 2005.

[75] S. Sardashti and D. A. Wood. UniFI: leveraging non-volatile memories for a unified fault tolerance and idle power management technique. In *Proc. of the 26th ACM international conference on Supercomputing*, ICS '12, pages 59–68.

[76] K. Sato, N. Maruyama, K. Mohror, A. Moody, T. Gamblin, B. R. de Supinski, and S. Matsuoka. Design and modeling of a non-blocking checkpointing system. In *Proc. of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '12, pages 19:1–19:10.

[77] D. J. Sorin, M. M. K. Martin, M. D. Hill, and D. A. Wood. SafetyNet: improving the availability of shared memory multiprocessors with global checkpoint/recovery. In *Proc. of the 29th annual international symposium on Computer architecture*, ISCA '02, pages 123–134.

[78] S. Sumimoto. Design and evaluation of fault-tolerant shared file system for cluster systems. In *Proc. of the The Twenty-Sixth Annual International Symposium on Fault-Tolerant Computing (FTCS '96)*, FTCS '96, pages 74–.

[79] H. Takizawa, K. Koyama, K. Sato, K. Komatsu, and H. Kobayashi. CheCL: Transparent Checkpointing and Process Migration of OpenCL Applications. In *Parallel Distributed Processing Symposium (IPDPS), 2011 IEEE International*, pages 864–876.

[80] Z. Tong, R. Y. Kain, and W. T. Tsai. Rollback Recovery in Distributed Systems Using Loosely Synchronized Clocks. *IEEE Trans. Parallel Distrib. Syst.*, 3(2):246–251, Mar. 1992.

[81] M. Vasavada, F. Mueller, P. H. Hargrove, and E. Roman. Comparing different approaches for incremental checkpointing: The showdown. In *Linux Symposium*, page 69, 2011.

[82] C. Wang, F. Mueller, C. Engelmann, and S. Scott. Hybrid Checkpointing for MPI Jobs in HPC Environments. In *Parallel and Distributed Systems (ICPADS), 2010 IEEE 16th International Conference on*, pages 524–533.

[83] Y.-M. Wang. Consistent global checkpoints that contain a given set of local checkpoints. *Computers, IEEE Transactions on*, 46(4):456 –468, apr 1997.

[84] Y.-M. Wang, P.-Y. Chung, I.-J. Lin, and W. Fuchs. Checkpoint space reclamation for uncoordinated checkpointing in message-passing systems. *Parallel and Distributed Systems, IEEE Transactions on*, 6(5):546 –554, may 1995.

[85] S. Yoo, C. Killian, T. Kelly, H. K. Cho, and S. Plite. Composable Reliability for Asynchronous Systems. In *Proc. of the USENIX Ann. Tech. Conf.*, USENIX ATC'12, pages 3–3, 2012.

[86] J. W. Young. A first order approximation to the optimum checkpoint interval. *Commun. ACM*, 17(9):530–531, Sept. 1974.

[87] H. Zhong and J. Nieh. CRAK: Linux checkpoint/restart as a kernel module. Technical report, Technical Report CUCS-014-01, Department of Computer Science, Columbia University, 2001.

[88] W. Zwaenepoel and D. Johnson. Sender-based message logging. In *17th International Symposium on Fault-Tolerant Computing*, 1987.