

## MODULARISING VERIFICATION OF DURABLE OPACITY

ELENI BILA <sup>a</sup>, JOHN DERRICK <sup>b</sup>, SIMON DOHERTY <sup>b</sup>, BRIJESH DONGOL <sup>a</sup>,  
GERHARD SCHELLHORN <sup>c</sup>, AND HEIKE WEHRHEIM <sup>d</sup>

<sup>a</sup> University of Surrey  
*e-mail address:* e.vafeiadibila@surrey.ac.uk  
*e-mail address:* b.dongol@surrey.ac.uk

<sup>b</sup> University of Sheffield  
*e-mail address:* j.derrick@sheffield.ac.uk  
*e-mail address:* s.doherty@sheffield.ac.uk

<sup>c</sup> University of Augsburg  
*e-mail address:* schellhorn@informatik.uni-augsburg.de

<sup>d</sup> University of Oldenburg  
*e-mail address:* heike.wehrheim@uol.de

**ABSTRACT.** Non-volatile memory (NVM), also known as persistent memory, is an emerging paradigm for memory that preserves its contents even after power loss. NVM is widely expected to become ubiquitous, and hardware architectures are already providing support for NVM programming. This has stimulated interest in the design of novel concepts ensuring correctness of concurrent programming abstractions in the face of persistency and in the development of associated verification approaches.

Software transactional memory (STM) is a key programming abstraction that supports concurrent access to shared state. In a fashion similar to linearizability as the correctness condition for concurrent data structures, there is an established notion of correctness for STMs known as opacity. We have recently proposed *durable opacity* as the natural extension of opacity to a setting with non-volatile memory. Together with this novel correctness condition, we designed a verification technique based on refinement. In this paper, we extend this work in two directions. First, we develop a durably opaque version of NOrec (no ownership records), an existing STM algorithm proven to be opaque. Second, we modularise our existing verification approach by separating the proof of durability of memory accesses from the proof of opacity. For NOrec, this allows us to *re-use* an existing opacity proof and complement it with a proof of the durability of accesses to shared state.

---

*Key words and phrases:* Nonvolatile memory, software transactional memory, opacity, formal verification.

Derrick and Doherty are supported by EPSRC project EP/R032351/1. Dongol is supported by EPSRC Grants EP/R032556/1, EP/V038915/1 and EP/R025134/2. Bila and Dongol are supported by VeTSS project “Persistent Safety and Security”. Wehrheim is partially supported by DFG grant WE2290/12-1.

## 1. INTRODUCTION

Non-volatile memory (NVM) promises the combination of the density and non-volatility of NAND Flash-based solid-state disks (SSDs) with the performance of volatile memory (RAM). The term *persistent memory* is used to describe an NVM technology that presents two characteristics: **(1)** directly byte-addressable access from the user space by using byte-addressable operations and **(2)** preservation of its contents even after system crashes and power failures. NVM is intended to be used as an intermediate layer between traditional volatile memory (VM) and secondary storage, and has the potential to vastly improve system speed and stability. Speed-ups of 2-3 orders of magnitude are likely to be feasible over and above hard disks. Furthermore, software that uses NVM has the potential to be more robust; in case of a crash, a system state before the crash may be recovered using contents from NVM, as opposed to being restarted from secondary storage. For these reasons alone, NVM is widely expected to become ubiquitous, and hardware architectures are already providing support for NVM programming.

However, writing correct NVM programs is extremely difficult, as the semantics of persistency can be unclear. Furthermore, because the same data is stored in both a volatile and non-volatile manner, and because NVM is updated at a slower rate than VM, recovery to a consistent state may not always be possible. This is particularly true for concurrent systems, where coping with NVM requires introduction of additional synchronisation instructions into a program. Such instructions are already supported by Intel-x86 and ARMv8.

This has led to work on the design of the first persistent concurrent programming abstractions, so far mainly concurrent data structures [ZFS<sup>+</sup>19, FHMP18, FPR21, VTS11, VTRC11]. To support the reasoning about correctness for these abstractions working over NVM, a coherent notion of correctness is needed. Such a notion for concurrent data structures has been defined by Izraelevitz et al. [IMS16] (known as *durable linearizability*) which naturally generalises the standard linearizability correctness condition [HW90]. A first proof technique for showing durable linearizability has been proposed by Derrick et al. [DDD<sup>+</sup>19].

In this paper we investigate another key programming abstraction known as *Software Transactional Memory* (STM) that supports concurrent access to shared state. STM is a mechanism that provides an illusion of atomicity in concurrent programs and aims to reduce the burden on programmers of implementing complicated synchronisation mechanisms. The analogy of STM is with database transactions, which perform a series of accesses/updates to shared data (via read and write operations) atomically in an all-or-nothing manner. Similarly with an STM, if a transaction commits, all its operations succeed, and in the aborting case, all its operations fail. STMs are now part of mainstream programming, e.g., the ScalaSTM library, a new language feature in Clojure that uses an STM implementation internally for all data manipulation and the G++ 4.7 compiler (which supports STM features directly in the compiler).

In a fashion similar to linearizability as the correctness condition for concurrent data structures, there is an established notion of correctness for STMs known as *opacity* [GK08]. Overall, opacity guarantees that committed transactions appear as if they are executed atomically, at some unique point in time, and aborted transactions, as if they did not execute at all. Amongst other things, opacity also guarantees that all reads that a transaction performs are valid with respect to a single memory snapshot.

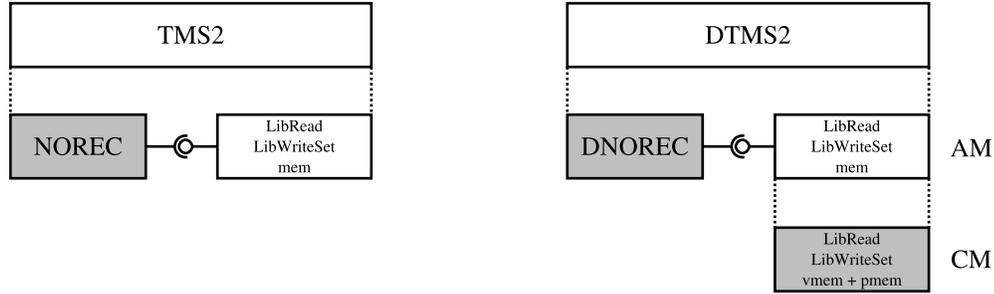


Figure 1: Original proof of opacity (left) vs. proof of durable opacity (right)

A fundamental challenge when developing STMs for persistent memory is to ensure a correct recovery after a crash. This requires that, at any point in the execution of the program, the persistent state must be sufficient to enable the recovery procedure to recreate an appropriate consistent state. Verification of STMs has to show that this is achieved by the proposed algorithm, i.e., that enough data is persisted and the recovery procedure correctly uses this data to guarantee opacity.

In this paper, we investigate STMs and their correctness via opacity on non-volatile memory architectures. Doing this entails a number of steps. First, the correctness criterion of opacity has to be adapted to cope with crashes in system executions. Second, STM algorithms have to be extended to deal with the coexistence of volatile and non-volatile memory during execution and need to be equipped with recovery operations. Third, proof techniques for opacity need to be re-investigated to make them usable for durable opacity.

In our prior work [BDD<sup>+</sup>20], we have addressed the steps above as follows. The first step is addressed by defining a notion of correctness called *durable opacity*, which generalises opacity in the same way that durable linearizability [IMS16] generalises linearizability for NVM architectures. Durable opacity requires executions of STMs to be opaque even if they are interspersed with crashes. The second step is addressed by *developing* a durable version of the Transactional Mutex Lock [DDS<sup>+</sup>10]. Finally, the third step is addressed by *proving* durable opacity of this new algorithm using a refinement-based approach.

This paper extends prior work [BDD<sup>+</sup>20] via the development of a *modular approach* to verifying durable opacity. Our new approach is inspired by the modularised verification of a filesystem for flash memory [PEB<sup>+</sup>17, BSR22]. The proof technique separates the proof of opacity (perceived atomicity of transactions) from the proof of durability (correct handling of non-volatile memory). Our proof technique assumes the existence of an STM that has been verified to be opaque by proving that it refines the specification TMS2 [DGLM13] (which itself has been shown to satisfy opacity [LLM12a]). This refinement proof is then *re-used* to construct a durably opaque version of the STM. We exemplify our technique by extending the No-Ownership-Records (NOREC) STM of Dalessandro et al. [DSS10] to create a durable NOREC.

Figure 1 illustrates our approach. The original NOREC algorithm is shown to the left. It has already been proven opaque by showing that it refines (dashed lines) the TMS2 automaton by Lesani et al. [LLM12a] using the PVS prover. The algorithm can be thought of as having an implicit interface to main memory *mem* (indicated by the  $\text{---}\ominus\text{---}$  symbol), which allows to read and write memory cells. Since NOREC is a lazy algorithm, writing is

confined to committing a write set to ensure that two transactions cannot commit their write set at the same time, producing a mixed result that would contradict opacity. NOREC enforces that there is at most one transaction committing a write set any time. Our approach will first make this interface with operations `LibRead` and `LibWriteSet` explicit and call it AM. The tricky bit in defining the interface is the enforcement of the constraint of a single committer as an ownership annotation for AM<sup>1</sup>. (This annotation parallels the use of an auxiliary variable in the original PVS proof of NOREC [LLM12a].)

It can then be observed that if a) all reads and writes to memory were directly to persistent memory and b) committing a write set is atomic, then the resulting algorithm is already durable opaque since the content of memory is preserved on a crash. Crashes in the middle of commits that could lead to a state that is not compatible with durable opacity are then avoided. As a consequence, we can reuse the original opacity proof with only minor adjustments. The main change is that using the abstract DTMS2 automaton (Figure 1, right) to express durable opacity adds the proof obligation that a crash does indeed not have any relevant effect. Since the original opacity proof is by far the most complex proof needed, reusing it saves a lot of work compared to verification from scratch. Of course, assumptions a) and b) above are not realistic assumptions when viewing AM as an *implementation*. However, AM can also be viewed as a *specification* of a library that can be refined to a non-atomic, concurrent implementation. We define such an implementation CM. It basically uses volatile memory *vmem* as a cache for persistent memory *pmem*. A logging mechanism ensures that a recovery procedure that runs on restarting from a crash can undo the effects of a partially completed transaction. The correctness proof for the refinement then is completely *separate* from the main proof. It shows that CM is a durable linearizable implementation of AM. We then prove in general, that two refinements constructed in this way together always give a proof of durable opacity for an algorithm which combines the two implementations shown in grey in the figure (written  $\text{dNOREC}[\text{CM}]$  for our case).

The approach of this paper therefore can be viewed as a blueprint for a modular strategy, that allows to transform an STM implementation that is opaque to a durably opaque one. In particular, we believe our modularisation technique can be used on any transactional memory algorithm that uses a write-log and serialises commits [DSS10, DSS06, SMvP08a].

The difficult bit for each algorithm will be the definition of an interface AM with suitable ownership conditions, that ensure that its implementation CM only has to deal with a suitably restricted form of concurrency (here: no two commits at the same time). However such restrictions must have already been relevant for proving opacity of the original algorithm, so similar to our case it should be possible to move them to constraints on AM, to reuse the original proof and to construct a separate refinement to CM.

We mechanise the proof of durable linearizability of the library in the theorem prover KIV [SBBR22]. KIV is also used to mechanise a general result on using refinement in a context which specialises to the result that the two refinement proofs together imply durable opacity of the final algorithm with library calls. These proofs are available online [BDD<sup>+</sup>21].

**Overview.** This paper is organised as follows. In section 2, we give background for this paper, and present the execution model and the formal definitions of (durable) linearizability and (durable) opacity. In section 3, we present the use of IOA to verify correctness of durable

<sup>1</sup>Other algorithms, like TL2 enforce disjoint write sets to ensure that there are no conflicts, which would result in a similar interface with a modified concept of ownership (in this case about memory locations).

concurrent objects. Our modular proof technique is described in section 4, which describes the transformation of an opaque algorithm to satisfy durable opacity, the modularisation of memory accesses using an abstract library and its fine-grained refinement of the abstract library to a concrete library. Both the modularisation and library refinement steps are guaranteed to preserve durable opacity. We cover related work in section 5.

## 2. FOUNDATIONS

We start by explaining some basic assumptions we make about the memory model and by explaining how persistent and volatile memory interact. We then define the correctness conditions relevant for our approach. These are *linearizability* [HW90] and *opacity* [GK08]. Linearizability (or better to say, its adaption for NVM) is part of our proof method, and the NVM-version of opacity, durable opacity, is the concurrent correctness criterion we intend to prove for STMs. Both correctness conditions formalise some form of atomicity in which a block of code executes seemingly atomically in an all-or-nothing manner. The difference lays in the level of atomicity: for linearizability blocks of code describe one operation of a concurrent data structure; for opacity we also have blocks of code for specific operations, and in addition group such operations into *transactions*.

**2.1. Memory model, crashes and recovery.** We assume that the shared state consists of a set  $Loc$  of locations and contains values from a set  $Val$ . Threads can concurrently access locations, and we assume these accesses to be sequentially consistent (SC [Lam79]).

Algorithms running on NVM architectures operate on two versions of memory: *persistent* and *volatile* memory (later denoted as  $pmem$  and  $vmem$ , respectively). In an NVM architecture, a write to some location  $l \in Loc$  first of all only modifies  $vmem(l)$ . Volatile memory is then occasionally *flushed* to persistent memory by the system. This updates the value of persistent memory to the value currently in volatile memory for location  $l$ . The programmer can also enforce such a flush to location  $l$  by executing `flush(l)`<sup>2</sup>, which is modelled by an update that sets  $pmem(l)$  to  $vmem(l)$ .

When a crash occurs, the contents of volatile memory is lost and that of persistent memory is kept. We assume that immediately after a crash  $vmem$  is (re)set to  $pmem$ , thus any writes to  $vmem$  that have not been flushed will be lost.

The implementations of concurrent data structures or STM algorithms have to ensure that shared memory is kept in a consistent state, despite these losses. To this end, they need to persist enough data (i.e., flush it) to be able to bring shared memory back to a consistent state after crashes. For our implementations, we assume that such a *recovery* step is automatically executed by the algorithms after every crash. In our models of the algorithms, we formalise this by a single atomic operation *crashRecovery*. Note that this is not a strict requirement of durable opacity, i.e., durable opacity (like durable linearizability) admits other algorithms in which the crash and recovery occur as two separate steps.

The execution with persistent and volatile memory applies to actual *implementations*, i.e., the low-level descriptions of STM algorithms with all the implementation details filled in. Implementations are one conceptual entity within our reasoning technique based on *refinement* [DB14]. Refinement compares abstract *specifications* to concrete implementations. The purpose of an abstract specification is to fix the allowed execution traces. Abstract

<sup>2</sup>We use typewriter font to refer to program code.

invocations	possible matching responses
$inv_t(\text{TMBegin})$	$res_t(\text{TMBegin(ok)}), res_t(\text{TMBegin(abort)})$
$inv_t(\text{TMCommit})$	$res_t(\text{TMCommit(commit)}), res_t(\text{TMCommit(abort)})$
$inv_t(\text{TMRead}(x))$	$res_t(\text{TMRead}(v)), res_t(\text{TMRead(abort)})$
$inv_t(\text{TMWrite}(x, v))$	$res_t(\text{TMWrite(ok)}), res_t(\text{TMWrite(abort)})$

Table 1: Events appearing in transactional histories, where  $t \in T$  is a transaction identifier,  $x \in Loc$  is a location, and  $v \in Val$  a value

specifications are hence not subject to specific forms of execution with volatile and persistent memory; they are allowed to (and should) abstract from implementation details. Thus, we often develop intermediate models that interact directly to NVM (bypassing volatile memory), with more realistic interactions between volatile and persistent memory only appearing in the final implementation (see Figure 11).

**2.2. Histories.** Both correctness conditions are formalised in terms of a *history*, which is a sequence of *events*. An event is either (1) an invocation (*inv*) or (2) a response (*res*) of an operation *op* out of a set of operations  $\Sigma$  or (3) a system-wide crash event *c*. Like durable linearizability, although crash events appear in the history, separate recovery operations do not explicitly appear in the histories. Invocation and response events of the same operation are said to *match*. Events are furthermore parameterised by thread or transaction identifiers from a set  $T$ . For simplicity, we do not distinguish between threads and transactions here. Invocation events may have input parameters and response events output parameters. We use the following notation on histories: for a history  $h$ ,  $h \upharpoonright t$  is the projection onto the events of transaction or thread  $t$  only, and  $h[i..j]$  the subsequence of  $h$  from  $h(i)$  to  $h(j)$  inclusive. We write  $hh'$  for the concatenation of two histories  $h$  and  $h'$ . We say that two histories  $h$  and  $h'$  are *equivalent*, denoted  $h \equiv h'$ , if  $h \upharpoonright t = h' \upharpoonright t$  for all  $t \in T$ . For a response event  $e$ , we let  $rval(e)$  denote the value returned by  $e$ . If  $e$  is not a response event, then we let  $rval(e) = \perp$ . We furthermore let *Res* be the set of all response and *Inv* the set of all invocation events.

We consider two types of histories, transactional and non-transactional histories. A transactional history only contains the invocation and response events of Table 1. STM algorithms allow for a concurrent access to shared memory (a set of locations *Loc*). Every transaction consists of an operation `TMBegin` followed by a number of operations `TMWrite` or `TMRead` and finally an operation `TMCommit`. All of these operations may also return `abort` meaning that the operation has not succeeded. We say that a transaction  $t$  is *committed* in a history  $h$  if  $res_t(\text{TMCommit(commit)})$  is contained in  $h$ . In non-transactional histories we only have invocations and responses of operations on an object (e.g., a data structure), where invocations appear before their corresponding responses, and there is no grouping of operations into transactions.

A (non-transactional) history is *sequential* if every invocation event (except for possibly the last event) is directly followed by its matching response. A transactional history is *transaction sequential* if it is sequential and there are no overlapping transactions. A history is *complete* if there are no pending operations, i.e., no invocations without a matching return. The function *complete* removes all pending operations from a history. A history is *well-formed* if  $h \upharpoonright t$  is sequential for every  $t \in T$ . A well-formed transactional history is furthermore *transaction well-formed* if for every  $t$ ,  $h \upharpoonright t = \langle e_0, \dots, e_m \rangle$  is a sequential

history such that  $e_0 = \text{inv}_t(\text{TMBegin})$ , and for all  $0 < i \leq m$ , event  $e_i \neq \text{inv}_t(\text{TMBegin})$  and for all  $0 < i < m$ ,  $\text{rval}(e_i) \notin \{\text{commit}, \text{abort}\}$ . This, in particular, implies that transaction identifiers cannot be re-used.

For a history  $h$  and events  $e_1, e_2$ <sup>3</sup>, we write (1)  $e_1 <_h e_2$  whenever  $h = h_0 e_1 h_1 e_2 h_2$ , and (2)  $e_1 \ll_h e_2$  if  $e_1 <_h e_2$  and  $e_1 \in \text{Res}, e_2 \in \text{Inv}$  (real-time order of operations). In a transactional history  $h$ , we furthermore write  $t_1 \prec_h t_2$  if the commit operation of transaction  $t_1$  completes before transaction  $t_2$  starts (real-time order of transactions).

**2.3. Linearizability and Durable Linearizability.** Linearizability of concurrent data structures is defined by comparing the (possibly concurrent) histories arising in usages of the data structure to sequential *legal* histories. Legality is defined by specifying sequential objects  $\mathbb{S}$ , i.e., sequential versions of a data structure (see Definition 3.3). These sequential versions define the “correct” behaviour, e.g. a queue data structure adhering to a FIFO protocol or not losing elements. For now, in the formal definition of linearizability, we simply assume that we are given the set of sequential legal histories  $H_{\mathbb{S}}$  (as generated by a sequential object). Later, we will give an abstract data type in the form of an IOA [LT87] as a specification of a sequential object.

A concurrent data structure is *linearizable* if all of its histories arising in usages of the data structure are linearizable.

**Definition 2.1** (Linearizability [HW90]). *A (concurrent) history  $h$  is linearizable (w.r.t. some set of sequential histories  $H_{\mathbb{S}}$ ) iff there exists some  $h_0 \in \text{Res}^*$  (completing some of the pending operations) such that for  $h' = \text{complete}(hh_0)$  there exists some  $h_s \in H_{\mathbb{S}}$  such that*

**L1** :  $h' \equiv h_s$  and

**L2** :  $e \ll_{h'} e'$  implies  $e \ll_{h_s} e'$ .

In this definition, we assume the history does not contain any crash events. Linearizability only considers executions of data structures without intervening system crashes.

For durable linearizability, we need to consider histories with crash events. Given a history  $h$ , we let  $\text{ops}(h)$  denote  $h$  restricted to non-crash events. The crash events partition a history into  $h = h_0 c_1 h_1 c_2 \dots h_{n-1} c_n h_n$ , such that  $n$  is the number of crash events in  $h$ ,  $c_i$  is the  $i$ th crash event and  $\text{ops}(h_i) = h_i$  (i.e.,  $h_i$  contains no crash events). We call the subhistory  $h_i$  the  $i$ -th era of  $h$ . For well-formedness of histories we now also require every thread identifier to appear in at most one era.

These definitions allow us to lift linearizability to durable linearizability.

**Definition 2.2** (Durable Linearizability [IMS16]). *A history  $h$  is durably linearizable iff it is well formed and  $\text{ops}(h)$  is linearizable.*

Durable linearizability will later be used to establish correctness of a library implementation that provides synchronised access to shared memory in the presence of NVM.

---

<sup>3</sup>Following Herlihy and Wing [HW90], we assume events are unique in a history by equipping them with a unique tag. For simplicity, these details are elided in our formalisation.

**2.4. Opacity and Durable Opacity.** Opacity [GK10, GK08] compares concurrent histories generated by an STM implementation to sequential histories. The difference to linearizability is that we need to (a) consider entire transactions and (b) deal with aborted transactions. The correctness criterion opacity guarantees that values written by aborted transactions (i.e., transactions with events with `abort` as response value) cannot be read by other transactions.

For opacity, we again compare concurrent histories against a set of legal sequential ones, but now we employ transaction sequential histories. Again, we assume the set of legal sequential transactional histories  $TH_{\mathbb{S}}$  to be given, and out of these define the *valid* ones.

**Definition 2.3** (Valid History). *Let  $hs$  be a sequential history and  $i$  an index of  $hs$ . Let  $hs'$  be the projection of  $hs[0..(i-1)]$  onto all events of committed transactions plus the events of the transaction to which  $hs(i)$  belongs. Then we say  $hs$  is valid at  $i$  whenever  $hs'$  is legal. We say  $hs$  is valid iff it is valid at each index  $i$ .*

We let  $VH_{\mathbb{S}}$  be the set of transaction sequential valid histories. With this at hand, we can define opacity similar to linearizability.

**Definition 2.4** (Opacity [GK08, GK10]). *A (concurrent) history  $h$  is end-to-end opaque iff there exists some  $h_0 \in Res^*$  (completing some of the pending operations) such that for  $h' = complete(hh_0)$  there exists some  $h_s \in VH_{\mathbb{S}}$  such that*

**O1 :**  $h' \equiv h_s$ , and

**O2 :**  $t_1 \prec_h t_2$  implies  $t_1 \prec_{h_s} t_2$ .

*A history  $h$  is opaque iff each prefix  $h'$  of  $h$  is end-to-end opaque.*

An STM algorithm itself is *opaque* iff its set of histories occurring during executions of the STM is opaque. For durable opacity, we simply lift this definition to histories with crashes.

Like durable linearizability, the purpose of durable opacity is to ensure that histories with crashes leave the shared state in a consistent state as defined by opacity. This means that any live transaction that has not yet started its commit operation will be treated as an aborting transaction. If a transaction has started its commit, then the commit could be completed by either a successful commit or an abort. For well-formedness of TM histories, we now also require every *transaction* identifier to appear in at most one era. This means that no transaction survives a crash.

**Definition 2.5** (Durable Opacity [BDD<sup>+</sup>20]). *A history  $h$  is durably opaque iff it is transaction well-formed and  $ops(h)$  is opaque.*

One of the guarantees of durable opacity (again like durable linearizability) is that it ensures every committed transaction is persisted. Thus, if a transaction's effects are globally visible, then this transaction is guaranteed to also survive any subsequent crashes. In other words, durable opacity ensures that for committed transactions, the *visibility order* (the order in which transactions are seen by other transactions) and the *persistent order* (the order in which transactions become durable) coincide.

Furthermore, durable opacity aims to transfer the atomicity property of opacity to the NVM setting. For opacity, this property has been shown via a study of a specification called TMS1 [DGLM13]. It is well known that TMS1 is both necessary and sufficient to ensure transactions are atomic [AGHR14]. Opacity is known to be stronger than TMS1 [LLM12b], thus also guarantees the sufficiency property. Durable opacity ensures transactional atomicity

even in the presence of crashes, thus ensures the same guarantees. However, the precise formulation of the atomicity problem in the setting of NVM deserves further study.

In this paper, we aim to develop a method for proving durable opacity of STM algorithms. For the proof, we develop a modular proof technique, which requires us to show durable linearizability of some library data structure providing access to shared memory.

### 3. USING IOA TO PROVE DURABLE OPACITY

Previous works [ADD17, DDD<sup>+</sup>16, DD15, AD17] have considered proofs of opacity using the operational TMS2 specification [DGLM13], which has been shown to guarantee opacity [LLM12b]. The proofs show refinement of the implementation against the TMS2 specification using either forward or backward simulation. In this, both implementation and specification are given as Input/Output automata (IOA) to enable use of a standard simulation-based proof technique. For durable opacity, we follow a similar strategy. We develop the dTMS2 operational specification, a durable version of the TMS2 specification, that we prove satisfies durable opacity. By proving a simulation relation to hold between a (durable) STM implementation and dTMS2 we can establish durable opacity of an STM.

In the following, we will first of all shortly explain IOA and simulations in general and thereafter develop dTMS2.

**3.1. IOA, Refinement and Simulation.** We use Input/Output Automata (IOA) [LT87] to model both STM implementations and the specification, dTMS2.

**Definition 3.1** (Input/Output Automaton (IOA)). *An Input/Output Automaton (IOA) is a labeled transition system  $A$  with a set of states  $states(A)$ , a set of actions  $acts(A)$ , a set of start states  $start(A) \subseteq states(A)$ , and a transition relation  $trans(A) \subseteq states(A) \times acts(A) \times states(A)$  (so that the actions label the transitions).*

The set  $acts(A)$  is partitioned into input actions  $input(A)$ , output actions  $output(A)$  and internal actions  $internal(A)$ . The internal actions represent events of the system that are not visible to the external environment. The input and output actions are externally visible, representing the IOA's interactions with its environment. Thus, we define the set of *external actions*,  $external(A) = input(A) \cup output(A)$ . We write  $s \xrightarrow{a}_A s'$  iff  $(s, a, s') \in trans(A)$ .

An *execution* of an IOA  $A$  is a sequence  $\sigma = s_0 a_0 s_1 a_1 s_2 \dots s_n a_n s_{n+1}$  of alternating states and actions, such that  $s_0 \in start(A)$  and for all states  $s_i$ ,  $s_i \xrightarrow{a_i}_A s_{i+1}$ . We write  $exec(A)$  for the set of all executions of  $A$  and  $first(\sigma) = s_0$  for the initial state of an execution  $\sigma$ . Whenever we have several IOAs, we use indices to distinguish between them, e.g.  $\sigma_A$  is used to denote an execution of  $A$ .

A *reachable* state of  $A$  is a state appearing in an execution of  $A$ . We let  $reach(A)$  denote the set of all reachable states of  $A$ . An *invariant* of  $A$  is any superset of the reachable states of  $A$  (equivalently, any predicate satisfied by all reachable states of  $A$ ). A *trace* of  $A$  is any sequence of (external) actions obtained by projecting the external actions of any execution of  $A$ . The set of traces of  $A$ , denoted  $traces(A)$ , represents  $A$ 's externally visible behaviour.

For IOA  $C$  and  $A$ , we say that  $C$  is a *refinement* of  $A$ , denoted  $C \leq A$ , iff  $traces(C) \subseteq traces(A)$ . Note that refinement is transitive. We typically show that  $C$  is a refinement of  $A$  by proving the existence of a *forward simulation*, which enables one to check step correspondence between the transitions of  $C$  and those of  $A$ . The definition of forward simulation we use is adapted from that of Lynch and Vaandrager [LV95].

**Definition 3.2** (Forward Simulation). *A forward simulation from a concrete IOA  $C$  to an abstract IOA  $A$  is a relation  $R \subseteq \text{states}(C) \times \text{states}(A)$  such that each of the following holds.*

Initialisation.  $\forall cs \in \text{start}(C). \exists as \in \text{start}(A). R(cs, as)$

External step correspondence.

$$\forall cs \in \text{reach}(C), as \in \text{reach}(A), a \in \text{external}(C), cs' \in \text{states}(C). \\ R(cs, as) \wedge cs \xrightarrow{a}_C cs' \Rightarrow \exists as' \in \text{states}(A). R(cs', as') \wedge as \xrightarrow{a}_A as'$$

Internal step correspondence.

$$\forall cs \in \text{reach}(C), as \in \text{reach}(A), a \in \text{internal}(C), cs' \in \text{states}(C). \\ R(cs, as) \wedge cs \xrightarrow{a}_C cs' \Rightarrow \\ R(cs', as) \vee \exists a' \in \text{internal}(A), as' \in \text{states}(A). R(cs', as') \wedge as \xrightarrow{a'}_A as'$$

Forward simulation is *sound* in the sense that if there is a forward simulation between  $A$  and  $C$ , then  $C$  refines  $A$  [LV95, Mül98].

**3.2. Canonical IOA for (durable) linearizability.** To prove linearizability the relevant set of sequential histories  $H_S$  are given as the histories of a sequential object  $\mathbb{S}$ , that defines a set atomic operations  $op_i$  that receive input, modify a state and return output.

**Definition 3.3** (Sequential Object). *A sequential object  $\mathbb{S}$  is a 4-tuple  $(\Sigma, Val, State, Init)$  where*

- *State is a set of states,  $Init \subseteq State$  is a set of initial states,*
- *Val is a set of values used as input and output,*
- *$\Sigma$  is a set of atomic operations  $op_i$  for some  $i \in I$ .*  
*Each operation is specified as a relation  $op_i \subseteq Val \times State \times State \times Val$ .*

Some operations may have no inputs/outputs and others may have several. This can be accommodated by including tuples including the empty tuple  $\epsilon$  in  $Val$ . We drop an empty input or output when writing an event. A sequential history of  $\mathbb{S}$  has the form

$$\text{inv}(op_{k_1}(in_1)), \text{res}(op_{k_1}(out_1)), \dots, \text{inv}(op_{k_n}(in_n)), \text{res}(op_{k_n}(out_n))$$

The history is a legal sequential history in  $H_S$ , iff there is a sequence  $s_0 \dots s_n$  of states, such that  $s_0 \in Init$  and  $(in_m, s_m, s_{m+1}, out_m) \in op_{k_m}$  for all  $m < n$ .

To prove durable linearizability of a concurrent implementation we will specify the concurrent program as an IOA  $C$  that generates a set of concurrent histories. Note that—as to mimic execution of an NVM architecture—this implementation IOA  $C$  would need to explicitly model persistent and volatile memory as well as its flushing discipline, i.e., when the implementation wants an update to a location to reach persistent memory. To prove that  $C$  is durably linearizable to  $H_S$ , it is then sufficient to prove that  $C$  refines the *canonical durable IOA*  $\text{DURAUT}(\mathbb{S})$  shown in Fig. 2 (see [DDD<sup>+</sup>19]). This IOA serves as an abstract specification of durable linearizability in the refinement proof: its traces are exactly the durably linearizable histories (of some sequential object).

The state of this IOA incorporates the state  $s$  of the sequential object  $\mathbb{S}$  and adds a program counter  $pc_t$  for every transaction  $t \in T$ . The possible values of this program counter include *notStarted*, *ready* and *crashed* to indicate that the transaction has not started (its initial value), is running but not currently executing an operation, or has crashed. The execution of an operation  $op$  is split into three steps: an invocation and a response of the

$inv_t(op(in))$ Pre: $pc_t = ready$ Eff: $pc_t := doOp(in)$	$do_t(op)$ Pre: $pc_t = doOp(in)$ Eff: $(s, out) := SOME(s', out').$ $op(in, s, s', out')$ $pc_t := resOp(out)$	$res_t(op(out))$ Pre: $pc_t = resOp(out)$ Eff: $pc_t := ready$
$run_t$ Pre: $pc_t = notStarted$ Eff: $pc_t := ready$	$crash$ Pre: $true$ Eff: $pc := \lambda t : T.$ <b>if</b> $pc_t \neq notStarted$ <b>then</b> crashed <b>else</b> $pc_t$	

Figure 2: Durable IOA DURAUT( $\mathbb{S}$ )

operation plus a *do*-step (where the actual effect of the operation takes place). Note that both *run* and *do* are internal actions and thus do not appear in the traces of the IOA.

- First, when  $pc_t = ready$  an invoke step with action  $inv_t(op(in))$  is executed. The input value of this step is arbitrary and gets stored in  $pc_t$  by setting it to  $doOp(in)$ .
- Second, a step with internal action  $do_t(op)$  is executed. This step will correspond to the linearization point of an implementation. The step modifies the state of  $op$  by choosing a new state and an output according to the specification of  $op$  (the step is not possible if there is no  $s', out'$  with  $op(in, s, s', out')$ ). The computed output is again stored in  $pc_t$  by setting it to  $resOp(out)$ .
- Finally, a response step, that returns the  $out$  value that was stored in  $pc_t$  by emitting an action  $resOp(out)$ . This step finishes the execution of  $op$  by setting  $pc_t$  to *ready*.

The durable canonical IOA (more details are in [DDD<sup>+</sup>19]) is an extension of the canonical IOA from [Lyn96] for linearizability to accommodate durable linearizability. It is the most general specification of concurrent runs that still allows us to construct an equivalent sequential history: the sequential history can be constructed as a sequence of invoke-response pairs from the sequence of executed  $do_t(op)$  steps. The IOA guarantees that this sequential history is obviously in  $H_{\mathbb{S}}$ .

The following theorem establishes a correspondence between the durable IOA and durable linearizability. For a sequential object  $\mathbb{S}$ , we let DURLIN( $\mathbb{S}$ ) be the set of histories that are durably linearizable with respect to  $\mathbb{S}$ .

**Theorem 3.4** [DDD<sup>+</sup>21]. *Let  $\mathbb{S}$  be a sequential object. Then  $traces(DURAUT(\mathbb{S})) = DURLIN(\mathbb{S})$ .*

As the durable IOA has durably linearizable histories only, it can serve as an abstract specification in a proof of durable linearizability via refinement.

**Lemma 3.5.** *Let  $C$  be an implementation IOA. If  $C$  refines DURAUT( $\mathbb{S}$ ), then  $C$  is durably linearizable to  $\mathbb{S}$ .*

Summarising, this gives us the following: Whenever we have an algorithm  $Alg$  which runs on an NVM architecture and the implementation IOA  $C$  models the executions of this algorithm on NVM (i.e., adequately represents persistent and volatile memory) and  $C$  refines DURAUT( $\mathbb{S}$ ), then the algorithm  $Alg$  is durably linearizable.

**3.3. Refinement in context.** For our modular proof technique, we will let an STM algorithm call a library in order to manage access to shared state. As both STM and library will be formalised in terms of an IOA, we need some notion of a *context* IOA (i.e., the STM IOA) using a *library IOA*. To this end, we employ the following definition of product IOA [Lyn96], which requires synchronisation of two IOA on shared external actions.

**Definition 3.6** (Product IOA). *Let  $A, B$  be two IOA with no shared internal actions. Then the product IOA  $A \times B$  is defined to have*

- $states(A \times B) = states(A) \times states(B)$ ,
- $start(A \times B) = start(A) \times start(B)$ ,
- $acts(A \times B) = acts(A) \cup acts(B)$ ,
- $(as, bs) \xrightarrow{a}_{A \times B} (as', bs')$  iff the following two properties hold:
  - if  $a \in actions(A)$ , then  $as \xrightarrow{a}_A as'$ , else  $as' = as$ ;
  - if  $a \in actions(B)$ , then  $bs \xrightarrow{a}_B bs'$ , else  $bs' = bs$ .

In the following we will use this product construction in an asymmetric way: the shared external actions of  $A$  (the library) and  $B[\cdot]$  (the STM algorithm) are the invocations and responses of library calls, and the library is required to have no further external actions. In such a setting, we write  $B[A]$  for the product of  $A$  and  $B[\cdot]$  (i.e., where IOA  $B[\cdot]$  uses library IOA  $A$ ).

Later we will develop two versions of the library which provides access to shared memory: one with and one without volatile memory. These two versions are shown to be a refinement of each other (more precisely, one version is shown to be durably linearizable w.r.t. the other), and we need to lift this result to STMs using the libraries. It is folklore knowledge that refinement of an abstract object by a concrete object implies refinement between an algorithm (a context) using the abstract object and the same algorithm using the concrete one. Theorems stating such a property have been proven in many settings, e.g. for data refinement in [dRE98]. The fundamental paper on linearizability [HW90] implicitly uses such a result when it assumes that the individual steps of algorithms are linearizable operations as well. We could however not find a formal proof of refinement in context for IOA, and thus both state and prove it in this setting.

**Theorem 3.7.** *If  $C \leq A$ , then  $B[C] \leq B[A]$ .*

Note that  $C \leq A$  and  $B[C] \leq B[A]$  implies that the external actions of  $C$  and  $A$  are the same and that they are a subset of external actions of  $B$ .

To prove refinement, we need to construct an execution  $\sigma_{B[A]}$  of  $B[A]$  with the same trace (i.e., the same external events) when given an execution  $\sigma_{B[C]}$  of  $B[C]$ . The idea is shown in Fig. 3 with an execution that executes three events  $a_1, a_2, a_3$ . The execution of  $B[C]$  contains an execution  $\sigma_C$  of  $C$  by projecting to the states of  $C$  and removing all steps (here:  $a_1$ ) where  $C$  is not involved. This execution can be split into finite segments of internal  $C$ -steps that each end with an external shared action (with possibly a final sequence of internal  $C$ -steps that is not used). In the example, there is one segment consisting of one internal action  $a_2$ , ending with the external action  $a_3$ . By refinement, there exists an execution  $\sigma_A$  of  $A$  with the same external actions. This execution can be split in the same way: in the example the new segment consists of internal actions  $\alpha$ , and ends with  $a_3$ . Now delete the internal  $C$ -steps from the combined execution of  $B[C]$ , and replace the  $C$ -step in each combined step of  $C$  and  $B$  with the corresponding  $A$ -step from the abstract execution. Add the sequence of internal  $A$ -steps (here:  $\alpha$ ) that leads to this step (here:  $a_3$ ) right before

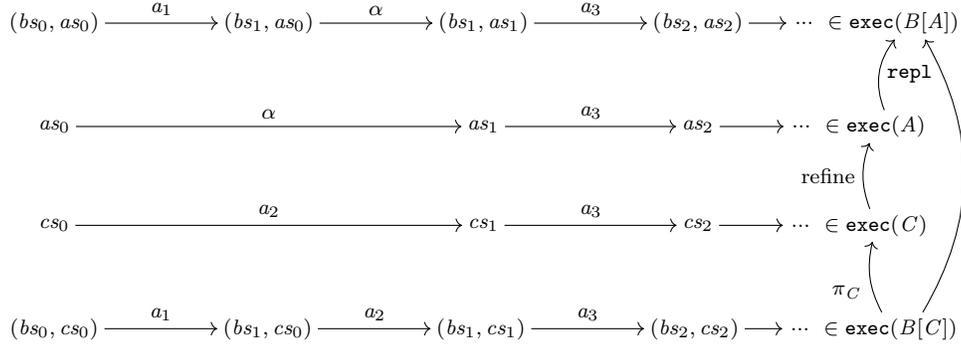


Figure 3: Construction of an execution of  $B[A]$  from an execution of  $B[C]$ .  $a_1 \in \text{acts}(B) \setminus \text{external}(C)$ ,  $a_2 \in \text{internal}(C)$ ,  $a_3 \in \text{external}(C)$  ( $= \text{external}(A)$ ,  $\subseteq \text{external}(B)$ ),  $\alpha \in \text{internal}(A)^*$ .

the combined step in the combined execution. The result is an execution  $\sigma_{B[A]}$  of  $B[A]$  which has the same trace as the original execution. Formally, the two steps are done by a projection function  $\pi_C$  and a **repl** function.

*Proof of Theorem 3.7:* To show that  $\text{traces}(B[C]) \subseteq \text{traces}(B[A])$  choose an arbitrary trace from  $\text{traces}(B[C])$ . For this trace an execution  $\sigma_{B[C]}$  of the form  $(bs_0, cs_0) a_0 (bs_1, cs_1) \dots a_n (bs_{n+1}, cs_{n+1})$  must exist. For such an execution the projection  $\pi_C : \text{exec}(B[C]) \rightarrow \text{exec}(C)$  to an execution of C can be defined recursively over its length  $n$ .

$$\begin{aligned} \pi_C((bs_0, cs_0)) &= cs_0 \\ \pi_C((bs_0, cs_0) a_0 \sigma'_{B[C]}) &= cs_0 a_0 \pi_C(\sigma'_{B[C]}) \text{ when } a_0 \in \text{acts}(C) \\ \pi_C((bs_0, cs_0) a_0 \sigma'_{B[C]}) &= \pi_C(\sigma'_{B[C]}) \text{ when } a_0 \in \text{acts}(B) \setminus \text{external}(C) \end{aligned}$$

In the second and third line  $\sigma'_{B[C]}$  is the rest of trace (of length  $n$ ) with the first state and action removed. The actions of  $\pi_C(\sigma_{B[C]})$  are those of  $\sigma_{B[C]}$  which are in  $\text{acts}(C)$ . Refinement then guarantees the existence of an execution  $\sigma_A \in \text{exec}(A)$  with  $\text{trace}(\pi_C(\sigma_{B[C]})) = \text{trace}(\sigma_A)$ .

This allows to define a function  $\text{repl} : \text{exec}(B[C]) \times \text{exec}(A) \rightarrow \text{exec}(B[A])$ . The result of  $\text{repl}(\sigma_{B[C]}, \sigma_A)$  is defined when  $\text{trace}(\pi_C(\sigma_{B[C]})) = \text{trace}(\sigma_A)$ . It replaces the steps of C in  $\sigma_{B[C]}$  with the corresponding ones in  $\sigma_A$ . Again, **repl** is defined recursively. For  $n = 0$  we simply have

$$\text{repl}((bs_0, cs_0), \sigma_A) = (bs_0, \text{first}(\sigma_A)) \quad (3.1)$$

When  $n > 0$  there are two cases: When  $a_0 \in \text{Acts}(B) \setminus \text{external}(C)$  then

$$\text{repl}((bs_0, cs_0) a_0 \sigma'_{B[C]}, \sigma_A) = (bs_0, \text{first}(\sigma_A)) a_0 \text{repl}(\sigma'_{B[C]}, \sigma_A) \quad (3.2)$$

Note that the first component of the pair  $\text{first}(\sigma'_{B[C]})$  is  $bs_0$  in this case, and that  $\text{trace}(\pi_C(\sigma'_{B[C]})) = \text{trace}(\sigma_A)$  is still true for the recursive call. Otherwise, when  $a_0 \in \text{external}(C)$  then  $a_0$  is in  $\text{external}(A)$  and  $\text{external}(B)$  as well, since the external actions of C and A are the same and shared with B. The trace of  $\pi_C(\sigma_{B[C]})$  then contains  $a_0$  as its first external action. By trace equality, the first external action of  $\sigma_A$  must be  $a_0$  as well.  $\sigma_A$  therefore has the form  $\sigma_A = as_0 a'_1 as_1 a'_2 \dots a'_m as_m a_0 \sigma'_A$  where  $m \geq 0$ , and  $a'_1 \dots a'_m \in \text{internal}(A)^*$  is a sequence of internal actions. The sequence  $\alpha$  in the example

of Fig. 3 is this sequence of actions. The resulting execution now first executes the internal steps, and finally the combined step, so  $\mathbf{repl}$  is defined in this case as

$$\begin{aligned} \mathbf{repl}((bs_0, cs_0) a_0 \sigma'_{B[C]}, \sigma_A) = \\ (bs_0, as_0) a'_1 (bs_0, as_1) \dots a'_m (bs_0, as_m) a_0 \mathbf{repl}(\sigma'_{B[C]}, \sigma'_A) \end{aligned} \quad (3.3)$$

Again,  $\mathit{trace}(\pi_C(\sigma'_{B[C]})) = \mathit{trace}(\sigma'_A)$  is still true for the recursive call. It is now easy to check inductively that  $\mathbf{repl}(\sigma_{B[C]}, \sigma_A)$  returns an execution of  $B[A]$ , since all steps of the constructed execution are steps of  $B[A]$ . The first step of (3.2) is a step of  $B$  with an unshared action of  $B$  that does not change the state of  $A$ , so it is a step of the product  $B[A]$  (last clause of Definition 3.6). The first  $m$  steps of (3.3) are steps with unshared internal actions of  $A$  that do not change the state of  $B$ , so they are steps of  $B[A]$  too. Finally, step  $m+1$  of (3.3) executes shared action  $a_0$  and changes both states according to the definition of  $A$  and  $B$ , so it is a step of  $B[A]$  too.

When  $\mathbf{first}(\sigma_{B[C]}) = (bs_0, cs_0)$ , then the first state of  $\mathbf{repl}(\sigma_{B[C]}, \sigma_A)$  is  $(bs_0, \mathbf{first}(\sigma_A))$ , which is initial, if the first states of  $\sigma_{B[C]}$  and  $\sigma_A$  are. The result  $\mathbf{repl}(\sigma_{B[C]}, \sigma_A)$  has the same trace as  $\sigma_{B[C]}$  since all external actions are preserved. This implies that the original  $\mathit{trace}(\sigma_{B[C]})$  the construction started with is also a trace of  $B[A]$ , finishing the proof.  $\square$

**Remark:** Although we do not need this generalisation here, the result holds as well if refinement is defined as trace inclusion for finite *as well as infinite* traces (see [LV95]). Both  $\pi_C$  and  $\mathbf{repl}$  are prefix-monotone, so the result of applying the functions to infinite traces can be defined as the limit of applying them to finite prefixes. The proof for this extended scenario has been formalised in KIV [BDD<sup>+</sup>21].

**3.4. IOA for dTMS2.** In this section, we describe the dTMS2 specification, an operational model that ensures durable opacity, which is based on TMS2 [DGLM13]. TMS2 itself has been shown to imply opacity [LLM12b], and hence has been widely used as an intermediate abstract specification in the verification of transactional memory implementations [ADD17, DDS<sup>+</sup>15, AD17, DDD<sup>+</sup>16]. dTMS2 is thus designed to play the rôle of an abstract specification for refinement proofs of durable opacity like  $\mathbf{DURAUT}(S)$  is for proofs of durable linearizability.

In the following, we let  $f \oplus g$  denote functional override of  $f$  by  $g$ , where we define  $f \oplus g = \lambda k \in \text{dom}(f). \mathbf{if } k \in \text{dom}(g) \mathbf{then } g(k) \mathbf{else } f(k)$ .

Formally, dTMS2 is specified by the IOA in Figure 4, which describes the required ordering constraints, memory semantics and prefix properties. Recall that we assume a set  $Loc$  of locations and a set  $Val$  of values. Thus, a memory is modelled by a function of type  $Loc \rightarrow Val$ . A key feature of dTMS2 (like TMS2) is that it keeps track of a *sequence* of memory states, one for each committed writing transaction. This makes it simpler to determine whether reads are consistent with previously committed write operations. Each committing transaction containing at least one write adds a new memory version to the end of the memory sequence. Note that dTMS2 is an IOA used for abstract specification in a refinement proof only; it is not an implementation that has to keep track of persistent and volatile memory.

The state space of dTMS2 has several components. The first,  $mems$  is a nonempty sequence of *memory* states, which initially contains one state. The original specification of TMS2 is parameterised by some initialisation predicate describing this initial memory state,

**State variables:**

$mems : seq(Loc \rightarrow Val)$ , initially satisfying  $\text{dom } mems = \{0\}$

$pc_t : PCVal$ , for each  $t \in T$ , initially  $pc_t = \text{notStarted}$  for all  $t \in T$

$beginIdx_t : \mathbb{N}$  for each  $t \in T$ , unconstrained initially

$rdSet_t : Loc \rightarrow Val$ , initially empty for all  $t \in T$ , where  $\rightarrow$  denotes a partial function

$wrSet_t : Loc \rightarrow Val$ , initially empty for all  $t \in T$

**Transition relation:**

$inv_t(\text{TMBegin})$	$res_t(\text{TMBegin(ok)})$
Pre: $pc_t = \text{notStarted}$	Pre: $pc_t = \text{beginPending}$
Eff: $pc_t := \text{beginPending}$	Eff: $pc_t := \text{ready}$
$beginIdx_t := \text{len}(mems) - 1$	
$inv_t(\text{TMRead}(l))$	$res_t(\text{TMRead}(v))$
Pre: $pc_t = \text{ready}$	Pre: $pc_t = \text{resRead}(v)$
Eff: $pc_t := \text{doRead}(l)$	Eff: $pc_t := \text{ready}$
$inv_t(\text{TMWrite}(l, v))$	$res_t(\text{TMWrite(ok)})$
Pre: $pc_t = \text{ready}$	Pre: $pc_t = \text{resWrite}$
Eff: $pc_t := \text{doWrite}(l, v)$	Eff: $pc_t := \text{ready}$
$inv_t(\text{TMCCommit})$	$res_t(\text{TMCCommit(commit)})$
Pre: $pc_t = \text{ready}$	Pre: $pc_t = \text{resCommit}$
Eff: $pc_t := \text{doCommit}$	Eff: $pc_t := \text{committed}$
$res_t(op(\text{abort}))$	$\text{DoWrite}_t(l, v)$
Pre: $pc_t \notin \{\text{notStarted}, \text{ready},$	Pre: $pc_t = \text{doWrite}(l, v)$
$\text{resCommit}, \text{committed}, \text{aborted}\}$	Eff: $pc_t := \text{resWrite}$
Eff: $pc_t := \text{aborted}$	$wrSet_t := wrSet_t \oplus \{l \rightarrow v\}$
$\text{DoCommitReadOnly}_t(n)$	$\text{DoCommitWriter}_t$
Pre: $pc_t = \text{doCommit}$	Pre: $pc_t = \text{doCommit}$
$\text{dom}(wrSet_t) = \emptyset$	$rdSet_t \subseteq \text{last}(mems)$
$\text{validIdx}(t, n)$	Eff: $pc_t := \text{resCommit}$
Eff: $pc_t := \text{resCommit}$	$mems := mems \wedge (\text{last}(mems) \oplus wrSet_t)$
$\text{DoRead}_t(l, n)$	$\text{crashRecovery}$
Pre: $pc_t = \text{doRead}(l)$	Pre: $\text{true}$
$l \in \text{dom}(wrSet_t) \vee \text{validIdx}(t, n)$	Eff: $pc := \lambda t : T.$
Eff: <b>if</b> $l \in \text{dom}(wrSet_t)$	<b>if</b> $pc_t \notin \{\text{notStarted}, \text{committed}\}$
<b>then</b> $pc_t := \text{resRead}(wrSet_t(l))$	<b>then</b> $\text{aborted}$
<b>else</b> $v := mems(n)(l)$	<b>else</b> $pc_t$
$pc_t := \text{resRead}(v)$	$mems := \langle \text{last}(mems) \rangle$
$rdSet_t := rdSet_t \oplus \{l \rightarrow v\}$	
<b>where</b> $PCEXternal \hat{=} \{\text{notStarted}, \text{ready}, \text{resCommit}, \text{resWrite}, \text{committed}, \text{aborted}\} \cup$	
$\{\text{resRead}(v) \mid v \in Val\}$	
$PCVal \hat{=} PCEXternal \cup \{\text{beginPending}, \text{doCommit}, \text{cancelPending}\}$	
$\cup \{\text{doRead}(l) \mid l \in L\} \cup \{\text{doWrite}(l, v) \mid l \in Loc, v \in Val\}$	
$\text{validIdx}(t, n) \hat{=} \text{beginIdx}_t \leq n < \text{len}(mems) \wedge rdSet_t \subseteq mems(n)$	
$op \in \{\text{TMBegin}, \text{TMRd}, \text{TMWr}, \text{TMCCommit}\}$	

Figure 4: The state space and transition relation of dTMS2, which extends TMS2 with a crash-recovery event

which we elide here for simplicity (and simply assume the implementation to employ the same initialisation). For each transaction  $t$  there is a program counter variable  $pc_t$ , which ranges over a set of *program counter values*, which are used to ensure that each transaction is well-formed, and to ensure that each transactional operation takes effect between its invocation and response. There is also a *begin index* variable  $beginIdx_t$ , that is set to the index of the most recent memory version when the transaction begins. This variable is critical to ensuring the real-time ordering property between transactions. Finally, there is a *read set*,  $rdSet_t$ , and a *write set*,  $wrSet_t$ , which record the values that the transaction has read and written during its execution, respectively.

The read set is used to determine whether the values that have been read by the transaction are consistent with the same version of memory (using  $validIdx$ ). The write set, on the other hand, is required because writes in DTMS2 are modelled using *deferred update* semantics: writes are recorded in the transaction’s write set, but are not published to any shared state until the transaction commits.

The *crashRecovery* action again models the effect of crashes and consecutive recoveries. It sets the program counter of every in-flight transaction to *aborted*, which prevents these transactions from performing any further actions in the era following the crash (for the generated history). Note that since transaction identifiers are not reused, the program counters of completed transactions need not be set to any special value (e.g., *crashed*) as with durable linearizability [DDD<sup>+</sup>19]. Moreover, after restarting, it must not be possible for any new transaction to interact with memory states prior to the crash. We therefore reset the memory sequence to be a singleton sequence containing the last memory state prior to the crash.

The external actions of DTMS2 are all invocation and response actions ( $inv_t$  and  $res_t$ ) plus the new *crashRecovery* action. The latter is the crash action  $c$  of histories. Note that the traces of DTMS2 hence take the form of histories.

The following theorem ensures that DTMS2 can be used as an intermediate specification in our proof method.

**Theorem 3.8.** *Each trace of DTMS2 is durably opaque.*

The proof of this theorem can be found in the appendix of [BDD<sup>+</sup>20]. Like durable linearizability, we have the following lemma, which allows us to establish durable opacity using refinement.

**Lemma 3.9.** *Let  $C$  be an implementation IOA. If  $C$  refines DTMS2, then  $C$  is durably opaque.*

#### 4. A MODULAR PROOF TECHNIQUE

In this section, we present a new approach to verifying durable opacity that allows one to leverage existing simulation-based proofs of opacity. An overview of the proof steps is shown in Figure 5. Given an existing opacity proof (step ①) that uses simulation against TMS2, the majority of the effort in the modularised proof method is the development of libraries AM and CM that handle memory operations and a proof of durable linearizability between the two. We exemplify this proof technique on a durable version of the STM algorithm NOREC which we newly develop below.

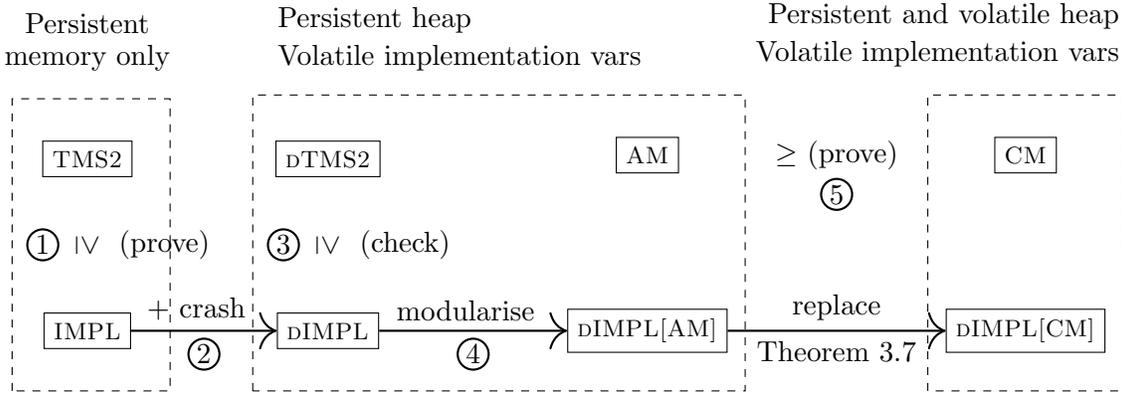


Figure 5: Overview of proof steps

**Overview of the proof technique.** The main idea behind the proof steps is to gradually introduce more complex (fine-grained) interactions between persistent and volatile memory, as outlined in Figure 5. We start with models (the implementation of the STM called IMPL and TMS2) in which all reads and writes interact directly with persistent memory, thus crashes have no effect.

Next, we introduce models DIMPL and dTMS2, where we assume memory is partitioned into two sorts: locations that the transactional memory implementations read from and write to (aka the *heap*), and variables that are used to implement the STM. E.g., for our running example (NOREC in Figure 6), variables such as  $glb$ ,  $loc_t$  and  $rdSet_t$  are implementation variables. In the models DIMPL and dTMS2, we assume that the operations on the heap are directly over persistent memory, whereas operations on implementation variables are over volatile memory. The former are assumed to be preserved upon a system crash, whereas the latter are lost. In our IOA models, the loss of implementation is modelled by setting the program counter of any running transactions to aborted when a crash occurs; since transactions are not restarted, this is equivalent to losing the local variables. Upon recovery, shared implementation variables must be reset since they are reused. For our example, the global counter  $glb$  is reset to 0 during recovery.

The next phase introduces DIMPL[AM], which uses a library AM that manages reads and writes to the heap. This model is a simple refactoring of DIMPL and hence has the same memory model: all reads and writes performed by AM are directly over persistent memory, whereas those performed by DIMPL[.] are on the transactional memory implementation and are hence volatile.

The final phase refines AM into CM such that CM is durably linearizable w.r.t. AM. Here, CM is a fine-grained implementation of AM, and hence we assume that it operates over both volatile and persistent memory.

We now describe the main steps of our proof method, as outlined in Figure 5 in more detail.

- ① Our modular proof method starts with an existing simulation-based proof between an STM implementation, IMPL, and TMS2, which establishes opacity of IMPL. In our example case study, we consider NOREC as our implementation, which has already been proven correct with respect to TMS2 [LLM12a]. Our primary motivation for the new

approach is to develop a durable version of an already opaque algorithm and avoid full re-verification of the durable version (against  $\text{dTMS2}$ ).

- ② To make sense of the adaptation of IMPL to persistent memory, we start by assuming IMPL runs directly on persistent memory, i.e., all memory accesses are persistent, and there is no notion of volatile memory. Hence, there are no flush operations that transfer memory contents from volatile to persistent memory. Of course, IMPL also does not contain a recovery operation since it has been designed to be opaque as opposed to durably opaque.

This step of the transformation therefore is to define a new version of IMPL,  $\text{dIMPL}$ , that extends IMPL with a “crash-recovery” operation. The purpose of this operation is to crash any live transactions so that they are no longer able to execute and to rollback the memory to a consistent state. We will require that the crash-recovery introduced into  $\text{dIMPL}$  is a refinement of the *crashRecovery* operation of  $\text{dTMS2}$  from Figure 4.

The introduction of a crash-recovery operation must be coupled with some small adjustments to the original algorithm. For instance, performing a write-back must be made atomic; a crash-recovery in the middle of a non-atomic write-back would leave the transactional memory heap in an inconsistent state. Details in the context of our running example are given in subsection 4.2.

- ③ The next step is to verify that the transformed algorithm  $\text{dIMPL}$  is durably opaque. To do this, we must adapt the existing simulation proof between IMPL and  $\text{TMS2}$  to prove simulation between  $\text{dIMPL}$  and  $\text{dTMS2}$ . Recall that the transformation from both IMPL to  $\text{dIMPL}$  and  $\text{TMS2}$  to  $\text{dTMS2}$  involves the introduction of a crash-recovery operation. Also recall that we assume both algorithms run directly on persistent memory, and no volatile memory is assumed in either case. Therefore, the effect of this crash-recovery operation in both cases is straightforward, and the adaptation of the proof is therefore straightforward too.
- ④ In the next step we adapt  $\text{dIMPL}$  so that it relegates all memory operations to an external library. We call this adapted algorithm  $\text{dIMPL}[\cdot]$ , which is  $\text{dIMPL}$  but with calls to external operations that manage memory interactions. We must additionally develop a data type (library IOA),  $\text{AM}$ , that handles these memory events. This enables one to define a composition  $\text{dIMPL}[\text{AM}]$  (via a product of two IOA), where the library used by  $\text{dIMPL}[\cdot]$  is  $\text{AM}$ .

There are two requirements for the library  $\text{AM}$ .

- (1) We must be able to show that the traces of  $\text{dIMPL}[\text{AM}]$  are a subset of the traces of  $\text{dIMPL}$ , i.e.,  $\text{dIMPL}[\text{AM}]$  is a *trace refinement* of  $\text{dIMPL}$ . Note that by transitivity of refinement this means that the composition  $\text{dIMPL}[\text{AM}]$  will be durably opaque.
- (2) We must allow  $\text{AM}$  to be *implemented* by a concrete memory library that uses both a volatile and persistent memory so that the algorithm can ultimately be implemented in a non-volatile memory architecture.

To satisfy both criteria while keeping the proof burden light, we make  $\text{AM}$  an *atomic* object, i.e., reads and write-backs are coarse-grained atomic operations (see Figure 9). Moreover,  $\text{AM}$  operates directly on persistent memory, i.e., it does not use volatile memory. With these restrictions in place, it becomes straightforward to show trace refinement between the traces of  $\text{dIMPL}$  and  $\text{dIMPL}[\text{AM}]$ .

- ⑤ In the last step, we develop  $\text{CM}$ , a *durably linearizable* implementation of  $\text{AM}$ , comprising a fine-grained concurrent memory library that operates over both volatile memory and persistent memory. The library  $\text{CM}$  manages (persistent) logging to undo partially

completed operations (in case of a crash and recovery) and flushing to ensure operations on volatile memory are made persistent.

More importantly (see Theorem 4.3 below), we obtain a trace refinement property: the traces of DIMPL[CM] projected onto the events of DIMPL only (i.e., ignoring the library calls) are a subset of the traces of DIMPL[AM] projected onto the events of DIMPL.

**4.1. Step 1: NOREC.** We start by instantiating IMPL to NOREC [DSS10], which is given by the algorithm in Figure 6.<sup>4</sup> NOREC employs a deferred update strategy: writes to shared state are first stored in a write set and at commit time written to main memory – if there are no conflicts with other transactions. For conflict detection, NOREC uses value-based validation (see the operation `TMValidate`).

To synchronise concurrent transactions, NOREC uses a global counter `glb` (initially 0) and a local variable `loc`, which is used to store a copy of `glb`. Each transaction maintains a local write set, `wrSet`, and a local read set, `rdSet`. Inside its `wrSet` a transaction records all the addresses that it attempts to update and their values. The actual update of the memory takes place inside the commit operation. An odd `glb` indicates that a live writing transaction attempts to commit. After a successful commit `glb` is incremented so that its value is once again even. Thus a live transaction can determine whether another writing transaction has performed a commit operation by checking whether the value of `glb` is equal to its local copy, `loc`. Inside its `rdSet`, a transaction records the addresses that it reads and their corresponding values. Every time a transaction attempts to read an address that is not inside its `wrSet`, if `loc`  $\neq$  `glb`, then the validation method is executed. The validation method waits until the global lock is not held (`glb` is even), then checks that the `rdSet` is still valid (w.r.t., the current memory state). In the case of writing transactions, the validity of `rdSet` is also checked at the commit stage.

Operation `TMBegin` copies the value of `glb` into its local variable `loc` and checks whether `glb` is even. If this is so, the transaction is started. Otherwise, a writing transaction is in progress, so the process attempts to start again by rereading `loc`. The operation `TMValidate` checks if the transaction's `rdSet` is consistent with the current state of memory to `glb`, and returns `time`.

`TMRead` first checks if the transaction has already written the address it attempts to read. In that case it returns the address' value from its `wrSet`. Otherwise it checks if `loc` is up to date, i.e., equal to `glb`. If it is not up to date, another transaction has updated the memory and `rdSet` should be checked to ensure that its values are consistent with the current state of the memory. The check is performed by calling the operation `TMValidate`. If the `rdSet` is found consistent, the `loc` is updated with the value of `time` that `TMValidate` returns (R4). The (address, value) pair is then added in `rdSet` for future validation (R6). Finally, the value of the read address is returned. If the `rdSet` is not found consistent, the transaction aborts.

`TMWrite` adds the (address, value) pair that is to be written at the `wrSet` of the transaction. The memory is updated at the commit stage. The operation `TMCommit` first checks if the transaction is a read-only transaction. If it is, then no further checking is required and the transaction commits at E1. If it is not, E2 checks whether a concurrent writing transaction has been committed. If no such commit has occurred, the CAS at E2

<sup>4</sup>In [DDD<sup>+</sup>21], we describe a procedure for transforming pseudocode into IOA, which we use here.

```

Init:
I1 glb := 0;

TMBegint:
B1 do loct := glb;
B2 until even(loct)
   return ok;

TMReadt(addr):
R1 if addr ∈ dom(wrSett) then
   return wrSett(addr);
R2 vt := *addr;
R3 if loct ≠ glb then {
R4   loct := TMValidatet;
R5   goto R2;
}
R6 rdSett.insert(addr, vt);
   return vt;

TMWritet(addr, val):
W1 wrSett.insert(addr, val);
   return ok;

TMCommitt:
E1 if wrSett.isEmpty()
   then return ok;
E2 while !cas(glb, loct, loct + 1)
E3   loct := TMValidatet;
E4 for ∀(addr, val) ∈ wrSett
E5   *addr := val;
E6 glb := loct + 2;
   return ok;

TMValidatet:
V1 while true
V2   timet := glb;
V3   if odd(timet) then goto V2;
V4   for ∀(addr, val) ∈ rdSett do
V5     if *addr ≠ val
       then abort;
V6   if timet = glb
       then return timet;

```

Figure 6: The NOREC algorithm. Line numbers for `return` statements are omitted.

succeeds, i.e.,  $loc = glb$ , so  $glb$  becomes odd (meaning that the writing transaction obtains the lock). If the CAS does not succeed a concurrent writing transaction has been committed and  $rdSet$  needs further validation. E3 validates  $rdSet$  and updates the value of  $loc$ . By this, it prepares the transaction for another commit attempt. When the commit has obtained the lock by making  $glb$  odd, memory is updated with all the values from the write set in the loop at E4 and E5. At E6 the transaction releases the lock by making the  $glb$  value even again.

Correctness of NOREC has been verified by Lesani et al. [LLM12a] using the theorem prover PVS. The proof proceeds via showing a refinement relationship (simulation relation) between NOREC and TMS2 [DGLM13]. For the proof, NOREC is first of all transformed into an IOA. Here, we only exemplify this on one operation, `TMWrite` with one statement W1. The operation is split into three actions, an invocation and a response action (both external) plus a do action (internal).

$inv_t(\text{TMWrite}(l, v))$	$\text{DoWrite}_t$	$res_t(\text{TMWrite}(\text{ok}))$
Pre: $pc_t = \text{ready}$	Pre: $pc_t = \text{doWrite}(l, v)$	Pre: $pc_t = \text{resWrite}$
Eff: $pc_t := \text{doWrite}(l, v)$	Eff: $pc_t := \text{resWrite}$	Eff: $pc_t := \text{ready}$
	$wrSet_t := wrSet_t \oplus \{l \rightarrow v\}$	

The proof of refinement between NOREC and TMS2 is carried out via the construction of a sequence of IOA in between TMS2 and NOREC, and a sequence of simulation proofs

```

TMReadt(addr):
R1 if addr ∈ dom(wrSett) then
    return wrSett(addr)
R2 atomic {
    if owns = t ∨ owns = ⊥
    then vt := *addr
    else vt := ? }
R3 if loct ≠ glb then {
R4   loct := TMValidatet
R5   goto R2
    }
R6 rdSett.insert(addr, vt);
    return vt

TMCommitt:
E1 if wrSett.isEmpty()
    then return ok;
E2 while !cas(glb, loct, loct + 1)
E3   loct := TMValidatet
E4 atomic {
    if owns = ⊥ then owns := t }
E5 atomic {
    for ∀ addr. (addr, val) ∈ wrSett
        *addr := val }
E6 atomic {
    if owns = t then owns := ⊥ }
E7 glb := loct + 2;
    return ok;

Recovery:
RC1 glb := 0
RC2 owns := ⊥

```

Figure 7: The TMRead, TMCommit and Recovery operations of DNOREC. Note that the loop at E5 now executes atomically.

from one to the next IOA on this sequence. We will not give all the details of this proof here, rather concentrate on the key concepts and their relationships to the durable version.

**4.2. Step 2: Defining NOREC with crash and recovery.** Next, we define an enhanced algorithm DNOREC. DNOREC plays the role of DIMPL in Figure 5. It differs from NOREC in three ways.

First, DNOREC has an additional operation which abstractly models the occurrence of a crash and the subsequent recovery operation.<sup>5</sup> In particular, it (1) simulates crashes by ensuring that no transaction “survives” crashes, i.e., the currently running transactions cannot continue their operations, and (2) performs a recovery to bring the metadata (in DNOREC, glb) back to the initial state. As discussed above, in DNOREC, we assume that the heap is persistent, thus the recovery part is almost empty. We give this additional operation directly as an IOA action:

*crashRecovery*

Pre: *true*

Eff:  $pc := \lambda t \in T. \text{if } pc_t \notin \{\text{notStarted}, \text{committed}\} \text{ then aborted else } pc_t$   
 $glb := 0$   
 $owns := \perp$

In Figure 7, we present pseudocode describing the Recovery procedure modelled by an atomic action.

<sup>5</sup>Note that crash and recovery could be modelled as two separate operations. However, we expect recovery to execute in our implementation immediately after a crash and before any new transactions are started, i.e., the crash and subsequent recovery are sequential. Thus, we simplify the model and combine the crash and recovery operations.

Second, dNOREC’s commit operation is different to that of NOREC to deal with the fact that a crash can occur at any time. dNOREC must therefore update the shared memory with its write set *atomically*. In a later step (see subsection 4.5) we show how the write-back can be safely made non-atomic when using both volatile and non-volatile memory.

Third, to be compatible with the abstract library in step (4), we introduce an ownership variable, `owns`, whose value is equal to a transaction iff that transaction currently has permission to write to the memory. In particular, `owns` is acquired by a transaction immediately prior to performing a write back (E4), and released immediately after (E6). A read from the (persistent) heap must return a random value in the presence of a concurrent writer since this indicates a potential data race between a reader and writer. In dNOREC (see Figure 7), the read at line R2 reads from memory only if  $\text{owns} = t \vee \text{owns} = \perp$  and otherwise returns a random value. A read returning a random value in the presence of another writing transaction is unproblematic from the perspective of (durable) opacity since such read operations will either be revalidated, or if a revalidation is not possible, the reading transaction will abort. In particular, for the implementation to be (durably) opaque, a read must never return an illegitimate value even if it reads this value from memory.

The idea of using ownership in an interface to enforce atomicity has been explored in prior work [SBPR20]. Variables akin to ownership are typically already present in correctness proofs of opacity since a transaction must have exclusive access to the shared memory during write back. For NOREC the `owns` variable is equivalent to the already existing auxiliary `commitLock` variable in the proof by Lesani et al [LLM12a].

**4.3. Step 3: Checking dNOREC refines dTMS2.** We prove durable opacity of dNOREC by showing that it refines dTMS2. This is straightforward to check for three reasons. (1) We make the write-back in `TMCommit` of NOREC atomic, and this trivially preserves behaviours of a non-atomic write-back. (2) The only new operation is *crashRecovery*, which preserves the original simulation relation used in the original proof by Lesani et al. (3) Reads from memory return an undefined value only when we know that the corresponding `TMRead` operation will fail.

**Lemma 4.1.**  $\text{dNOREC} \leq \text{dTMS2}$ .

The proof has been mechanised in PVS [BDD<sup>+</sup>21], and is an adaptation of the mechanised proof by Lesani et al. [LLM12a] that shows that NOREC refines TMS2. Their proof is structured into four layers as shown in Figure 8. This structure keeps each refinement proof small, and design details of the NOREC algorithm are incrementally introduced. The most abstract is the TMS2 specification, which is shown to be refined by the next layer, `NORECATOMICCOMMITVALIDATE`, where read validation and commit write back are atomic. The next layer, `NORECDERIVED`, introduces a fine-grained write back operation, but leaves the read validation atomic. Finally, NOREC is shown to be a refinement of `NORECDERIVED`, where reads and validation are split into separate atomic steps.

The three changes needed between NOREC and dNOREC must be reflected in each of these layers as shown in Figure 8, i.e., we obtain `dNORECATOMICCOMMITVALIDATE` and `dNORECDERIVED`, which are analogues of `NORECATOMICCOMMITVALIDATE` and `NORECDERIVED`, respectively. We have the following changes as highlighted in Figure 8.

- `dNORECATOMICCOMMITVALIDATE` is obtained from `NORECATOMICCOMMITVALIDATE` by introducing a crash-recovery operation, which, like dNOREC resets `glb` to 0. No other changes are necessary since transactional read and write operations are atomic.

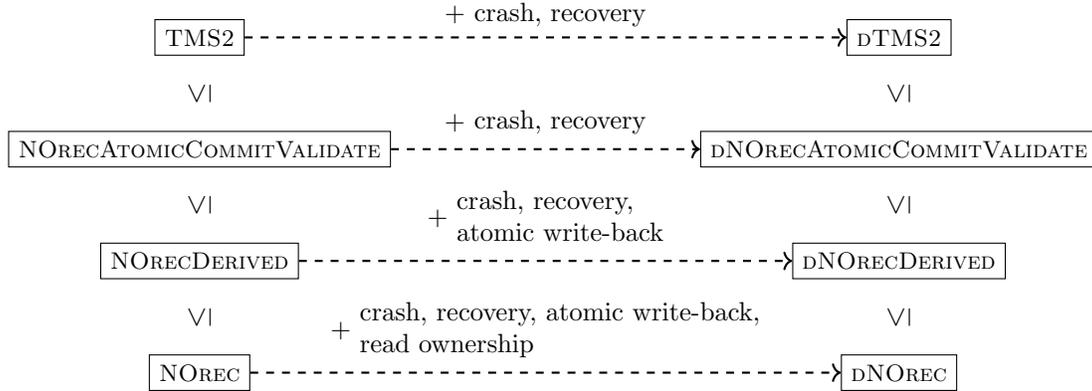


Figure 8: Adapting Lesani et al's [LLM12a] proof steps

- `DNORECDERIVED` is derived from `NORECDERIVED` by introducing the crash-recovery operation described above, and additionally reintroducing an atomic write-back (since `NORECDERIVED` uses a fine-grained commit loop). From the perspective of the simulation proof, this introduces a minor change to the verification, whereby the linearization point is shifted. In `NORECDERIVED`, the line corresponding to the successful `cas` at line `E3` can be used as the linearization point since this is the point at which the commit lock (aka ownership) is taken. In the context of durable opacity, linearizing the commit at a successful `cas` is no longer valid since the operation could still crash even after the `cas` is successful. Thus, in the revised proof, we shift the linearization point to the atomic write-back itself.
- The differences between `NOREC` and `DNOREC` are already described above; the modified operations `TMRead` and `TMCommit` are shown in Figure 7. Since this level splits the atomicity of `TMRead`, in addition to the changes described for `DNORECDERIVED`, we must allow reads to return a random value if the read is destined to fail (as discussed above). This allows one-one compatibility with the abstract library introduced in the next step. Use of a read that returns a random value is unproblematic from the perspective of the proof since a key invariant for `NOREC` is that no other transaction is performing its write back when a transaction is reading.

**4.4. Step 4: Modularising `DNOREC` and defining AM.** Next, we modularise `DNOREC` by calling a library instead of directly accessing shared memory. We start with a sequential library specification  $\mathbb{L}$  (Figure 9, left), which we convert into a concurrent durable IOA (Figure 9, right) using the technique described in subsection 2.3.

The modularised algorithm `DNOREC[AM]` is given in Figure 10, where reads from memory occur through calls to the `LibRead` operation, the write-back occurs via `LibWriteSet`, and acquire/release of ownership via `LibAcquire` and `LibRelease`, respectively. In the first instance, we start with an abstract library `AM` (see Figure 9) that matches the code in `DNOREC` exactly. Technically, moving atomic steps of `DNOREC` to library calls that execute the same atomic step does not change the algorithm. Its traces are unchanged if the external invoke and response actions calling and returning from library operations are hidden.

Note that it is crucial for the library to include operations for acquiring and releasing ownership. The specification  $\mathbb{L}$  directly expresses that concurrent `LibWriteSet` calls are

<b>LibAcquire<sub>t</sub></b>	$inv_t(\text{LibRead}(addr))$
Pre: $true$	Pre: $lpc_t = \text{ready}$
Eff: <b>if</b> $owns = \perp$ <b>then</b> $owns := t$	Eff: $lpc_t := \text{doLibRead}(addr)$
<b>LibRead<sub>t</sub>(addr; v)</b>	<b>DoLibRead<sub>t</sub>(addr)</b>
Pre: $true$	Pre: $lpc_t = \text{doLibRead}(addr)$
Eff: <b>if</b> $owns = t \vee owns = \perp$ <b>then</b> $v := mem(addr)$ <b>else</b> $v := ?$	Eff: <b>if</b> $owns = \perp \vee owns = t$ <b>then</b> $v := mem(addr)$ <b>else</b> $v := ?$ $lpc_t := \text{resLibRead}(v)$
<b>LibRelease<sub>t</sub></b>	$res_t(\text{LibRead}(v))$
Pre: $true$	Pre: $lpc_t = \text{resLibRead}(v)$
Eff: <b>if</b> $owns = t$ <b>then</b> $owns := \perp$	Eff: $lpc_t := \text{ready}$
<b>LibWriteSet<sub>t</sub>(wrset)</b>	
Pre: $owns = t$	
Eff: $mem := mem \oplus wrset$	
<b>LibRecovery</b>	
Pre: $true$	
Eff: $lpc := \lambda t : T.$ <b>if</b> $lpc_t \neq \text{notStarted}$ <b>then</b> $\text{crashed}$ <b>else</b> $lpc_t$	

Figure 9: Sequential Specification  $\mathbb{L}$  (left) of the library and transitions for LibRead of the IOA  $\text{AM} = \text{DURAUT}(\mathbb{L})$  (right)

impossible since the precondition of **LibWriteSet** requires that the calling thread is the current owner. This is exploited when developing a concurrent implementation such as CM defined in the next subsection. In particular, the correctness proof of CM does not have to prove linearizability for two concurrent calls of **LibWriteSet**, which would have been necessary for a library that only offers **LibRead** and **LibWriteSet** without mentioning ownership. Ownership therefore is used as a way to formalise “linearizability under constraints of not calling specific operations concurrently” as ordinary linearizability (here: durable linearizability).

Using the notation from subsection 3.3,  $\text{DNOREC}[\text{AM}]$  denotes the program in Figure 10 using the abstract memory in Figure 9. The traces of  $\text{DIMPL}[\text{AM}]$  include, as external actions, the external actions of both **DIMPL** and those of the library **AM**. Let  $\text{traces}(\text{DIMPL}[\text{AM}])|_{\text{DIMPL}}$  denote the traces restricted to just **DIMPL**. The next lemma establishes durable opacity of  $\text{DNOREC}[\text{AM}]$  by stating that it refines **DNOREC** which we know to refine **DTMS2** (by Lemma 4.1) which itself is durably opaque (by Theorem 3.8).

**Lemma 4.2.**  $\text{traces}(\text{DNOREC}[\text{AM}])|_{\text{DNOREC}} = \text{traces}(\text{DNOREC})$ .

Note that we prove that the trace sets are equal, not just subset. To do this we prove that the preconditions of **AM** operations are always satisfied at their call sites in  $\text{DNOREC}[\text{AM}]$ . A client that violates the precondition of an **AM** call at some call site would be deadlocked due to the semantics of IO Automata: a violated precondition of a transition means that

```

TMBegint:
B1 do loct := glb;
B2 until even(loct)
   return ok;

TMReadt(addr):
R1 if addr ∈ dom(wrSett) then
   return wrSett(addr)
R2 vt := LibReadt(addr)
R3 while loct ≠ glb
R4   loct := TMValidatet
R5   vt := LibReadt(addr)
R6 rdSett.insert(addr, vt);
   return vt

Recovery:
RC1 atomic {
   LibRecovery;
   glb := 0; }

TMWritet(addr, val):
W1 wrSett.insert(addr, val);
   return ok;

TMCommitt:
E1 if wrSett.isEmpty()
   then return ok;
E2 while !cas(glb, loct, loct + 1)
E3   loct := TMValidatet
E4 LibAcquiret
E5 LibWriteSett(wrSett)
E6 LibReleaset
E7 glb := loct + 2;
   return ok;

TMValidatet:
V1 while true
V2   timet := glb
V3   if odd(timet) then goto V2
V4   for ∀ (addr, val) ∈ rdSett do
V5     if LibReadt(addr) ≠ val
       then abort
V6   if timet = glb
       then return timet

```

Figure 10: The DNOREC[·] algorithm with library calls that relegate memory operations to a library

it is disabled. The refinement would still be correct for such a client, since IO automata refinement as well as (durable) linearizability/opacity does not guarantee any liveness. In the extreme an empty implementation that has no transitions enabled at all is correct, though not useful. Proving that preconditions of calls hold, together with the fact that a sequential program, when translated to an IO automaton always has its next step enabled, guarantees that deadlocks are avoided in our case study.

**4.5. Step 5: Defining CM and proving durable linearizability.** So far, the read/write operations on transactional variables have existed entirely on persistent memory. Our final task, therefore, is to develop a concrete library, CM, that is durably linearizable w.r.t. AM and manages low-level read/write operations across volatile and persistent memory (see Figure 5). The following theorem ensures that it is safe to perform such a replacement without violating durable opacity.

**Theorem 4.3.** *If (1) CM is durably linearizable w.r.t.  $\mathbb{L}$ , (2) AM equals  $\text{DURAUT}(\mathbb{L})$  and (3)  $\text{DIMPL}[\text{AM}]$  is a refinement of  $\text{DTMS2}$ , then  $\text{DIMPL}[\text{CM}]$  is durably opaque.*

*Proof.* The proof is by applying Theorem 3.7. Durable linearizability of CM to  $\mathbb{L}$  is equivalent to CM being a refinement of the canonical IOA, AM, which has been shown in Lemma 4.4.  $\text{DIMPL}[\text{CM}]$  and  $\text{DIMPL}[\text{AM}]$  can be constructed as the product IOA of  $\text{DIMPL}[\cdot]$  and

CM/AM, respectively. The shared external actions and steps between both IOA are the invocations and responses of library operations, together with the crash. We assume that the  $run_t$  action of AM and CM is synchronised with the  $inv_t(\mathbf{TM}\mathbf{Begin})$  action of the  $\mathbf{DNOREC}[\cdot]$  which starts a transaction. The theorem states that  $\mathbf{DIMPL}[\mathbf{CM}]$  is a refinement of  $\mathbf{DIMPL}[\mathbf{AM}]$ , implying that  $traces(\mathbf{DNOREC}[\mathbf{CM}])_{|\mathbf{DNOREC}} \subseteq traces(\mathbf{DNOREC}[\mathbf{AM}])_{|\mathbf{DNOREC}}$  hiding invocations and responses of library operations. Since  $\mathbf{DIMPL}[\mathbf{AM}]$  refines  $\mathbf{DTMS2}$  (Lemmas 4.1 and 4.2), and by transitivity of refinement we get that  $\mathbf{DIMPL}[\mathbf{CM}]$  refines  $\mathbf{DTMS2}$ , implying durable opacity.  $\square$

We now describe the instance of CM that we use (see Figure 11). CM implements AM on an architecture with persistent and volatile memory: instead of writing directly to persistent memory  $mem$  (as in the case of AM), CM first writes to the concrete volatile memory,  $vmem$ , and this is later flushed to the concrete persistent memory,  $pmem$ .

Functional correctness is not affected when all transactions read and write to  $vmem$  instead of  $mem$ . However, after a crash the data in  $vmem$  is lost, and computation resumes from the state of  $pmem$ . Therefore, to ensure durable opacity, we have to ensure that  $pmem$  is updated during a commit so that the memory snapshot that results from the successful commit is available even after a crash. For a lazy STM implementation like NOREC, committing the write set is the only place in the code which writes to memory, so the implementation must update both  $vmem$  and  $pmem$  during a commit write back.

In NOREC, a crash occurring partway through a commit write back may result in an inconsistent memory state, i.e., one that is not a snapshot of the successfully completed transactions. We treat transactions that crash during (or before) a commit write back to be an aborted transaction, thus any memory updates performed by a partially completed write back operation must be reverted. To make this possible, we keep a *persistent* log  $plog$  that stores old values for those locations of the write set that have already been committed.

This leads to the following algorithm for committing the given write set  $wrSet$ :

```

for  $\forall$  ( $addr, val$ )  $\in$   $wrSet$  do
  oldv := *addr;
  plog := plog  $\oplus$  { $addr \mapsto oldv$ };
  *addr := val;
  flush(addr);
plog :=  $\emptyset$ ;

```

In KIV, the abstract code “for  $\forall$  ( $addr, val$ )  $\in$   $wrSet$ ” is realised as a while loop, that iterates over the write set. Translating to steps of an IOA, this gives the steps shown in Figure 11. The first action  $W1$  is the loop test, that checks whether  $wrSet$  is empty. In case it is not, an  $addr$  is chosen in step  $W2$ , and the four instructions of the loop body above are executed as steps  $W3$  to  $W6$ . Flushing moves  $vmem(addr)$  to  $pmem(addr)$ .  $addr$  is then removed from  $wrSet$  in step  $W7$  which jumps back to the loop test. When  $wrSet$  is empty, the loop is left and step  $W8$  resets the persistent log. In addition to the program steps the IO automaton for CM includes a  $flush(l)$  step (with an internal action) that models flushing a memory location  $l$  that is possible at any time.

On a crash, the log is used to undo the partial commit<sup>6</sup>. When the write set has been fully committed, the log is cleared, and clearing the log at  $W8$  becomes the linearization

<sup>6</sup>In the IOA, the recovery executes  $vmem := pmem \oplus plog$ ; the KIV specification uses a recovery program that writes each log entry separately in a loop.

point of the implementation of commit. After this point the transaction has successfully committed.

$inv_t(\text{LibWriteSet}(wrSet))$ Pre: $lpc_t = \text{ready} \wedge \text{owns} = t$ Eff: $lpc_t := W1(wrSet)$	$W1_t$ Pre: $lpc_t = W1(wrSet)$ Eff: $lpc_t := \text{if } (wrSet \neq \emptyset) \text{ then } W2(wrSet) \text{ else } W8$
$W2_t$ Pre: $lpc_t = W2(wrSet)$ $SOME \text{ addr}. \text{addr} \in \text{dom}(wrSet)$ Eff: $lpc_t := W3(wrSet, \text{addr})$	$W3_t$ Pre: $lpc_t = W3(wrSet, \text{addr})$ Eff: $lpc_t := W4(wrSet, \text{addr})$ $oldv_t := \text{vmem}(\text{addr})$
$W4_t$ Pre: $lpc_t = W4(wrSet, \text{addr})$ Eff: $lpc_t := W5(wrSet, \text{addr})$ $plog := plog \oplus \{\text{addr} \rightarrow oldv_t\}$	$W5_t$ Pre: $lpc_t = W5(wrSet, \text{addr})$ Eff: $lpc_t := W6(wrSet, \text{addr})$ $\text{vmem}(\text{addr}) := wrSet(\text{addr})$
$W6_t$ Pre: $lpc_t = W6(wrSet, \text{addr})$ Eff: $lpc_t := W7(wrSet, \text{addr})$ $\text{pmem}(\text{addr}) := \text{vmem}(\text{addr})$	$W7_t$ Pre: $lpc_t = W7(wrSet, \text{addr})$ Eff: $lpc_t := W1(wrSet \setminus \{\text{addr}, wrSet(\text{addr})\})$
$W8_t$ Pre: $lpc_t = W8$ Eff: $lpc_t := \text{resLibWriteSet}$ $plog := \emptyset$	$res_t(\text{LibWriteSet}())$ Pre: $lpc_t := \text{resLibWriteSet}$ Eff: $lpc_t := \text{ready}$
$R1_t$ Pre: $lpc_t = \text{doLibRead}(\text{addr})$ Eff: $v := \text{vmem}(\text{addr})$ $lpc_t := \text{resLibRead}(v)$	$run_t$ Pre: $lpc_t = \text{notStarted}$ Eff: $lpc_t := \text{ready}$
$\text{LibRecovery}$ Pre: $true$ Eff: $lpc := \lambda t : T. \text{if } lpc_t \neq \text{ready} \text{ then } \text{crashed} \text{ else } lpc_t$ $\text{owns} := \perp$ $\text{vmem} := \text{pmem} \oplus plog$ $\text{pmem} := \text{pmem} \oplus plog$ $plog := \emptyset$	$\text{flush}(l)$ Pre: $true$ Eff: $\text{pmem}(l) := \text{vmem}(l)$

Figure 11: Transition relation of CM. Transitions for Acquire/Release, as well as invoke and response transitions for read are the same as in AM.

For the concrete library CM we have shown the following result.

**Lemma 4.4.** CM *refines* AM.

Since AM is  $\text{DURAUT}(\mathbb{L})$  for the sequential specification  $\mathbb{L}$  of the library given in Figure 9, we get the following corollary of Lemma 3.5.

**Corollary 4.5.** *CM is durably linearizable to  $\mathbb{L}$ .*

Lemma 4.4 has been mechanically proven in the theorem prover KIV [SBBR22]. Both the IOA for AM and for CM are specified in KIV by giving labelled programs which generate a predicate logic specification of the transition relation. Specifications and proofs are online at [BDD<sup>+</sup>21].

The refinement from AM to CM is proven in two steps. First an invariant for CM is proven which overapproximates the set of reachable states. The invariant is then used in place of  $reach(C)$  in the proof of a forward simulation according to Def. 3.2.

The invariant of CM consists of a global invariant and local assertions. The global invariant

**if**  $owns = \perp$  **then**  $vmem = pmem \wedge plog = \emptyset$  **else**  $vmem \oplus plog = pmem \oplus plog$

states that as long as there is no writer that commits a write set, volatile and persistent memory agree, and the log is empty. Otherwise overriding the volatile and persistent memory with the log gives the same result: both result in the memory snapshot at the start of the commit.

The local assertions give formulas that hold when a transaction  $t$  is at a specific program counter,  $lpc_t$ . (We use  $lpc$  here to distinguish the library program counter.) As an example, while  $t$  is executing the write operation ( $lpc_t$  is one of  $W1$  to  $W8$ ) it has write ownership ( $owns = t$ ). The KIV specification specifies this implication (and many more) as pairs of a label range and a formula. The full invariant is generated as a conjunction of all local assertions, that is universally quantified over all  $t$ , together with the global invariant. To have a thread-modular proof of the assertions, a rely predicate  $rely(t, s, s')$  is specified that the steps of all other transactions  $t' \neq t$  (from state  $s$  to  $s'$ ) and flush steps of the system must satisfy. Assertions for thread  $t$  and the global invariant are shown to be stable with respect to this predicate. In our case the rely predicate consists of three formulas.

$$\begin{aligned} owns = t &\rightarrow vmem = vmem' \wedge owns = owns' \wedge plog = plog' \\ owns = t &\rightarrow between(pmем, pmем', vmem) \\ owns \neq t &\rightarrow owns' \neq t \end{aligned}$$

The first ensures that while a transaction  $t$  is writing other transactions will leave  $vmem$ ,  $owns$  and  $plog$  unchanged. The second asserts that while a writer is running,  $pmем$  may only be changed by system flushes:  $between(pmем, pmем', vmem)$  asserts that all values  $pmем'(l)$  will either still be  $pmем(l)$  or be the flushed value  $vmem(l)$ . The third guarantees that steps of other threads cannot make  $t$  the writer.

The specification of CM in KIV also fixes the linearization points of CM, by defining non- $\tau$  actions for such steps. Reading linearizes at `doLibRead`, when the value is read from volatile memory. Committing linearizes at  $W8$ , when the log is set to empty.

The forward simulation between CM and AM consists of a global part and a local part for every transaction  $t$  too. The global part simply states that  $owns$  of AM and CM are identical and that the abstract memory  $mem$  of AM is always equal to  $vmem \oplus plog$  (the invariant implies that it is then equal to  $pmем \oplus plog$  as well).

The local part of the simulation for thread  $t$  gives a mapping between program counter values of CM and AM in the obvious way. As an example, since the linearization point of commit is at  $W6$ , all  $lpc_t$  values before and including  $W6$  are mapped to `doLibWriteSet( $wrSet_t$ )`, while  $lpc_t = \text{resLibWriteSet}$  is mapped to `resLibWriteSet`. The local part of the simulation

also ensures that input received by an operation of CM that is stored in a local variable is equal to the corresponding input of AM, and similar for the outputs.

With this forward simulation, the proof has to show a commutativity for every step of CM according to Definition 3.2. The proofs of this refinement in KIV are simple. Three days of work were required to set up the specifications and to do the proofs, which have ca. 300 interactive steps.

The final step is to combine the steps above instantiating Theorem 4.3, resulting in the corollary below.

**Corollary 4.6.** *If (1) CM is durably linearizable w.r.t.  $\mathbb{L}$ , (2) AM equals  $\text{DURAUT}(\mathbb{L})$  and (3)  $\text{DNOREC}[\text{AM}]$  is a refinement of  $\text{DTMS2}$ , then  $\text{DNOREC}[\text{CM}]$  is durably opaque.*

Note, that in the implementation CM the *owns* variable is an auxiliary variable, that has no effect on computations. Therefore the final program code equivalent to the IOA  $\text{DIMPL}[\text{CM}]$  can omit the variable together with the calls to `LibAcquire` and `LibRelease`.

## 5. RELATED WORK

The literature around persistent memory has grown remarkably quickly. Below we provide a snapshot of some related work, focussing in particular on correctness and atomicity.

**5.1. Correctness conditions.** Constructing robust shared objects for NVM requires the development of criteria that provide meaningful guarantees in the presence of crashes. Linearizability [HW90], is one of the most well-known, broadly used, correctness conditions for concurrent objects. Several correctness conditions attempt to adapt linearizability to histories that include crash events.

As mentioned before, durable linearizability [IMS16] extends the events that can appear in an abstract concurrent history with crash events. Crashes are considered global events. Durable linearizability expects that no thread survives after a crash, thus a thread can operate only in one crash-free region. On the contrary, strict linearizability [AF03], consider crashes to be local to the threads that they occur. Under this condition, operations that are not subjected to a failure can take effect between their invocation and response. In the case that a thread crashes while executing an operation, it requires this operation to take effect between its invocation and the crash, but not after the crash. Operations that are disrupted by a crash either take effect or abort when a crash occurs. Guerraoui and Levy [GL04] have defined two more correctness conditions that extend linearizability, *persistent atomicity*, and *transient atomicity*. Persistent atomicity requires that, in the event of a crash, every pending operation on the crashed thread either takes effect or aborts before a subsequent operation of the same thread is invoked, noting that an operation may take effect after a crash. Transient atomicity relaxes this condition further, by allowing an incomplete operation to take effect before a subsequent write response of the same thread. Berryhill et al. [BGT16] have proposed *recoverable linearizability*, which requires every pending operation on a thread to take effect or abort before the thread linearizes another operation. This condition does not provide consistency around the crash — a thread can perform an operation on some other object before coming back to the pending operation causing “program order inversion”. The main disadvantage of strict linearizability, persistent atomicity and transient atomicity is that they are not compositional. On the other hand, durable and recoverable linearizability are compositional.

Several models, both hardware and software specific, aim to define the correctness of the order in which writes are persisted in NVM. Pelley et al. [PCW14] described various such low-level models including *strict persistency* and relaxed persistency models such as *epoch persistency* and *strand persistency*. Those models consider hardware to be able to track persist dependences and perform flushes in a manner described by the persistency model. Izraelevitz et al. [IMS16] gave formal semantics to epoch persistency which corresponds to real-world explicit ISAs, where flushes are issued explicitly with dedicated instructions by the respective application. Raad et al. [RWV19] developed declarative semantics that formalise the persistency semantics of ARMv8 architecture. [RWNV20] propose persistency semantics for the Intel-x86 Architecture and [CLRK21] provides view-based and axiomatic persistency models for Intel-x86 and ARMv8.

On the language level, Kolli et al. [KGS<sup>+</sup>17] proposed an acquire and release persistency model based on the acquire-release consistency of C++11. Furthermore, Raad et al. [RLV20] and Bila et al. [BDL<sup>+</sup>22] have developed program logics for reasoning about persistent programs on Intel-x86, based on the Owicki-Gries proof system.

Regarding transactional memory [HLR10], not many correctness conditions have been adapted to the persistent memory setting. Raad et al [RWV19] base their framework for formalising ARMv8 to a persistent variant of serializability (PSER) under relaxed memory. Even though serializability provides simple intuitive semantics, it does not handle aborted transactions. TimeStone [KKM<sup>+</sup>20] and Pisces [GYW<sup>+</sup>19] are recent persistent transactional memories that guarantee snapshot isolation [BBG<sup>+</sup>95], which is weaker than serializability, and hence opacity.

**5.2. Persistent Transactional Memory (PTM).** Mnemosyne [VTS11] provides a low-level interface to persistent memory with high-level transactions based on TinySTM [FFR08]. The Mnemosyne transaction system combines lazy version management with the eager conflict detection that encounter-time locking provides. The lazy version management is implemented with a redo log, which has been chosen to reduce ordering constraints. The new writes to persistent memory are kept in a redo log and are buffered in the volatile memory. When a write transaction commits, it flushes the log to the persistent memory and optionally writes back the new values. Unlike TMs that use undo logging, the write transactions do not update the memory until they commit. This adds an overhead to read transactions, since they should recognise the modified values, but not yet committed values, and then return them from the buffer. Moreover, the size of the log increases proportionally to the size of the transaction, potentially making commits time consuming. Mnemosyne uses a global array of volatile locks to implement encounter-time locking. Every memory location is associated with a lock. Prior to accessing a memory location, the transaction identifies its associated lock and tries to acquire it. In the case that the operation succeeds, it adds the lock to the lock-set. Otherwise, it aborts and releases all the locks contained in its lock-set.

NV-heaps [CCA<sup>+</sup>11] is a persistent object system that aims to integrate persistent objects into conventional programs, and furthermore seeks to prevent safety bugs that occur in predominantly persistent memory models, such as multiple frees, pointer errors etc. NV-heaps only handle updates to persistent memory inside transactions and critical sections. It uses ACID transactions to guarantee the consistency of persistent objects in the face of system failures. Specifically, NV-heaps rely on atomic sections that log all the updates of the non-volatile memory to provide fine-grain consistency. Each transaction keeps a volatile read log and a non-volatile write log. NV-heaps provide eager conflict detection for writes.

The system keeps a copy of the objects that are going to be modified by write transactions in an undo log. In this way, the modifications to an object can be rolled back in the case of an atomic section abort or a system failure. Each log update needs an epoch barrier, which affects the overall performance. Before a transaction tries to modify an object, its atomic section attempts to take ownership of the object by acquiring a volatile lock in a table of ownership records. In case of success, the entire object is copied into the write undo log and the transaction proceeds to modify the object. Otherwise, the atomic section retries. To read an object, NV-heaps store a pointer to the object and its current version number in the read log. The version numbers help in detecting read conflicts at access time.

Unlike persistent transactional memories that provide durable transactions via undo and redo logs, Romulus [CFR18] provides durable transactions by keeping two copies (main and back) of the data in non-volatile memory and ensuring that at any time at least one of the copies is consistent. When a transaction begins, any modification of the data that is caused by the user-code is immediately flushed to the main copy. Before a transaction ends, the modifications in the main copy are copied to the back copy. If a failure occurs when the copying from main to back takes place, then the recovery procedure copies the contents of main to back. In the same way, if a failure occurs while the modification of main is taking place, the recovery procedure copies the contents of back to main. In order to avoid full replication of the data of the main to the back, Romulus introduces a volatile redo log that tracks the addresses of the modified data. At the end of the transaction, only those addresses are flushed to the back. There are two implementations of Romulus available, one with a scalable reader-writer lock and another that uses a universal construct and supports wait-free read-only transactions.

OneFile [RCFC19] is a wait-free PTM that supports durably linearizable transactions. It uses a redo log for durability and a time based concurrency control. In this design, every thread maintains a redo log as write set that can be read by other threads in order to help the completion of the ongoing transaction, but does not maintain a read set. All writing transactions are associated with a unique sequence number that allows their serialization. A technique that is similar to TL2 [DSS06] and also uses sequence numbers is applied to ensure consistency of read operations. The design of OneFile allows write transactions to run concurrently with read-only transactions. There are two variants of the OneFile available: one with lock-free progress and providing bounded wait-freedom.

QSTM [BCWS20] is a non-blocking persistent transactional memory. Its design is based on RingSTM [SMvP08b] enhanced with a redo log based on the persistent lock-free queue of Friedman et al. [FHMP18]. Each transaction maintains a read and a write filter. The entries of the redo log represent the live transactions. Each entry consists of a pointer to the transaction's write set (this allows any thread to perform the writes of a committed transaction), a unique timestamp associated with the represented transaction and its write filter. The validation mechanism is taking place within the read operation. Each transaction while reading is checking if its read filter conflicts with any write filter of the committed transactions. If so, it aborts. Queue entries are deleted only when their respective writes are persisted. Beadle et al [BCWS20] provide several correctness arguments of QSTM. Specifically, they argue that QSTM is linearizable as a single concurrent object, durably linearizable, and lock-free. Compared to OneFile, QSTM uses significantly less space due to the fact that it does not require modifications in data declaration or the use of `cas` and `LL/SC` instructions. However OneFile achieves higher throughput than QSTM, due to QSTM's global log.

**5.3. Generic Approaches to Persistency.** Apart from PTMs, several generic frameworks have been developed to tackle the problem of consistency under persistent memory. Indicatively, Naama Ben-David et al. [BDBFW19] developed a system that can transform programs that consist of read, write and `cas` operations in shared memory, to persistent memory. The system aims to create concurrent algorithms that guarantee consistency after a fault. This is done by introducing persist checkpoints, which record the current state of the execution and from which the execution can continue after a fault.

Izraelevitz et al. [IKK16] develop and implement a logging mechanism based on undo and redo log properties named JUSTDO logging and introduce the concept of FASE (failure-atomic sections). This mechanism aims to reduce the memory size of log entries while preserving data integrity after crash occurrences. Unlike optimistic transactions [CBB14], JUSTDO logging resumes the execution of interrupted FASEs at their last store instruction, and then executes them until completion. One disadvantage of this strategy is that the FASEs cannot be rolled back after a system failure. As a consequence, there is no tolerance of bugs inside the FASEs. In this system, it is assumed that the cache memory is persistent, and the system also requires that all load/store instructions access persistent data. A small log is maintained for each thread, that records its most recent store within a FASE. The small per thread logs simplify the log management and reduce the memory requirements.

## 6. CONCLUSION

In this paper, we use *durable opacity* as a correctness condition for STMs running on non-volatile hardware architectures. We have proposed an abstract specification dTMS2 which is durably opaque and have shown how this can be employed in refinement-based proofs of durable opacity. We have furthermore developed a *modular* proof technique for such refinement proofs, separating out the proof of durability from that of opacity. We have exemplified this proof technique on the STM NOREC for which we have – to this end – developed a version adequate for non-volatile memory.

Our proof technique is inspired by work on the verification of a Flash file system by Schellhorn et al. [SBPR20]. Although this prior work does not target NVM or STMs, it also uses ownership as a mechanism for restricting concurrency at the interface of a library. The development, which refines an abstract POSIX-compatible file system specification in several steps to the Linux interface MTD for flash hardware, uses intermediate layers (similar to the library AM used here) to introduce caches for flash pages [PEB<sup>+</sup>17] and for file-content [BSR22], though the correctness criteria defined in these papers are strictly weaker than durable linearizability.

Our new proof technique (outlined in Figure 5) describes a similar technique in the setting of STMs. In particular, we provide an abstract library that operates directly on persistent memory and a concrete implementation that uses both volatile and persistent memory. The original STM is placed in an execution context that could have system crashes and uses the abstract and concrete library to perform memory operations. Given an STM that already refines TMS2 (and hence is opaque), the bulk of the verification effort using our method is focused on verifying durable linearizability between the abstract and concrete libraries. This proof is the only one that has to consider the distinction between volatile and persistent memory. In our case it is not very difficult when the correct ownership annotations are used. In contrast, a non-modular proof would have to re-do the already complex opacity

proof in the more complex setting where the distinction between volatile and persistent memory is present.

**Future work.** We conjecture that the libraries AM and CM that we have defined could be used to transform other opaque algorithms into durably opaque algorithms when the STMs use a lazy write-back commit with mutual exclusion between the write-back operations. Other types of STMs, e.g., TL2 [DSS06] (which uses a per-location lock to allow concurrent write-backs) and TML [DDS<sup>+</sup>10] (which uses an eager write-back mechanism) cannot use the libraries AM and CM that we have developed directly. Whether the modular library based approach presented in this paper applies to these other algorithms as well remains a future research topic.

The memory model that we have assumed is strong, only making a distinction between volatile and persistent memory. The writes themselves are assumed to be sequentially consistent, and no intra-thread reordering is possible. In reality, programs executed in platforms such as persistent x86-TSO [RLV20, CLRK21], in which instructions may be reordered due to the effects of both persistency and Total Store Order (TSO). In future work, we aim to extend our methods to additionally take such effects into account. In particular, it would be interesting to know whether CM could be further refined and integrated with the remainder of the system in a modular manner. For the persistent TSO model, such proofs could build on existing logics, e.g., [RLV20, BDL<sup>+</sup>22], but may require new theories for refinement.

## REFERENCES

- [AD17] A. Armstrong and B. Dongol. Modularising opacity verification for hybrid transactional memory. In A. Bouajjani and A. Silva, editors, *FORTE*, volume 10321 of *LNCS*, pages 33–49. Springer, 2017.
- [ADD17] A. Armstrong, B. Dongol, and S. Doherty. Proving opacity via linearizability: A sound and complete method. In A. Bouajjani and A. Silva, editors, *FORTE*, volume 10321 of *LNCS*, pages 50–66. Springer, 2017.
- [AF03] M. K. Aguilera and S. Frølund. Strict linearizability and the power of aborting. *Technical Report HPL-2003-241*, 2003.
- [AGHR14] H. Attiya, A. Gotsman, S. Hans, and N. Rinetzky. Safety of live transactions in transactional memory: TMS is necessary and sufficient. In F. Kuhn, editor, *DISC*, volume 8784 of *LNCS*, pages 376–390. Springer, 2014.
- [BBG<sup>+</sup>95] H. Berenson, P. Bernstein, J. Gray, J. Melton, E. O’Neil, and P. O’Neil. A critique of ansi sql isolation levels. *ACM SIGMOD Record*, 24(2):1–10, 1995.
- [BCWS20] H Alan Beadle, Wentao Cai, Haosen Wen, and Michael L Scott. Nonblocking persistent software transactional memory. In *2020 IEEE 27th International Conference on High Performance Computing, Data, and Analytics (HiPC)*, pages 283–293. IEEE, 2020.
- [BDBFW19] N. Ben-David, G. E. Blelloch, M. Friedman, and Y. Wei. Delay-free concurrency on faulty persistent memory. In *The 31st ACM Symposium on Parallelism in Algorithms and Architectures*, pages 253–264, 2019.
- [BDD<sup>+</sup>20] E. Bila, S. Doherty, B. Dongol, J. Derrick, G. Schellhorn, and H. Wehrheim. Defining and verifying durable opacity: Correctness for persistent software transactional memory. In A. Gotsmanj and A. Sokolova, editors, *FORTE 2020*, volume 12136 of *Lecture Notes in Computer Science*, pages 39–58. Springer, 2020. doi:10.1007/978-3-030-50086-3\3.
- [BDD<sup>+</sup>21] E. Bila, J. Derrick, S. Doherty, B. Dongol, G. Schellhorn, and H. Wehrheim. Verification of a durable opaque version of NOREC with KIV and PVS, 2021. URL: <http://www.informatik.uni-augsburg.de/swt/projects/DNOREC.html>.

- [BDL<sup>+</sup>22] Eleni Vafeiadi Bila, Brijesh Dongol, Ori Lahav, Azalea Raad, and John Wickerson. View-based Owicki-Gries reasoning for persistent x86-TSO. In Ilya Sergey, editor, *ESOP*, volume 13240 of *Lecture Notes in Computer Science*, pages 234–261. Springer, 2022. doi:10.1007/978-3-030-99336-8\\_9.
- [BGT16] R. Berryhill, W. Golab, and M. Tripunitara. Robust shared objects for non-volatile main memory. In *19th International Conference on Principles of Distributed Systems (OPODIS 2015)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.
- [BSR22] S. Bodenmüller, G. Schellhorn, and W. Reif. Verification of Crashsafe Caching in a Virtual File System Switch. *Formal Aspects of Computing (FAC)*, 2022. to appear.
- [CBB14] D. R. Chakrabarti, H.-J. Boehm, and K. Bhandari. Atlas: Leveraging locks for non-volatile memory consistency. *ACM SIGPLAN Notices*, 49(10):433–452, 2014.
- [CCA<sup>+</sup>11] Joel Coburn, Adrian M Caulfield, Ameen Akel, Laura M Grupp, Rajesh K Gupta, Ranjit Jhala, and Steven Swanson. Nv-heaps: making persistent objects fast and safe with next-generation, non-volatile memories. *ACM SIGARCH Computer Architecture News*, 39(1):105–118, 2011.
- [CFR18] Andreia Correia, Pascal Felber, and Pedro Ramalhete. Romulus: Efficient algorithms for persistent transactional memory. In *Proceedings of the 30th on Symposium on Parallelism in Algorithms and Architectures*, pages 271–282, 2018.
- [CLRK21] Kyeongmin Cho, Sung-Hwan Lee, Azalea Raad, and Jeehoon Kang. Revamping hardware persistency models: view-based and axiomatic persistency models for Intel-x86 and Armv8. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*, pages 16–31, 2021.
- [DB14] John Derrick and Eerke A. Boiten. *Refinement in Z and Object-Z - Foundations and Advanced Applications (2. ed.)*. Springer, 2014. doi:10.1007/978-1-4471-5355-9.
- [DD15] B. Dongol and J. Derrick. Verifying linearisability: A comparative survey. *ACM Comput. Surv.*, 48(2):19:1–19:43, 2015.
- [DDD<sup>+</sup>16] S. Doherty, B. Dongol, J. Derrick, G. Schellhorn, and H. Wehrheim. Proving opacity of a pessimistic STM. In P. Fatourou, E. Jiménez, and F. Pedone, editors, *OPODIS*, volume 70 of *LIPICs*, pages 35:1–35:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016.
- [DDD<sup>+</sup>19] J. Derrick, S. Doherty, B. Dongol, G. Schellhorn, and H. Wehrheim. Verifying correctness of persistent concurrent data structures. In *FM*, volume 11800 of *Lecture Notes in Computer Science*, pages 179–195. Springer, 2019.
- [DDD<sup>+</sup>21] John Derrick, Simon Doherty, Brijesh Dongol, Gerhard Schellhorn, and Heike Wehrheim. Verifying correctness of persistent concurrent data structures: a sound and complete method. *Formal Aspects of Computing*, 2021. Online first. URL: <https://link.springer.com/article/10.1007/s00165-021-00541-8>.
- [DDS<sup>+</sup>10] L. Dalessandro, D. Dice, M. L. Scott, N. Shavit, and M. F. Spear. Transactional mutex locks. In P. D’Ambra, M. R. Guarracino, and D. Talia, editors, *Euro-Par (2)*, volume 6272 of *LNCS*, pages 2–13. Springer, 2010.
- [DDS<sup>+</sup>15] J. Derrick, B. Dongol, G. Schellhorn, O. Travkin, and H. Wehrheim. Verifying opacity of a transactional mutex lock. In *FM*, volume 9109 of *LNCS*, pages 161–177. Springer, 2015.
- [DGLM13] S. Doherty, L. Groves, V. Luchangco, and M. Moir. Towards formally specifying and verifying transactional memory. *Formal Asp. Comput.*, 25(5):769–799, 2013.
- [dRE98] W. P. de Roever and K. Engelhardt. *Data Refinement: Model-oriented Proof Theories and their Comparison*, volume 46 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press, 1998.
- [DSS06] Dave Dice, Ori Shalev, and Nir Shavit. Transactional locking ii. In Shlomi Dolev, editor, *Distributed Computing*, pages 194–208, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.
- [DSS10] L. Dalessandro, M. F. Spear, and M. L. Scott. N0rec: streamlining STM by abolishing ownership records. In R. Govindarajan, D. A. Padua, and M. W. Hall, editors, *PPoPP*, pages 67–78. ACM, 2010.
- [FFR08] Pascal Felber, Christof Fetzer, and Torvald Riegel. Dynamic performance tuning of word-based software transactional memory. In *Proceedings of the 13th ACM SIGPLAN Symposium on Principles and practice of parallel programming*, pages 237–246, 2008.
- [FHMP18] M. Friedman, M. Herlihy, V. J. Marathe, and E. Petrank. A persistent lock-free queue for non-volatile memory. In A. Krall and T. R. Gross, editors, *ACM SIGPLAN Symposium on*

- Principles and Practice of Parallel Programming, PPOPP*, pages 28–40. ACM, 2018. URL: <http://doi.acm.org/10.1145/3178487.3178490>.
- [FPR21] Michal Friedman, Erez Petrank, and Pedro Ramalhete. Mirror: making lock-free data structures persistent. In Stephen N. Freund and Eran Yahav, editors, *PLDI '21: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, Virtual Event, Canada, June 20-25, 2021*, pages 1218–1232. ACM, 2021. doi:10.1145/3453483.3454105.
- [GK08] R. Guerraoui and M. Kapalka. On the correctness of transactional memory. In S. Chatterjee and M. L. Scott, editors, *PPOPP*, pages 175–184. ACM, 2008.
- [GK10] R. Guerraoui and M. Kapalka. *Principles of Transactional Memory*. Synthesis Lectures on Distributed Computing Theory. Morgan & Claypool Publishers, 2010.
- [GL04] R. Guerraoui and R. Levy. Robust emulations of shared memory in a crash-recovery model. In *24th International Conference on Distributed Computing Systems, 2004. Proceedings.*, pages 400–407. IEEE, 2004.
- [GYW<sup>+</sup>19] J. Gu, Q. Yu, X. Wang, Z. Wang, B. Zang, H. Guan, and H. Chen. Pisces: a scalable and efficient persistent transactional memory. In *USENIX Annual Technical Conference*, pages 913–928, 2019.
- [HLR10] T. Harris, J. Larus, and R. Rajwar. Transactional memory. *Synthesis Lectures on Computer Architecture*, 5(1):1–263, 2010.
- [HW90] M. Herlihy and J. M. Wing. Linearizability: A correctness condition for concurrent objects. *ACM TOPLAS*, 12(3):463–492, 1990.
- [IKK16] Joseph Izraelevitz, Terence Kelly, and Aasheesh Kolli. Failure-atomic persistent memory updates via justdo logging. *ACM SIGARCH Computer Architecture News*, 44(2):427–442, 2016.
- [IMS16] J. Izraelevitz, H. Mendes, and M. L. Scott. Linearizability of persistent memory objects under a full-system-crash failure model. In C. Gavoille and D. Ilcinkas, editors, *DISC*, volume 9888 of *LNCS*, pages 313–327. Springer, 2016.
- [KGS<sup>+</sup>17] A. Kolli, V. Gogte, A. Saidi, S. Diestelhorst, P. M. Chen, S. Narayanasamy, and T. F. Wenisch. Language-level persistency. In *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, pages 481–493. IEEE, 2017.
- [KKM<sup>+</sup>20] M. R. Krishnan, J. Kim, A. Mathew, X. Fu, A. Demeri, C. Min, and S. Kannan. Durable transactional memory can scale with timestone. In *ASPLOS*, pages 335–349, 2020.
- [Lam79] L. Lamport. How to make a multiprocessor computer that correctly executes multiprocess programs. *IEEE Trans. Computers*, 28(9):690–691, 1979.
- [LLM12a] M. Lesani, V. Luchangco, and M. Moir. A framework for formally verifying software transactional memory algorithms. In M. Koutny and I. Ulidowski, editors, *CONCUR 2012*, volume 7454 of *LNCS*, pages 516–530, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
- [LLM12b] M. Lesani, V. Luchangco, and M. Moir. Putting opacity in its place. In *Workshop on the Theory of Transactional Memory*, 2012.
- [LT87] N. A. Lynch and M. R. Tuttle. Hierarchical correctness proofs for distributed algorithms. In *PODC*, pages 137–151, New York, NY, USA, 1987. ACM.
- [LV95] N. Lynch and F. Vaandrager. Forward and backward simulations. *Information and Computation*, 121(2):214 – 233, 1995.
- [Lyn96] N. A. Lynch. *Distributed Algorithms*. Morgan Kaufmann, 1996.
- [Mül98] O. Müller. I/O Automata and beyond: Temporal logic and abstraction in Isabelle. In J. Grundy and M. Newey, editors, *TPHOLs*, pages 331–348, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.
- [PCW14] S. Pelley, P. M. Chen, and T. F. Wenisch. Memory persistency. In *2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*, pages 265–276. IEEE, 2014.
- [PEB<sup>+</sup>17] J. Pfähler, G. Ernst, S. Bodenmüller, G. Schellhorn, and W. Reif. Modular verification of order-preserving write-back caches. In N. Polikarpova and S. A. Schneider, editors, *iFM*, volume 10510 of *Lecture Notes in Computer Science*, pages 375–390. Springer, 2017. doi:10.1007/978-3-319-66845-1\_25.
- [RCFC19] Pedro Ramalhete, Andreia Correia, Pascal Felber, and Nachshon Cohen. Onefile: A wait-free persistent transactional memory. In *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 151–163. IEEE, 2019.

- [RLV20] A. Raad, O. Lahav, and V. Vafeiadis. Persistent Owicki-Gries reasoning: a program logic for reasoning about persistent programs on Intel-x86. *Proceedings of the ACM on Programming Languages*, 4(OOPSLA):1–28, 2020.
- [RWNV20] Azalea Raad, John Wickerson, Gil Neiger, and Viktor Vafeiadis. Persistency semantics of the Intel-x86 architecture. *Proc. ACM Program. Lang.*, 4(POPL):11:1–11:31, 2020. doi:10.1145/3371079.
- [RWV19] A. Raad, J. Wickerson, and V. Vafeiadis. Weak persistency semantics from the ground up: Formalising the persistency semantics of ARMv8 and transactional models. *Proceedings of the ACM on Programming Languages*, 3(OOPSLA):1–27, 2019.
- [SBBR22] G. Schellhorn, S. Bodenmüller, M. Bitterlich, and W. Reif. Software & System Verification with KIV. In *The Logic of Software. A Tasting Menu of Formal Methods*, volume 13360 of *LNCS*. Springer, 2022. to appear.
- [SBPR20] G. Schellhorn, S. Bodenmüller, J. Pfähler, and W. Reif. Adding concurrency to a sequential refinement tower. In *International Conference on Rigorous State-Based Methods*, pages 6–23. Springer, 2020.
- [SMvP08a] M. F. Spear, M. M. Michael, and C. von Praun. RingSTM: scalable transactions with a single atomic instruction. In *Proceedings of the twentieth annual symposium on Parallelism in algorithms and architectures*, pages 275–284. ACM, 2008.
- [SMvP08b] Michael F. Spear, Maged M. Michael, and Christoph von Praun. Ringstm: scalable transactions with a single atomic instruction. In Friedhelm Meyer auf der Heide and Nir Shavit, editors, *SPAA 2008: Proceedings of the 20th Annual ACM Symposium on Parallelism in Algorithms and Architectures, Munich, Germany, June 14-16, 2008*, pages 275–284. ACM, 2008. doi:10.1145/1378533.1378583.
- [VTRC11] Shivaram Venkataraman, Niraj Tolia, Parthasarathy Ranganathan, and Roy H. Campbell. Consistent and durable data structures for non-volatile byte-addressable memory. In Gregory R. Ganger and John Wilkes, editors, *9th USENIX Conference on File and Storage Technologies, San Jose, CA, USA, February 15-17, 2011*, pages 61–75. USENIX, 2011. URL: <http://www.usenix.org/events/fast11/tech/techAbstracts.html#Venkataraman>.
- [VTS11] Haris Volos, Andres Jaan Tack, and Michael M Swift. Mnemosyne: Lightweight persistent memory. *ACM SIGARCH Computer Architecture News*, 39(1):91–104, 2011.
- [ZFS<sup>+</sup>19] Y. Zuriel, M. Friedman, G. Sheffi, N. Cohen, and E. Petrank. Efficient lock-free durable sets. *PACMPL*, 3(OOPSLA):128:1–128:26, 2019.