Pot: Deterministic transactional execution

TIAGO M. VALE and JOÃO A. SILVA, NOVA LINCS, DI, FCT, Universidade NOVA de Lisboa
RICARDO J. DIAS, SUSE Linux GmbH and NOVA LINCS
JOÃO M. LOURENÇO, NOVA LINCS, DI, FCT, Universidade NOVA de Lisboa

This paper presents Pot, a system that leverages the concept of preordered transactions to achieve deterministic multithreaded execution of programs that use Transactional Memory. Preordered transactions eliminate the root cause of nondeterminism in transactional execution: they provide the illusion of executing in a deterministic serial order, unlike traditional transactions which appear to execute in a nondeterministic order that can change from execution to execution. Pot uses a new concurrency control protocol that exploits the serialization order to distinguish between fast and speculative transaction execution modes in order to mitigate the overhead of imposing a deterministic order. We build two Pot prototypes: one using STM and another using off-the-shelf HTM. To the best of our knowledge, Pot enables deterministic execution of programs using off-the-shelf HTM for the first time. An experimental evaluation shows that Pot achieves deterministic execution of TM programs with low overhead, sometimes even outperforming nondeterministic executions, and clearly outperforming the state of the art.

ACM Reference Format:
Tiago M. Vale, João A. Silva, Ricardo J. Dias, and João M. Lourenço. Pot: Deterministic transactional execution. *ACM Trans. Architec. Code Optim.* 13, 4, Article 52 (December 2016), 25 pages.
DOI: 10.1145/3017993

1. INTRODUCTION

Over the last decade, Transactional Memory (TM) [Herlihy and Moss 1993; Shavit and Touitou 1997] emerged as a viable mechanism to synchronize concurrent accesses to shared state due to an interesting trade-off between ease of use and performance. With TM, programmers specify which portions of code should be atomic (transactions) *without* worrying how to enforce such atomicity. A concurrency control protocol (implemented either in software (STM), hardware (HTM), or a mixture of both) enforces atomicity at runtime, providing the illusion that transactions execute one at a time. TM is becoming mainstream, as processors from Intel and IBM already provide support for HTM [Cain et al. 2013; Yoo et al. 2013], the GCC has experimental support for TM (using either STM or HTM) [Free Software Foundation 2014], and there is ongoing work in integrating TM language constructs in C/C++ [C++ Committee SG5 2015].

Although TM provides a simple programming model it inherits the nondeterministic behavior of multithreaded execution. Specifically, the order in which transactions appear to execute depends on the nondeterministic interleavings of threads at runtime, so different executions of the same program with the same inputs can yield different

---

New Paper, Not an Extension of a Conference Paper.
This work is supported by Fundação para a Ciência e Tecnologia, Ministério da Ciência, Tecnologia, e Ensino Superior, under grants SFRH/BD/84497/2012 and PEst/UID/CEC/04516/2013.
Authors' addresses: T. Vale, J. Silva, R. Dias, and J. Lourenço, Departamento de Informática, FCT/UNL, Quinta da Torre, 2829-516 Caparica, Portugal.

DOI: 10.1145/3017993

thread 1	**thread 2**
if *test it* **then**	
	change it
assume it	

(a) Atomicity violation.

thread 1	**thread 2**
	use it
initialize it	

(b) Order violation.

Fig. 1: Example of the most common concurrency bugs [Lu et al. 2008], with transactions in *italic*. In (a) the assumption of a predicate is not atomic with its test. In (b) thread 2 uses a resource before thread 1 initializes it.

outcomes. In this work we focus on building a TM system that ensures that data race-free programs execute according to a deterministic transaction serialization order.[1]

Having a system that ensures a deterministic transaction serialization order has at least two benefits: (1) we can execute multiple replicas of a multithreaded application for fault tolerance [Schneider 1990], and (2) it helps debugging, or prevents, the most common concurrency bugs [Lu et al. 2008]. Executing multiple replicas for fault tolerance relies on the assumption that correct replicas always yield the same outputs. With a deterministic transaction serialization order this assumption is *not* broken under multithreaded execution, so replicas do *not* need to fall back to sequential execution to ensure correctness. Consequently, replicas potentially make better use of the available resources such as multicore processors. Regarding concurrency bugs, Fig. 1 depicts the two most common concurrency bugs (amounting to 97% of the non-deadlock bugs) found in a study of 4 real-world applications [Lu et al. 2008], with transactions highlighted in *italic*. Fig. 1a shows an example of an atomicity violation. Thread 1 tests some predicate, and then executes code that assumes that it is true. Thread 2 executes code that changes the predicate's outcome. If thread 2 interleaves thread 1 after the predicate test, but before the "then branch," thread 1 will execute code that assumes the predicate is true while it is not, which can result in unexpected behavior. Fig. 1b shows an example of an order violation. Thread 1 initializes some resource that thread 2 uses, but at runtime thread 2 attempts to use the resource before thread 1 initializes it. These concurrency errors are sensitive to thread interleavings, and in the particular case of TM, only manifest themselves in particular transaction serialization orders. Since the transaction serialization order is nondeterministic, the errors are difficult to reproduce and debug. With a deterministic transaction serialization order, the aforementioned errors either manifest themselves in every execution, or not at all, greatly simplifying the developer's work.

In this paper we present Pot, a system that enables deterministic multithreading of TM-based applications. While existing work [Ravichandran et al. 2014] also ensures a deterministic transaction serialization order of TM-based applications, Pot: (1) performs better, which is important when executing multiple replicas for fault tolerance, (2) is equally helpful when dealing with the most common concurrency bugs such as atomicity and order violations, and (3) is applicable to both STM and HTM.

In principle, lock-based deterministic multithreading techniques [Bergan et al. 2010; Liu et al. 2011; Olszewski et al. 2009; Lu et al. 2014; Cui et al. 2013] could be used to achieve deterministic execution of STM programs (if, and only if, the STM concurrency control protocol is implemented using deterministic locks). However, such an approach has several drawbacks: (a) it cannot be applied to HTM, because the concurrency control is implemented in the hardware, (b) it fails to exploit the semantics of transactions to reduce the overhead of ensuring determinism, because determinism is enforced with locks, which are at a level of abstraction lower than transactions, and (c) many practi-

[1]This property is known as weak determinism [Olszewski et al. 2009].

cal STMs directly use atomic primitives such as compare-and-swap rather than locks. Instead, Pot uses the concept of *preordered transactions* as a principled approach to ensure a deterministic transaction serialization order. While traditional transactions provide the illusion of executing one at a time in *any* order, preordered transactions appear to execute in a specific, predefined, order.

To realize preordered transactions, Pot must address two key challenges: (1) guarantee that the predefined serial order is the same across executions, and (2) that the outcome of executing transactions is as if they executed serially in the predefined order. To ensure (1), Pot's *sequencer* assigns a sequence number to each new transaction. The sequence number reflects the transaction's place in a deterministic transaction serialization order. To ensure (2) efficiently, Pot executes transactions concurrently and relies on a new concurrency control protocol that guarantees that the outcome is equivalent to the order defined by the sequencer. Pot's concurrency control protocol relies on two key techniques: *ordered commits* and *transaction modes*. Ordered commits force transactions to commit according to the predefined serialization order. Transaction modes leverage the key insight that, at any given time, there is always one transaction that is "the next allowed to commit." Pot's concurrency control protocol executes that particular transaction as fast as possible, with virtually no concurrency control overhead (*fast mode*) while executing the other transactions using regular mechanisms to maintain correctness in the presence of the fast-mode transaction (*speculative mode*).

We built two Pot prototypes, one using STM and another using off-the-shelf HTM, and evaluate them with the popular STAMP benchmark suite [Minh et al. 2008] and STMBench7 [Guerraoui et al. 2007]. Our Pot STM implementation clearly outperforms the state of the art in STM-based deterministic execution while simultaneously achieving deterministic execution with low overhead, providing promising evidence that using both STM and determinism to ease multithreaded programming may be practical. To the best of our knowledge, Pot also advances the state of the art by enabling deterministic execution of off-the-shelf HTM-based multithreaded programs for the first time.

The rest of the paper is structured as follows. §2 presents Pot's design, namely its sequencer (§2.1) and concurrency control protocol (§2.2); §3 highlights the challenges and details our implementation of Pot in an STM and an off-the-shelf HTM system; §4 reports an experimental evaluation of Pot; we discuss the related work and conclude the paper in §5 and §6.

2. DESIGN

The standard Transactional Memory (TM) correctness criterion is opacity [Guerraoui and Kapalka 2008]. Traditional concurrency control protocols used to implement opaque transactions, such as Two-phase Locking [Bernstein et al. 1987] or Optimistic Concurrency Control [Kung and Robinson 1981], embrace opacity's flexibility and perform two tasks simultaneously while transactions are executing: (a) they compute the transaction serialization order (ordering), and (b) control the concurrent execution of transactions to respect that serialization order (concurrency control). Since ordering is intertwined with concurrency control, the final transaction serialization order depends on the nondeterministic interleavings that occur at runtime between transactions and thus varies from one execution to the next. We refer to this execution model as *traditional transactions*.

With *preordered transactions* the serialization order is independent of the interleavings that may occur between transactions because, unlike traditional transactions, preordered transactions already have a place in the serialization order before they are executed. Conceptually, preordered transactions have a two-phase execution model: (1) the *ordering phase* which defines every transactions' place in the serializa-

tion order, and (2) the *execution phase* where transactions execute concurrently in such a way that the outcome is equivalent to their sequential execution in the predefined order. Traditional concurrency control protocols *cannot* be used in the execution phase, because they implement both ordering and concurrency control. This paper proposes a novel concurrency control protocol that can be used in the execution phase (§2.2).

2.1. Ordering phase: Pot sequencer

A consequence of decoupling ordering and concurrency control is that both the ordering and execution phase, where concurrency control occurs, can be performed separately by two different components. Ordering is performed by a *sequencer* component that computes some total order over the set of all transactions.

At first glance it seems that the sequencer needs to know which transactions will execute ahead of time, but we can devise generic sequencers that compute the transaction order on-the-fly by defining an order over the application threads and deriving the transaction order from it. For example, take threads t and u, with transactions $(a; b; c)$ and $(d; e; f)$ in their code, respectively. Consider a sequencer that orders threads using a round-robin scheme, i.e. $(t; u)$. This sequencer defines the transaction order $(a; d; b; e; c; f)$. Now consider that thread t only executes transaction c depending on some condition. The condition may be defined over global state, thread-private state, or a mixture of both. If the condition is over global state, the respective state must have been read within a transaction, e.g. transaction b, so the condition is always tested over the state resulting from the order $(a; d; b)$, yielding a deterministic result.[2] If thread t decides not to execute transaction c the order is $(a; d; b; e; f)$. If thread t's logic is "execute c or g" instead, the order is $(a; d; b; e; g; f)$.

The only requirement of a generic sequencer that derives the transaction order from the thread order is that the events of starting and stopping threads must be processed deterministically by the sequencer with respect to the transaction order. To do so, since transactions appear to execute in a deterministic order, Pot treats thread start/stop events as if they are transactions. Take threads t and u, with transactions $(a; b; c)$ and $(d; e; f)$, respectively, where transaction b is the creation of a new thread v with transactions $(g; h)$. If we organize threads in a tree where the main thread is the root, the remaining threads are children of the thread that spawned them, and let the tree's post-order traversal specify the thread order, a round-robin sequencer defines the transaction order $(a; d; b; e; g; c; f; h)$.

It is also possible to use application-specific sequencers. For example, we may record the transaction commit order in a nondeterministic execution and then feed it to a sequencer to replay the recorded execution. We can also have sequencers that explicitly define a transaction order, e.g. $(a; b; c; d; e; f)$, but these need to take care because if a thread decides not to execute a transaction in the order then the program would hang waiting for it to execute. (We can detect this situation and abort the application with an error.)

Our design works best for workloads in which threads perform transactions regularly. Optimizing for workloads with very heterogeneous thread behaviors is an open problem left for future work.

2.2. Execution phase: Pot Concurrency Control

Transactions may execute once they go through the ordering phase. At the core of the execution phase is a concurrency control protocol that guarantees equivalence to the serialization order defined in the ordering phase. The straightforward way to imple-

[2]Assuming the only source of nondeterminism is the transaction serialization order. Techniques to deal with other sources, e.g. randomness, are complementary to this work.

```
1:  when txn_start(t)
2:
3:  when txn_write(t, o, v)
4:    deferred_update(o, v, W_t)
5:  when txn_read(t, o)
6:
    consistent_read(o, R_t, W_t)
    or abort
7:  when txn_commit(t)
8:    atomically
9:      if validate(R_t)
10:       writeback(W_t)
11:     else
12:       abort
```

(a) OCC.

```
1:  when txn_start(t, sn)
2:    sn_t ← sn
3:  when txn_write(t, o, v)
4:    deferred_update(o, v, W_t)
5:  when txn_read(t, o)
6:    consistent_read(o, R_t, W_t) or
    abort
7:  when txn_commit(t)
8:    wait until sn_c = pred(sn_t)
9:    if validate(R_t)
10:     writeback(W_t)
11:     sn_c ← sn_t
12:   else
13:     abort
```

(b) Speculative PCC.

```
1:  when sn_c = pred(sn_t)
2:    if validate(R_t)
3:      writeback(W_t)
4:    else
5:      abort
6:  when txn_write(t, o, v)
7:    direct_update(o, v)
8:  when txn_read(t, o)
9:    read(o)
10: when txn_commit(t)
11:   sn_c ← sn_t
```

(c) Fast PCC.

Fig. 2: Methodology to transform Optimistic Concurrency Control (OCC) into Pot Concurrency Control (PCC). Fig. 2a models a typical OCC transaction. Figs. 2b and 2c model a PCC transaction in speculative and fast mode, respectively. sn_c represents the sequence number of the last committed transaction. sn_t, R_t and W_t represent the sequence number, read set, and write set of transaction t, respectively.

ment such concurrency control protocol is to simply execute transactions sequentially. However this approach is clearly suboptimal as it does not take advantage of the inherent parallelism present in today's multicore architectures.

This section describes Pot Concurrency Control (PCC), a new protocol that executes transactions concurrently while guaranteeing equivalence to the serial order defined by the sequencer. We design PCC by modifying Optimistic Concurrency Control (OCC), which works as follows. An OCC transaction consists of one, or more, speculative executions. A speculative execution is divided into three phases: (1) the read phase, (2) the validation phase, and (3) the write phase. The read phase records the objects read by the transaction in the transaction's read set. Write operations do not modify the shared state; instead the transaction defers its updates and logs them in its write set. Therefore locations that are both read and modified occur in both the read and the write sets. After the read phase, the transaction undergoes a validation phase where it checks whether any concurrently committed transaction's updates overlap with its read set. If so the transaction is aborted to respect opacity, and can be retried; otherwise it proceeds to the next phase. Finally, the transaction enters the write phase where it atomically updates all objects in its write set with the values buffered during the read phase.

We have chosen OCC as the base for PCC because OCC is suitable for dynamic transactions, i.e. transactions for which it is very difficult (or even impossible) to identify their read/write sets in advance. Dynamic transactions are common in general-purpose TM-based programs due to aliasing and the unstructured nature of the heap. In fact, most STM and all existing HTM concurrency control protocols are optimistic.

Next, we present PCC incrementally. First, we describe the baseline OCC protocol in §2.2.1, and then present our methodology to transform the baseline OCC protocol into PCC by applying two key techniques: ordered commits, in §2.2.2, and transaction modes, in §2.2.3.

2.2.1. Baseline protocol. Consider the protocol depicted in Fig. 2a, modeling a typical OCC scheme [Dice et al. 2006; Kung and Robinson 1981]. The read phase occurs after txn_start and before either txn_commit or txn_abort, and consists of invo-

cations to txn_read and/or txn_write. Both the validation and write phase occur during txn_commit.

Read phase. Write operations intending to update object o's value to v, buffer the update in W_t (Fig. 2a, line 4). Read operations on an object o log the access in the transaction's read set R_t and return (a) the buffered value for o in the write set W_t, if existing, or (b) read a value of o from the shared state consistent with the rest of the read set (line 6). If it is not possible to read a consistent value the transaction aborts. For example, take two objects x and y, both initially 0. Transaction t observes $x = 0$. Meanwhile, another transaction commits and sets both x and y to 1. If transaction t attempts to read y it can either return 0 or abort, but it must never return 1, because $x = 0$ and $y = 1$ is not possible under opacity.

Validation phase. The validation phase iterates the read set and checks that the observed values are still coherent, i.e., all the observed values remain the same (line 9).

Write phase. If validation is successful then transaction t enters its write phase and directly updates the objects in its write set with the values buffered during the read phase, creating a new version of the shared state (line 10).

Correctness. This protocol guarantees opacity mainly due to the atomicity of the validation and write phases (lines 8–12). If the validation phase is successful then none of the read objects have been modified since the transaction's read phase. This means that the read phase happens in the same logical instant of the validation phase. Since the validation and write phase occur atomically, the write phase also happens in the same logical instant of the read phase. Therefore, transaction t appears to have been the sole transaction executing. Hence t is serialized after all the transactions that wrote the values t observed, and before any transactions that eventually observe the values t wrote.

2.2.2. Ordered commits. The OCC protocol described in the previous section provides the illusion that transactions execute one at a time. However, the order in which transactions appear to execute is not deterministic because it depends on the interleavings between transactions' operations that will occur at runtime.

To adhere to the serial order predefined in the ordering phase, we make two key observations: (a) OCC transactions only modify shared state during their write phase, and (b) each transactions' place in the serialization order depends on the relative order in which each transaction (atomically) performs its validation and write phase. If we restrict transactions to execute their validation and write phases in the order defined by the sequencer, we guarantee that the outcome is equivalent to the respective ordered sequential execution.

To transform the OCC protocol described in the previous section into PCC, we start by updating the txn_start operation to have an additional parameter, a sequence number sn, that reflects the order of transaction t in the serialization order defined by the sequencer (Fig. 2b, line 1). Transaction t is preordered after the transaction with sequence number $predecessor(sn_t)$ and before the transaction with sequence number $successor(sn_t)$. We force transactions to commit according to the predefined order by inserting a conditional wait in txn_commit. When transaction t wants to commit, it waits until the transaction with sequence number $predecessor(sn_t)$ commits (line 8). To this end, transactions communicate via a sn_c object whose value is the sequence number of the last committed transaction (line 11).

Correctness. In the original OCC protocol correctness is guaranteed by atomically executing both the validation and write phase. However, the order in which active transactions execute those phases depends on their nondeterministic multithreaded execution. To conform with the predefined order the atomic block is replaced with a conditional wait that restricts the order in which transactions are allowed to commit.

Specifically, a transaction t that finishes its read phase is only allowed to perform the validation and write phases *after* the transaction that directly precedes t in the serial order has completed. Since transactions are totally ordered, *only one transaction at a time* can escape the conditional waiting on line 8. Correctness is maintained because the conditional wait also guarantees atomicity. The atomicity scope is between the wait condition (line 8) and updating sn_c (line 11).

2.2.3. Transaction modes. OCC employs a set of techniques to guarantee correctness, such as read and write sets, read set validation and deferred updates. With OCC *all* transactions are executed using the aforementioned techniques because *any* transaction *may* become the next transaction in the serialization order, which is being defined as transactions execute. Using such techniques imposes additional overhead when compared with an execution without any concurrency control.

However, unlike in OCC, in PCC the serialization order is predefined. Since PCC restricts the order in which transactions commit, they may now have to wait for their turn to commit, leading to a loss of parallelism. To mitigate this loss of parallelism, we make the key observation that at any moment there is always a single transaction, which we refer to as *fast*, which is the next transaction that is allowed to commit. We exploit the fact that the fast transaction is the next transaction allowed to commit to execute it without most concurrency control overheads. Hence, we distinguish between two types of transactions: fast and speculative. We describe both fast and speculative modes below.

Fast transaction. A fast transaction t is the *only* active transaction whose predecessors are all completed. A fast transaction is the next, and only, transaction allowed to commit. It can be executed more efficiently by merging the read and write phases and completely removing the validation phase, thus eschewing most of the traditional OCC techniques and associated overhead. Fast transactions execute according to the protocol in Fig. 2c.

Read phase. Write operations no longer perform deferred updates; instead they use direct updates (line 7). Since updates are installed in place during the now combined read-write phase, read operations are reduced to simply reading the current object's value with no additional consistency checks or read set tracking (line 9).

Validation phase. Fast transactions are guaranteed to execute to completion without interference from other active transactions, thus the validation phase is unnecessary. (Transactions that switch on the fly to fast mode need to validate the speculative execution done up to that point; we elaborate below.)

Write phase. The write phase is implicitly executed during the read phase due to the direct update strategy, therefore the "write back" step is also completely eliminated.

Correctness. Our argument for correctness is the same as for the ordered commits technique. However a fast transaction does not speculatively perform the read phase and wait for its turn to transition to the validation and write phases. Instead the fast transaction executes the now combined read-write phase when it is already its turn to commit. A fast transaction is effectively given exclusive write permission to the shared state until it commits, so merging the read and write phases by replacing deferred with direct updates, and removing the validation phase, does not affect correctness.

Speculative transaction. A transaction whose turn to commit has not yet come is a speculative transaction, and it follows the ordered commit protocol (§2.2.2).

Live promotion. Since fast transactions bypass most concurrency control overhead, a live speculative transaction t, i.e. still executing its read phase, immediately switches to fast mode as soon as $sn_c = predecessor(sn_t)$ holds (line 1). Upon a live promotion, transaction t eagerly validates the portion of the read phase it has executed so far (line 2). If the validation is successful then t applies any pending writes to

the shared state, *without* updating sn_c, and executes its remaining operations in fast mode (line 3). Otherwise t aborts and retries in fast mode (line 5).

Explicit aborts. If the transaction API has an explicit txn_abort operation to abort the current transaction, fast transactions must keep the write set as an undo log, i.e. remember the values they overwrite to restore them upon abort. The txn_abort operation may allow the developer to specify a "no retry" policy, i.e. abort the transaction without retrying it afterwards. If so, these "no retry" aborts must comply with the predefined order as they are equivalent to committing the current transaction as read only. This is done by processing a "no retry" explicit abort as a commit. For example, a speculative tansaction waits for its turn, validates its read set, and updates sn_c if validation is sucessful, or retries if not. A fast transaction restores the write set (undo log) and updates sn_c.

Multiple simultaneous fast transactions. Multiple fast transactions can safely execute in parallel given additional knowledge about transactions. A string of successive transactions that do not have read-write nor write-write conflicts between themselves can all execute simultaneously as fast transactions, because the final outcome is independent of the order in which they commit. To implement multiple simultaneous fast transactions the runtime requires a compatibility matrix of all transactions. When a transaction becomes fast it publishes its information: transaction identifier, sequence number, and that it is active. Using this scheme, a transaction knows it can switch to fast mode if: (1) its predecessor is already fast (active or finished), and (2) it is compatible with all currently active fast transactions. If both conditions hold, the transaction can switch to fast mode.

3. IMPLEMENTATION

We implemented a Pot prototype consisting of an implementation of a sequencer and two concurrency control protocols: one where transactions execute using STM and another where transactions execute using HTM. Our sequencer implementation is generic and derives the transaction order from a round-robin thread order (§2.1). Next, we describe our STM (§3.1) and HTM (§3.2) implementations.

3.1. Software Transactional Memory

The ordered commits technique ensures that only one transaction executes its commit procedure at a time. In NOrec [Dalessandro et al. 2010] commits are also sequential. While this similarity makes NOrec a potential baseline for Pot, NOrec eschews per-memory location metadata and uses value-based validation instead. Consequently, speculative transactions are unable to identify which particular memory location is written when the fast transaction performs a write. As such, implementing fast transactions while still preserving opacity would require that, every time a fast transaction performs a write, all speculative transactions would have to validate their entire read set, regardless of which specific memory location was written by the fast transaction. Instead, our Pot STM protocol is based on TL2 [Dice et al. 2006], a popular STM that uses per-memory location metadata, so that speculative transactions do not have to perform incremental validation on reads.

Baseline STM transaction. In a nutshell, TL2 works as follows. There is a global version and a table of versioned locks, i.e., a version and a lock bit implemented as a single value—vlocks for short. Odd versions are locked and even versions are unlocked. Each memory address is mapped to one vlock. When a transaction starts, it samples the global version gv to rv_t and performs an acquire fence (Fig. 3a, lines 2–3). The transaction can safely read any value whose version is less than or equal to its rv_t sampling. The fence with acquire semantics ensures that this transaction observes all the memory writes performed by the transaction that updated gv's value

```
1:  when txn_start(t)
2:    rv_t ← gv
3:    acquire-fence
4:  when txn_write(t, addr, val)
5:    add (addr, val) to W_t
6:  when txn_read(t, addr)
7:    if addr ∈ W_t
8:      return W_t(addr)
9:    v1 ← get-vlock(addr)
10:   acquire-fence
11:   value ← read(addr)
12:   acquire-fence
13:   v2 ← get-vlock(addr)
14:   if unlocked(v1) ∧ v1 ≤ rv_t ∧ v1 = v2
15:     add addr to R_t
16:     return value
17:   else abort
18: when txn_commit(t)
19:   for each (addr, −) ∈ W_t
20:     if try-lock(addr) fails
21:       abort
22:   wv_t ← atomic-add-fetch(gv, 2)
23:   for each addr ∈ R_t
24:     v ← get-vlock(addr)
25:     if locked-by-other(v) ∨ v > rv_t
26:       abort
27:   for each (addr, val) ∈ W_t
28:     write(addr, val)
29:   release-fence
30:   for each (addr, _) ∈ W_t
31:     set-and-unlock(addr, wv_t)
```

(a) Original TL2.

```
1:  when txn_start(t)
2:    rv_t ← gv
3:    acquire-fence
4:    if first attempt
5:      wv_t ← get-seq-no(tid)
6:  when txn_write(t, addr, val)
7:    add (addr, val) to W_t
8:  when txn_read(t, addr)
9:    if addr ∈ W_t
10:     return W_t(addr)
11:   v1 ← get-version(addr)
12:   acquire-fence
13:   value ← read(addr)
14:   acquire-fence
15:   v2 ← get-version(addr)
16:   if v1 ≤ rv_t ∧ v1 = v2
17:     add addr to R_t
18:     return value
19:   else abort
20: when txn_commit(t)
21:   wait until gv = wv_t − 1
22:   acquire-fence
23:   for each addr ∈ R_t
24:     v ← get-version(addr)
25:     if v > rv_t
26:       abort
27:   for each (addr, val) ∈ W_t
28:     set-version(addr, wv_t)
29:     release-fence
30:     write(addr, val)
31:   release-fence
32:   gv ← wv_t
```

(b) Speculative PCC.

```
1:  when gv = wv_t − 1
2:    acquire-fence
3:    for each addr ∈ R_t
4:      v ← get-version(addr)
5:      if v > rv_t
6:        abort
7:    for each (addr, val) ∈ W_t
8:      set-version(addr, wv_t)
9:      release-fence
10:     write(addr, val)
11: when txn_write(t, addr, val)
12:   set-version(addr, wv_t)
13:   release-fence
14:   write(addr, val)
15: when txn_read(t, addr)
16:   return read(addr)
17: when txn_commit(t)
18:   release-fence
19:   gv ← wv_t
```

(c) Fast PCC.

Fig. 3: Pot Concurrency Control (PCC) STM implementation.

to rv_t. Write operations are buffered in the write set (line 5). Read operations return the value of a buffered write if there is any (line 7–8). Otherwise, they perform a consistent read by: (1) reading the address' vlock to $v1$ (line 9), (2) performing an acquire fence (line 10), (3) reading the memory address (line 11), (4) performing another acquire fence (line 12), and (5) reading the vlock again to $v2$ (line 13). The first fence ensures that the memory address value is at least as recent as $v1$. (If $v1$ is 42, then the value read has version 42 or newer.) The second fence ensures that if the value is newer than $v1$, then $v2$ is at least as recent as the value's version. (If the value read has version 43, $v2$ is 43 or newer.) If $v1$ is not locked, and $v1 \leq rv_t$, and $v1 = v2$, then the read successfully returns a consistent value; otherwise, the transaction aborts (lines 14–17).

The commit operation locks every address in the write set by performing a compare-and-swap on their vlocks. If any of the compare-and-swap operations fails, then the transaction releases any acquired locks and aborts (lines 19–21). After successfully acquiring the vlocks, the transaction performs an atomic add-and-fetch by 2 on gv and stores gv's new value in wv_t (line 22). Then, the transaction validates its read set by checking whether all memory addresses read are unlocked and their version is still compatible with rv_t. If any check fails then the transaction restores any acquired locks and aborts (lines 23–26). Note that the atomic add-and-fetch operation ensures that: (1) any other transaction that starts meanwhile and observes $gv = wv_t$ will at least observe all the write set vlocks as acquired, and (2) if any transactions committed since this transaction started, i.e. $wv_t > rv_t + 2$, and wrote to a memory address read by this transaction, then the read set validation will observe vlocks as locked or with a version newer than rv_t.

At this point the transaction successfully commits. It writes back any buffered writes, performs a release fence, and unlocks the write set, setting every vlock to wv_t. The release fence ensures that if any transaction observes a vlock with version wv_t then it also observes the value written by the transaction.

Speculative STM transaction. To implement PCC, we leverage the fact that TL2 uses a global version and retrofit sequence numbers directly as versions. Thus, transactions communicate the commit order via gv. A consequence of ordered commits is that we no longer require locks, just versions, as they were only needed due to concurrent commits.

When a transaction starts for the first time, it requests its sequence number wv_t from the sequencer by supplying the thread's identifier tid (Fig. 3b, lines 4–5). Read operations are similar to TL2 except that we no longer test if the address is locked (line 16). When the transaction attempts to commit, if necessary it waits until $gv = wv_t - 1$ (line 21). Once $gv = wv_t - 1$, we perform an acquire fence that ensures that the following read set validation observes the newest version of the addresses read (line 23–26). In the write back step, we first update the address' version, perform a release fence, and then write the new value (lines 27–30). As discussed before, the release fence ensures that if any transaction observes the written value, it also observes the new version number. Finally, the transaction updates gv, signaling the next transaction that it is its turn to commit (line 32). The update of gv is preceded by a release fence to ensure that all transactions that see the new value of gv will also see the new values for the objects written in the write back.

Fast STM transaction. The fast mode write operation is equivalent to the write back step of a speculative transaction, i.e. updates the version number, performs a release fence, and writes the new value (Fig. 3c, lines 12–14). The read operation is reduced to a regular load from memory (line 16), and the commit operation simply updates gv (line 19).

Live promotion. A speculative STM transaction t changes to fast on the fly when it detects that it is its turn, i.e. $gv = wv_t - 1$ (Fig. 3c, lines 1–10). In our implementation we check whether the condition holds whenever the speculative transaction begins, reads, or writes.

3.2. Hardware Transactional Memory

Implementing PCC in HTM poses unique challenges when compared with an STM implementation. Existing HTMs use the cache to maintain the read and write set, and rely on the cache coherence protocol to detect conflicts. HTMs are also *best effort*, i.e., hardware transactions are not guaranteed to eventually commit, even in the absence of conflicts, be it because the transaction's footprint exceeds the cache capacity, or due to the execution of an illegal instruction, an interrupt, a page fault, etc. Therefore, we must always provide a software fallback to guarantee progress. These characteristics pose three challenges, namely: (a) how to ensure that transactions eventually progress, (b) how to implement ordered commits without inducing false conflicts, and (c) how to implement fast transactions.

In our prototype we ensure progress using the most common fallback that achieves opacity: resorting to a global lock. Every time a transaction acquires the global lock, all hardware transactions abort and only retry when the lock is released.

In HTM, the commit operation is implemented entirely in hardware. This poses a challenge on how to implement ordered commits because we introduce conflicts if transactions signal each other whose turn it is to commit using a shared variable. For example, imagine two non-conflicting transactions t_1 and t_2, serialized in that order. Transaction t_2 attempts to commit before t_1. It reads the commit-order variable, sn_c, and observes that it is still not its turn so it waits, e.g., because t_2 can only

```
1:  when txn_start(t)
2:    if first attempt
3:      path_t ← HW
4:      tries_t ← 10
5:    wait while locked(gl)
6:    tbegin
7:    if locked(gl)
8:      abort
9:    execute app. code
10: when txn_abort(t)
11:   if persistent
12:     tries_t ← 0
13:   else
14:     tries_t ← tries_t − 1
15:   if tries_t = 0
16:     lock(gl)
17:     path_t ← SW
18:     execute app. code
19: when txn_commit(t)
20:   if path_t = HW
21:     tcommit
22:   else
23:     unlock(gl)
```

(a) Standard HTM.

```
1:  when txn_start(t)
2:    if first attempt
3:      path_t ← HW
4:      tries_t ← 10
5:      sn_t ← get-seq-no(tid)
6:    wait while locked(gl)
7:    if sn_c = sn_t
8:      tbegin(ROT)
9:      switch to fast mode
10:   else
11:     sn ← sn_c
12:     tbegin
13:     if locked(gl)
14:       abort
15:   execute app. code
16: when txn_abort(t)
17:   if persistent
18:     wait until sn_c = sn_t − 1
19:   else
20:     wait until sn_c > sn
21: when txn_commit(t)
22:   tsuspend
23:   wait until sn_c = sn_t
24:   tresume
25:   tcommit
26:   sn_c ← sn_t
```

(b) Speculative PCC.

```
1:  when txn_start(t)
2:    tbegin(ROT)
3:    execute app. code
4:  when txn_abort(t)
5:    if persistent
6:      tries_t ← 0
7:    else
8:      tries_t ← tries_t − 1
9:    if tries_t = 0
10:     lock(gl)
11:     path_t ← SW
12:     execute app. code
13: when txn_commit(t)
14:   if path_t = HW
15:     tcommit
16:   else
17:     unlock(gl)
18:   sn_c ← sn_t
```

(c) Fast PCC.

Fig. 4: Pot Concurrency Control (PCC) HTM implementation.

commit when $sn_c = 1$. When transaction t_1 commits it sets $sn_c = 1$, triggering a conflict in t_2 because it observed a (now) stale value. Implementing fast transactions is also challenging because all concurrency control is performed by the hardware.

To implement our prototype we looked at the existing HTMs from Intel [Yoo et al. 2013] and IBM [Cain et al. 2013]. To implement ordered commits without inducing false aborts we require the possibility to perform non-transactional accesses, i.e. that do not trigger transactional conflicts. Unfortunately, Intel provides no support for non-transactional accesses. However, IBM's HTM has two instructions, `tsuspend` and `tresume`, that allow the possibility to suspend, and resume, transactional execution inside a hardware transaction. (While in suspended mode accesses are performed non-transactionally.)

IBM's HTM also provides a special kind of transaction called Rollback-only Transaction (ROT). According to IBM, ROTs are intended to be used for single thread algorithmic speculation [Cain et al. 2013]. For this reason, ROTs also buffer transactional writes in the cache but do not maintain a read set. Furthermore, ROTs do not observe buffered transactional writes from other transactions and all writes performed by a ROT become visible to other transactions atomically, making them a prime choice to implement fast transactions. However, note that ROTs may nevertheless abort due to write-write conflicts with other concurrent transactions. For these reasons we implemented our prototype on IBM's HTM. It is still possible to implement Pot with Intel's HTM, albeit with ordered commits inducing false aborts and fast transactions being regular transactions.

Baseline HTM transaction. The relevant IBM's HTM instructions are `tbegin` and `tcommit`, to start and commit a hardware transaction, respectively. We initialize two important variables, $path_t$ and $tries_t$, when the application starts a transaction,

by invoking our usual txn_start operation. $path_t$ is either HW or SW depending on whether the transaction will execute as a hardware transaction or by software using the global lock fallback. $tries_t$ holds the number of remaining attempts to execute the transaction in hardware until we fallback to software (Fig. 4a, lines 2–4). (We retry 10 times like in GCC's experimental implementation.) If there is an ongoing transaction executing in software, we wait until the global lock is free, otherwise the hardware transaction may observe an inconsistent state and violate opacity (line 5). After the lock is free we start a hardware transaction by issuing the `tbegin` instruction (line 6). From this point on, every memory access is performed transactionally. Finally, we subscribe the global lock by checking if it is locked before proceeding with the actual application code (lines 7–8). By checking if the lock is taken it becomes part of the transaction's read set, so if any transaction falls back to software, any active hardware transaction is immediately aborted.

Committing a transaction depends on whether it executed in hardware ($path_t = \text{HW}$) or software (SW). We commit a hardware transaction using the `tcommit` instruction, whereas for a software transaction we simply release the global lock (lines 20–23). Note that the `tcommit` operation may still trigger an abort if the transaction fails to commit.

When a hardware transaction aborts, the control flow jumps to the txn_abort handler. First, we check whether the abort is expected to be persistent by inspecting the IBM's TEXASR register, which contains several hints about the reason why the transaction aborted. For example, an abort due to capacity restrictions is persistent. If the abort is persistent, we fallback to software by acquiring the global lock and execute the transaction's application code (lines 11, 15–18). Otherwise, we decrement the number of remaining attempts and control flow jumps back to txn_start.

Speculative HTM transaction. Like in our STM implementation, when a transaction starts for the first time it requests a sequence number from the sequencer (Fig. 4b, line 5). After waiting until the global lock is free, we check whether it is the transaction's turn to commit. If so, we begin a ROT and switch to fast mode (lines 8–9). Otherwise, we sample the sequencer number of the current fast transaction sn (we explain why shortly), then begin a hardware transaction, and subscribe to the global lock (lines 11–13). To commit a transaction we: (1) issue the `tsuspend` instruction to suspend transactional execution (line 22), (2) wait for our turn to commit (line 23), (3) issue the `tresume` instruction to resume transactional execution (line 24), and (4) issue the `tcommit` instruction to commit (line 25). If the commit is successful we update the sn_c variable accordingly (line 26).

If the speculative transaction aborts due to persistent reasons, there is no point in retrying the transaction until it is it's turn to commit (line 18). Otherwise, we wait until the concurrent fast transaction commits to retry the speculative transaction (line 20)—recall that we sampled its sequence number sn when we started the aborted hardware transaction (line 11). The rationale for waiting for the concurrent fast transaction to commit is to minimize the chances of aborting the fast transaction via write-write conflicts.

Fast HTM transaction. As previously stated, fast transactions execute as ROTs. Unlike regular hardware transactions, ROTs do not maintain a read set so they enjoy an increased capacity limit that can be used exclusively for writes. Transactions that previously exceeded capacity constraints and had to fallback to software might now be able to commit in hardware. This has the potential to increase the parallelism in the system because falling back to software effectively "stops the world."

Committing a fast transaction is essentially equivalent to the standard HTM transaction with an additional update to sn_c (Fig. 4c lines 14–18). If a fast transaction aborts due to capacity restrictions it falls back to software (lines 5–6, 9–12).

Benchmark	Parameters
Bayes	`-v 32 -r 4096 -n 10 -p 40 -i 2 -e 8 -s 1`
Genome	`-g 65536 -s 32 -n 16777216`
Intruder	`-a 10 -l 2048 -n 8192 -s 1`
Kmeans−	`-m 40 -n 40 -t 0.00001 -i inputs/random-n65536-d32-c16.txt`
KMeans+	`-m 15 -n 15 -t 0.00001 -i inputs/random-n65536-d32-c16.txt`
Labyrinth	`-i inputs/random-x512-y512-z7-n512.txt`
SSCA2	`-s 20 -i 1.0 -u 1.0 -l 3 -p 3`
Vacation−	`-n 8 -q 90 -u 98 -r 1048576 -t 4194304`
Vacation+	`-n 8 -q 10 -u 90 -r 1048576 -t 4194304`
Yada	`-a 15 -i inputs/ttimeu1000000.2`
STMBench7 (r)	`-t true -w r`
STMBench7 (rw)	`-t true -w rw`
STMBench7 (w)	`-t true -w w`

Fig. 5: Parameters used in STAMP and STMBench7.

Note that the hardware ensures that the fast transaction's reads do *not* observe the writes of concurrent speculative transactions. Moreover, if the fast transaction reads a memory location that has been written by a concurrent speculative transaction, the hardware aborts the speculative transaction immediately if it is executing, or when it issues the `tresume` instruction if it is suspended.

4. EXPERIMENTAL EVALUATION

All experiments were run on a 10-core IBM POWER8 with a total of 128GB RAM. We highlight the fact that the machine has a NUMA architecture. Particularly, the memory latencies in our experiments are as follows: with 1 to 4 threads memory latencies are uniform, while with 8 or more threads memory latencies increase up to $2\times$.

We evaluate Pot using the popular STAMP 0.9.10 benchmark suite [Minh et al. 2008] and STMBench7 [Guerraoui et al. 2007], using the parameters listed in Fig. 5. STAMP consists of 8 representative applications from different domains, e.g. online transaction processing, iterative clustering algorithms, and Delaunay mesh refinement (Vacation, KMeans, and Yada, resp.) [Minh et al. 2008]. Some STAMP benchmarks, such as Labyrinth, KMeans, and Yada, output non-deterministic results using STM. The benefits of Pot in these benchmarks are that the computed Labyrinth's solution, KMeans' clusters, and Yada's mesh, are always the same across executions. STMBench7 is a more complex benchmark suggestive of CAD, CAM or CASE software [Guerraoui et al. 2007]. Results are the average of five runs. The GCC version is Red Hat 5.1.1-4.

4.1. Software Transactional Memory

In this section we evaluate our Pot STM prototype. We seek to answer the following questions:

Are fast transactions effective? (§4.1.1.) Yes, they successfully reduce concurrency control overheads and execute faster than regular transactions. Our experiments show that fast transactions already execute faster than regular transactions even when they perform as little as 1 read and 1 write access, despite the addition work performed regarding the sequencer and switching modes (Fig. 6).

Does Pot ensure determinism efficiently? (§4.1.2.) We argue that it does. Our experiments show that Pot ensures deterministic execution across all of STAMP's benchmarks with an average slowdown over nondeterministic execution of less than $2\times$ (geometric mean of Fig. 7), and it is $\approx 5\times$ *faster* on average than the nondeterministic baseline in STMBench7 (geometric mean of Fig. 8).

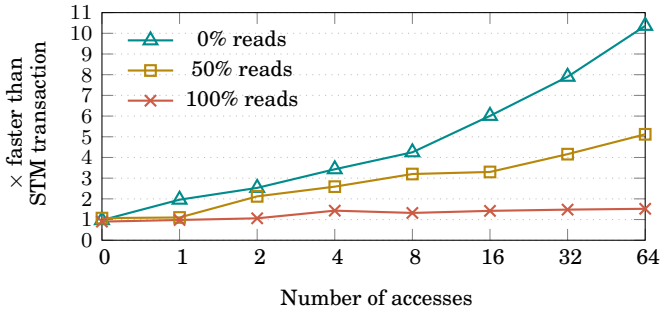


Fig. 6: Speedup achieved by a Pot fast transaction over the baseline STM transaction.

Does Pot improve upon the state of the art? (§4.1.2.) Yes, Pot successfully lowers the overheads of ensuring determinism when compared with DeSTM [Ravichandran et al. 2014]. Our experiments show that, when compared to DeSTM, Pot is up to $\approx 3\times$ faster than DeSTM on average across the STAMP benchmarks (geometric mean of Fig. 7) and up to $\approx 9\times$ faster on average in STMBench7 (geometric mean of Fig. 8), and scales better with the number of threads (Fig. 11 and 12).

4.1.1. Effectiveness of fast transactions. The fast transaction's objective is to reduce concurrency control overheads in order to mitigate the potential loss of parallelism introduced by ordered commits. To measure how effective is the fast execution mode we executed a microbenchmark that consists of a simple key-value data structure implemented with an array of counters. We use a single thread, and vary the number of accesses performed by transactions, and the accesses' read/write ratio.

Fig. 6 shows how much faster the Pot fast transaction protocol is than the baseline STM transaction. Transactions with 0 accesses consist of txn_begin immediately followed by txn_commit. This allows us to measure the overhead imposed by the additional work performed by the sequencer, ordered commits, and transaction modes, which is negligible. By increasing the number of accesses we observe that, as expected, fast transactions perform increasingly better than the baseline. We also observe that write operations contribute more to the achieved speedup. This is due to the fact that in the baseline STM write operations impose overhead on reads because reads must query the write set for possible buffered values. Write operations also impose overhead on the commit operation due to the need to lock the write set, perform the write back, and unlock the write set. Fast transactions bypass all these sources of overhead. However, fast transactions do not achieve observable gains when transactions are read-only. This is because read-only transactions in the baseline STM do not need to validate the read set at commit time—they are serialized at begin time. Overall, fast transactions are successful in minimizing concurrency control overheads, even for transactions that perform as little as 1 read and 1 write.

4.1.2. Comparison with the state of the art. In this section we evaluate deterministic execution using Pot in the popular STAMP 0.9.10 benchmark suite [Minh et al. 2008], and STMBench7 [Guerraoui et al. 2007]. We also compare Pot against DeSTM, a state of the art system in deterministic execution of STM programs. To perform an apples-to-apples comparison, we implemented DeSTM in our own prototype.[3] Both Pot and DeSTM are based on the same baseline STM protocol and use the exact same se-

[3]DeSTM is not publicly available. We asked the authors for the source code via e-mail but got no response.

quencer. See §5 for a comparison of Pot with DeSTM. We also implemented a deterministic and non-speculative solution based on a global lock that transactions acquire according to the order defined by the sequencer, i.e. transactions acquire a global lock at txn_begin and release it at txn_commit (PoGL, as in Preordered Global Lock). The rationale is that PoGL is a "trivial" implementation of PCC without any speculation. We show results for DeSTM, PoGL, ordered commits only (Pot$-$), ordered commits and transaction modes (Pot$*$), and ordered commits, transaction modes, and live promotion (Pot).

Performance. Fig. 7 quantifies the cost of deterministic multithreading when using DeSTM, PoGL, and Pot, on STAMP. With it we seek to answer the following question: "How much slower is the execution with x threads if we want determinism?" The Figure reports the execution time normalized to the baseline nondeterministic STM execution (y axis) of every benchmark of the STAMP suite, when executed with DeSTM, PoGL, Pot$-$ (ordered commits), Pot$*$ (ordered commits and transaction modes) and Pot (ordered commits, transaction modes, and live promotion) using from 2 to 16 threads (x axis). In these plots lower is better, and values below 1 mean that the deterministic execution was *faster* than standard nondeterministic execution. Four observations stand out: (a) the cost of ensuring determinism increases with the number of threads, (b) Pot outperforms DeSTM in all benchmarks, (c) Pot is at most $\approx 3\times$ slower than the nondeterministic baseline, while DeSTM suffers from a slowdown of up to $\approx 11\times$, (d) Pot is even *always faster* than the baseline STM execution on Genome, and (e) although PoGL works well in some workloads, Pot achieves the best of both worlds: Pot is comparable to PoGL on the workloads PoGL works well, and considerably outperforms PoGL on the remaining workloads (e.g. $\approx 2.5\times$ on Intruder, $\approx 3\times$ on Labyrinth and Vacation+, and $\approx 5\times$ on Vacation$-$).

The fact that the cost of ensuring determinism increases with the number of threads is unsurprising; the probability of a transaction t attempting to commit before its turn increases with the number of threads, particularly if there are transactions ordered before t that take longer than t. Pot's ordered commits and transaction modes minimize these situations to increase the probability of transactions not having to wait for their turn to commit. Fig. 9 supports this claim. It shows, for each benchmark/thread combination, how much time DeSTM transactions "waste" to enforce determinism, on average, when compared to Pot. We can observe that in general DeSTM transactions spend more time waiting for their turn to commit. Fig. 10 shows two example scenarios that highlight the differences between DeSTM and Pot. In DeSTM time is divided into rounds, and in each round each thread executes one transaction. A transaction cannot start if some transaction from the previous round has not finished yet (Fig. 10a), and cannot commit, even on its turn, if some transaction from the same round has not started yet (Fig. 10b). In contrast, Pot realizes that rounds are not necessary to respect a predefined serial order, so transactions never wait to start, nor to commit on their turn.

Pot also accelerates the execution of the next transaction to commit according to the serial order. From Fig. 6 we deduce that the benefits of the fast mode should be more apparent in benchmarks with bigger transactions with higher write-to-read ratio, and/or higher contention. Fast transactions (Pot$*$) further improve performance over ordered commits in all benchmarks (Fig. 7). However, the overhead of our implementation of live promotion (Pot) only pays off in Genome, Vacation+, and Yada.

We also experimented with STMBench7. Fig. 8 shows the throughput of DeSTM, PoGL, Pot$-$, Pot$*$, and Pot, normalized to the throughput of the baseline STM. Because STMBench7 features a more diverse set of transaction profiles, with more complex read-write transactions, live promotion is very effective at boosting Pot's through-

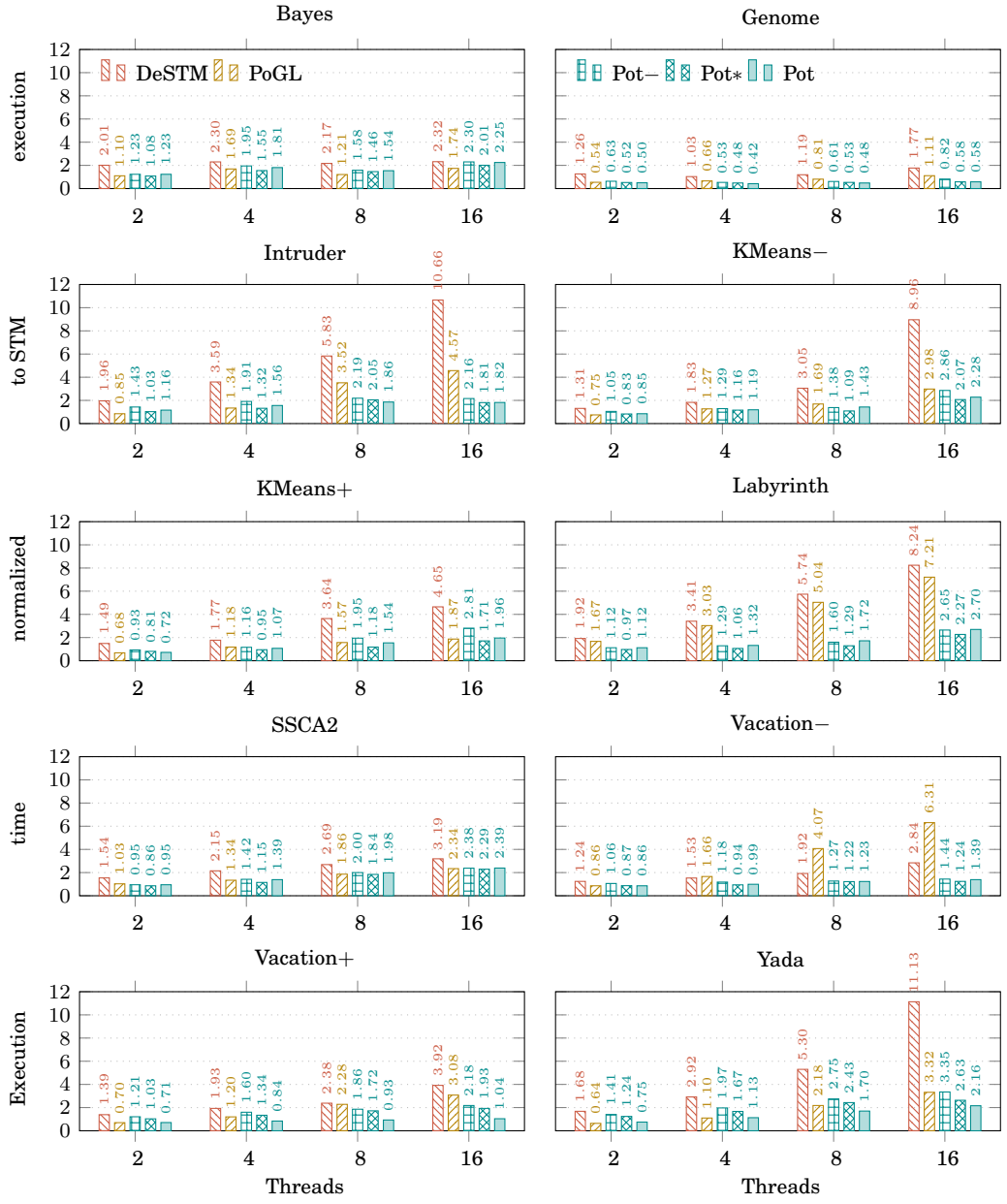


Fig. 7: **How much slower is the execution of each STAMP benchmark with x threads if we want determinism?** The y axis measures the execution time using DeSTM, PoGL (preordered global lock), Pot− (ordered commits), Pot∗ (ordered commits and transaction modes), and Pot (ordered commits, transaction modes, and live promotion), normalized to the nondeterministic execution using the baseline STM (lower is better). − and + refer to the relative levels of contention in the configuration.

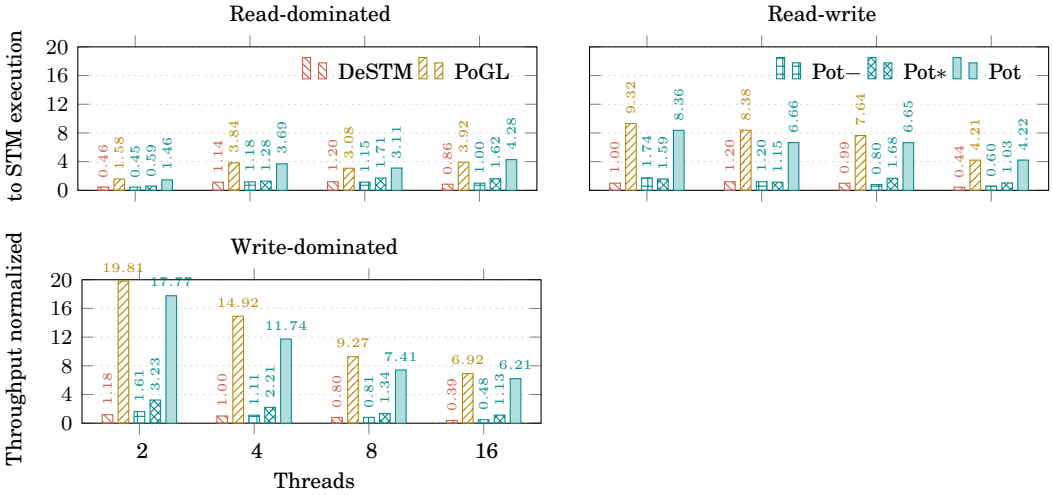


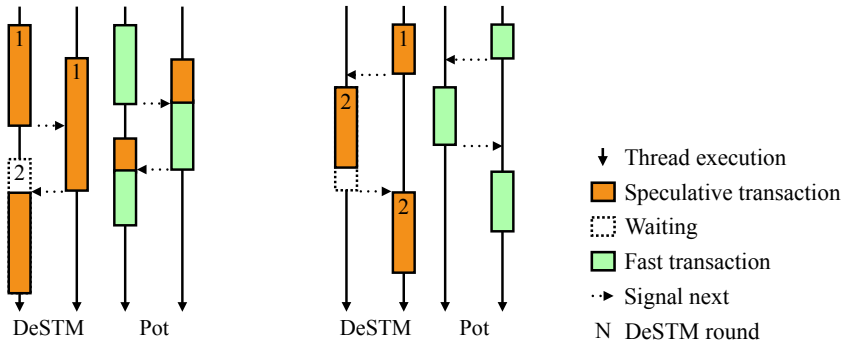
Fig. 8: **How much faster is the execution of STMBench7 with x threads if we want determinism?** The y axis measures the throughput using DeSTM, PoGL (pre-ordered global lock), Pot− (ordered commits), Pot∗ (ordered commits and transaction modes), and Pot (ordered commits, transaction modes, and live promotion), normalized to the nondeterministic execution using the baseline STM (higher is better). The titles indicate the workload type.

Benchmark	Threads			
	2	4	8	16
Bayes	1.88×	1.03×	0.95×	0.68×
Genome	3.24×	3.99×	3.92×	3.24×
Intruder	2.92×	3.11×	2.19×	1.74×
Kmeans−	4.16×	2.54×	2.22×	1.68×
KMeans+	3.62×	2.76×	1.97×	1.16×
Labyrinth	6.61×	5.31×	2.67×	0.77×
SSCA2	4.29×	1.62×	1.36×	1.34×
Vacation−	5.52×	4.01×	3.51×	3.60×
Vacation+	5.91×	4.93×	4.41×	5.29×
Yada	3.00×	3.26×	2.23×	1.69×
STMBench7 (r)	2.90×	3.53×	2.34×	4.73×
STMBench7 (rw)	8.02×	5.96×	7.16×	7.82×
STMBench7 (w)	15.10×	11.77×	9.17×	11.31×

Fig. 9: Time that DeSTM transactions spend waiting to enforce determinism compared to Pot, in STAMP. A value of $2\times$ means that, on average, DeSTM transactions spend $2\times$ more time waiting for their turn (hence higher is better for Pot).

put: in fact, Pot is *always faster* than the nondeterministic baseline, usually by more than $3\times$.

To conclude, in our experiments Pot is a good general solution because it achieves the best of both worlds: when speculation is effective Pot provides superior performance, and when speculation is not effective Pot's performance is very close to PoGL's (Figs. 7 and 8). Pot's excellent results compared to DeSTM's are explained by both the speedups that fast transaction can achieve, as observed in Fig. 6, and the decrease of the time transactions spend waiting for their turn, as we observe in Fig. 9.



(a) DeSTM waits to start. (b) DeSTM waits to commit.

Fig. 10: Examples of the difference between DeSTM and Pot. In DeSTM time is divided into rounds, and in each round each thread executes one transaction. A transaction cannot start if some transaction from the previous round has not finished yet (a), and cannot commit, even on its turn, if some transaction from the same round has not started yet (b). In contrast, Pot realizes that rounds are not necessary to respect a predefined serial order, so transactions never wait to start, nor to commit on their turn. Pot also accelerates the execution of the next transaction to commit according to the serial order.

Pot marks a significant advance over the state of the art in performance, and provides promising evidence that using both STM and determinism to enable multithreaded replicas for fault tolerance, and/or to ease multithreaded programming, may be practical.

Scalability. We further evaluate Pot's scalability compared to a singlethread execution using the baseline STM on STMBench7 and all of the STAMP benchmarks. For comparison we also show results for DeSTM and the baseline STM itself. The baseline's behavior serves as a guide for what to expect from Pot and DeSTM's implementation: we don't expect them to scale if the baseline does not scale. However, ideally we should expect the Pot and DeSTM implementation to scale, even if shyly, despite the overheads required to ensure determinism, particularly the need to wait to enforce the deterministic commit order. Figs. 11 and 12 show the results for STAMP and STMBench7, respectively. We observe that DeSTM fails to scale, whereas Pot is able to scale up to some point, notably in Genome, Intruder and Vacation. Pot shows better results than the baseline on STMBench7 because Pot inherently provides stronger progress guarantees to the more complex transactions in the benchmark: while they struggle to commit in the baseline STM, in Pot they eventually do when it is their turn, and even have their execution sped up by the fast mode.

As threads increase it becomes increasingly challenging to mask the overhead required to ensure determinism, but nonetheless Pot manages to keep up with the baseline up to a point. As part of future work we plan to address this issue by taking advantage of commutativity: if two successive transactions in the predefined serial order commute they can both execute simultaneously as fast transactions. The knowledge of whether two transactions commute can either be fed by the programmer via some sort of annotations, or inferred via analysis.

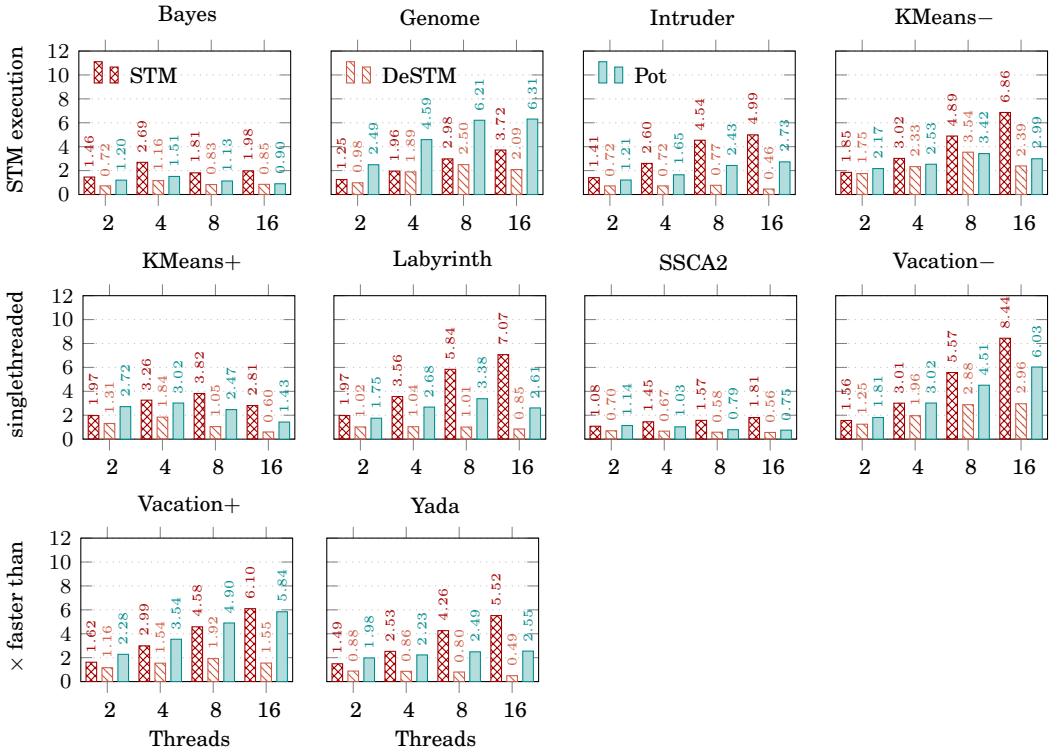


Fig. 11: Scalability of deterministic execution using DeSTM and Pot on STAMP. The y axis measures the speedup over a singlethread baseline STM execution. A value of 1 means the execution time was the same as the baseline, a value greater than 1 means the execution time was faster (better), and a value less than 1 means the execution time was slower (worse).

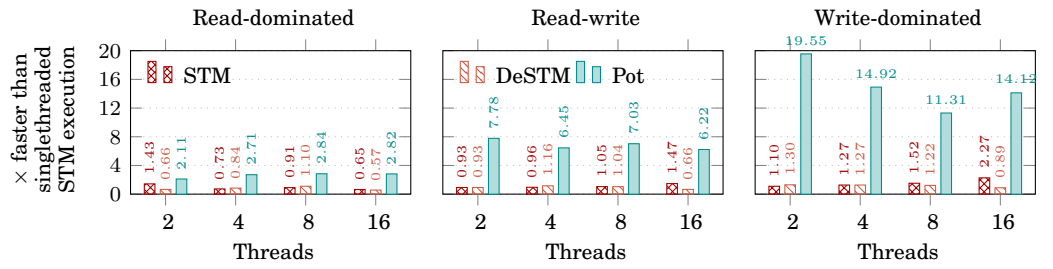


Fig. 12: Scalability of deterministic execution using DeSTM and Pot on STMBench7. The y axis measures the speedup over a singlethreaded baseline STM execution. A value of 1 means the throughput was the same as the baseline, a value greater than 1 means the throughput was greater (better), and a value less than 1 means the throughput was lower (worse). The titles indicate the workload type.

4.2. Hardware Transactional Memory

We also evaluate our Pot HTM implementation using the STAMP benchmark suite. We are interested in answering the following questions: (1) how effective are fast transactions, and (2) what is the cost that Pot incurs in to ensure deterministic execution.

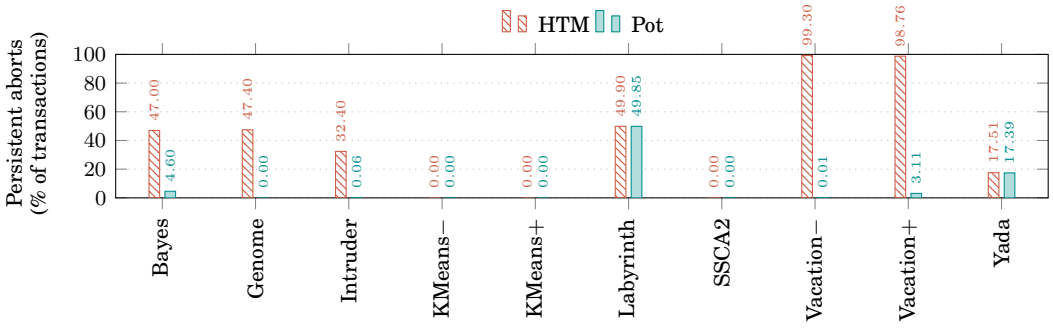


Fig. 13: Percentage of transactions that experience persistent aborts using baseline HTM transactions and Pot fast HTM transactions on the STAMP benchmarks (lower is better). − and + refer to the relative levels of contention in the configuration.

Are fast transactions effective? (§4.2.1.) Yes, Pot fast transactions enjoy increased capacity limits when compared to regular transactions. Our experiments show that for 4 of the STAMP benchmarks, Pot fast transactions greatly reduce the need to fall back to software (Fig. 13).
What is the cost that Pot incurs in to ensure determinism? (§4.2.2.) Our experiments show that Pot ensures deterministic execution across all of STAMP's benchmarks with moderate overhead (Fig. 14.)

4.2.1. Effectiveness of fast transactions. While our Pot STM fast transaction is able to reduce concurrency control overheads, implementing a HTM fast transaction that effectively reduces concurrency control overheads would require hardware support that is currently unavailable in existing processors. However, by exploiting IBM's Rollback-only Transactions (ROTs), Pot HTM fast transactions enjoy increased capacity limits, which increases the chance of committing more transactions entirely in hardware without falling back to the global lock.

We executed each benchmark with regular HTM and Pot using a single thread. Since there is only one thread executing there are no aborts due to concurrency, however transactions may still abort for spuriously; thus we only count aborts that the hardware hints to be persistent—we collect this information from the TEXASR register as we discuss in §3.2. Fig. 13 shows that the transactions that the baseline HTM cannot accommodate in both Labyrinth and Yada are also not accommodated by Pot's fast transaction. The transactions of KMeans and SSCA2, on the other hand, can all execute without problem. The rest of the benchmarks have a mix of transactions that can and cannot execute in hardware. In these we can clearly see the benefit of Pot's fast transactions: for example, in Bayes around 47% of the transactions can not be accomodated by the baseline HTM but this number falls to around 5% with Pot. Indeed, with Pot the number of transactions that are not accomodated by the hardware falls down from more than 30% to less than 5%. This means that Pot fast HTM transactions are successful at avoiding to fall back to the global lock. Thus, fast transactions manage to regain some of the parallelism lost to ordered commits when the baseline HTM falls back to software.

4.2.2. Performance. Fig. 14 shows the overhead of deterministic multithreading using Pot HTM. It has less overhead on benchmarks where the baseline often falls back to software (Bayes, Genome, Vacation). In Genome Pot always outperforms the nondeterministic execution. Vacation's results in Fig. 14 may seem unintuitive given than the

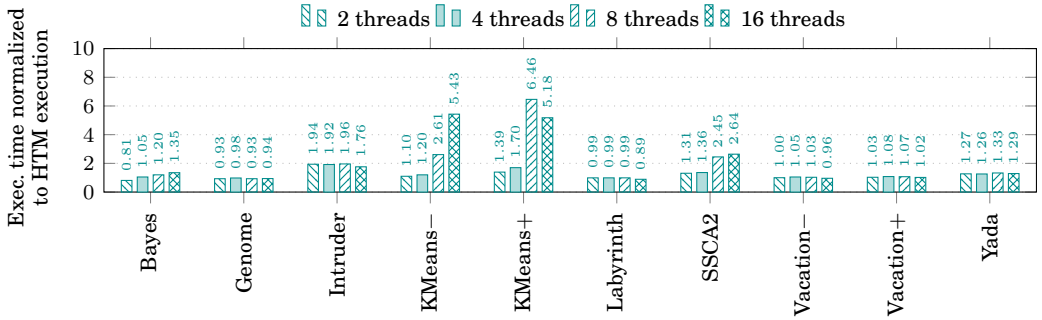


Fig. 14: Deterministic execution of STAMP using Pot. The y axis measures the execution time normalized to the nondeterministic execution using the baseline HTM (lower is better). − and + refer to the relative levels of contention in the configuration.

baseline HTM practically always falls back to the global lock while Pot mostly executes without resorting the global lock (Fig. 13). However, note that since all transactions executing in speculative mode exceed the hardware capacity, Pot is also executing one transaction at a time, albeit in fast mode instead of needing to fall back to the global lock.

Arguably the more interesting benchmarks are the ones where the baseline performs well, i.e. falls back less to the global lock (Intruder, KMeans, SSCA2, and Yada from Fig. 13). In Intruder and Yada, in Fig. 14, Pot achieves modest overheads of up to $2\times$. KMeans and SSCA2 are optimal for the baseline HTM, featuring small transactions with few accesses and conflicts. These characteristics make it difficult to mask the overheads of ensuring determinism. KMeans also features an abundant use of thread synchronization via barriers which amplifies the overhead caused by the sequencer while assigning sequence numbers deterministically. Also note that since fast transactions are not sped up in HTM, there is a noticeable drop in performance from 4 to 8 threads and even more from 8 to 16 threads due to the increased memory latencies of the NUMA architecture and hardware oversubscribing.

To the best of our knowledge, Pot advances the state of the art by enabling deterministic execution of HTM-based multithreaded programs for the first time. Overall, Pot achieves deterministic execution with lower overhead at lower thread counts, but increased memory latencies lead to a drop in performance relatively to the nondeterministic baseline. Efficiently achieving deterministic execution in the presence of non-uniform memory accesses represents an interesting future research avenue. The results achieved by Pot STM fast transactions suggest that hardware support for fast transactions that do not abort due to conflicts with other transactions may be worthwhile.

5. RELATED WORK

Preordered transactions. The idea of preordering transactions has also been used in the database community with the objective of reducing the costs of distributed transactions [Thomson et al. 2012]. In our work we advocate its use to achieve deterministic multithreading of TM-based programs. The aforementioned works also describe a concurrency control protocol to ensure that the predefined order is respect. However, that protocol could not be used in the context of our work, particularly in HTM, because it is tailored for transactions whose read/write sets can be determined a priori and is

based on Two-phase Locking. Ours is based on OCC and more general because it fully supports both transactions with static and dynamic read/write sets.

Deterministic multithreading. Many deterministic multithreading systems for lock-based programs have been proposed [Bergan et al. 2010; Liu et al. 2011; Olszewski et al. 2009; Lu et al. 2014; Cui et al. 2013]. If transactions are implemented with locks, deterministic transactions could be implemented using such systems. However this approach cannot be applied to off-the-shelf HTM, because the concurrency control is implemented in the hardware, and fails to exploit the semantics of transactions to reduce the overhead of ensuring determinism, because determinism is enforced with locks, which are at a level of abstraction lower than transactions. Both DeTrans [Smiljkovic et al. 2014] and DeSTM [Ravichandran et al. 2014] adapt the double barrier technique used by many deterministic lock-based systems to STM. DeTrans provides strong determinism [Olszewski et al. 2009] while DeSTM, like Pot, provides weak determinism [Olszewski et al. 2009]. One of the key differences between DeSTM and Pot is that in Pot the sequencer establishes a deterministic transaction serialization order that is enforced, i.e., the final outcome is as if transactions executed in the serial order defined by the sequencer. DeSTM, on the other hand, uses a token-passing scheme that defines a deterministic order in which threads attempt to commit transactions. Thus, the final outcome is always equivalent to the same transaction serialization order, although that order is unknown beforehand. As a consequence of this design, DeSTM orders both aborts and commits and requires conflicts to be deterministic. Pot only orders commits, and works whether conflicts are deterministic or not. Pot's design allows it to achieve better performance than DeSTM (which is important when replicating applications for fault tolerance), and from the application developers point of view, they are equally helpful with concurrency bugs such as atomicity and order violations. Furthermore, Pot's sequencer enables more uses cases, e.g. record-replay. However, if concurrency bugs lie in the STM implementation then DeSTM is better due to its requirement of deterministic conflicts. Grace [Berger et al. 2009] ensures deterministic execution of fork-join programs by using a custom STM that uses a technique similar to ordered commits. DMP-TM [Devietti et al. 2009] also ensures determinism using hardware transactions and ordered commits. Pot works with both STM and HTM, is not limited to fork-join-style parallelism, and takes advantage of the deterministic order to improve efficiency via fast transactions. DMP-TMFwd [Devietti et al. 2009] proposes to speculatively forward values written from transactions to their successors in the order. However current hardware does not support this functionality, and it is unclear how this can be done without violating opacity.

Parallelizing sequential code. FastPath [Spear et al. 2009], IPOT [von Praun et al. 2007] and TEPO [Gonzalez-Mesa et al. 2014] present a programming model to parallelize sequential code, e.g. loops, using transactions. Like DMP-TMFwd, they also order commits and propose to speculatively forward values. IPOT relies on unavailable hardware support to do so, TEPO does not implement it, and neither discusses how to do it while perserving opacity. Pot ensures deterministic execution of multithreaded TM programs, exploits the deterministic order using fast transactions, and preserves opacity.

Transaction modes. Executing transactions with the guarantee that they do not abort has been used to support I/O inside transactions [Spear et al. 2008; Welc et al. 2008], and to improve the performance of STMs at low thread counts [Wamhoff et al. 2013]. Pot fast transactions are similar in spirit to these works, however, their intent is to minimize the overhead of ensuring determinism. Unlike the aforementioned works, multiple fast transactions can safely execute in parallel by exploiting the existence of a predefined serialization order, e.g. a string of successive transactions that do not have read-write nor write-write conflicts between them can execute simultaneously as fast

transactions. PhTM [Lev et al. 2007] can execute transactions in different modes but only one mode is active at a time, while Pot executes different transactions in different modes simultaneously and transactions can switch modes at runtime.

6. CONCLUSIONS

We presented Pot, a system that uses the concept of preordered transactions as a principled approach to achieve deterministic multithreaded execution of transactions. At Pot's core is a novel concurrency control protocol that efficiently enforces a predefined transaction serialization order using two techniques: ordered commit and transaction modes. Pot advances the state of the art by: (1) to the best of our knowledge, enabling deterministic execution of off-the-shelf HTM-based multithreaded programs, and (2) clearly outperforming the state of the art in STM-based deterministic execution while simultaneously achieving determinism with low overhead, providing promising evidence that using both STM and determinism to enable multithreaded replicas for fault tolerance, and/or ease multithreaded programming, may be practical.

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for their helpful comments and suggestions, and to Koen De Bosschere for acting as a proxy for additional discussion with the reviewers. We also thank Tomas Vojnar, Daniel Horák, Jakub Cajka, Jaromir Capik, and the folks at Red Hat in Brno, Czech Republic, for providing us access to the infrastructure used in the evaluation. This research is supported by SFRH/BD/84497/2012 and PEst/UID/CEC/04516/2013.

REFERENCES

Tom Bergan, Owen Anderson, Joseph Devietti, Luis Ceze, and Dan Grossman. 2010. CoreDet: A compiler and runtime system for deterministic multithreaded execution. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. DOI:http://dx.doi.org/10.1145/1736020.1736029

Emery D. Berger, Ting Yang, Tongping Liu, and Gene Novark. 2009. Grace: Safe Multithreaded Programming for C/C++. In *ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*. DOI:http://dx.doi.org/10.1145/1640089.1640096

Philip Bernstein, Vassos Hadzilacos, and Nathan Goodman. 1987. *Concurrency control and recovery in database systems*. Addison-Wesley.

C++ Committee SG5. 2015. Technical Specification for C++ Extensions for Transactional Memory. open-std.org/jtc1/sc22/wg21/docs/papers/2015/n4514.pdf. (2015).

Harold Cain, Maged Michael, Brad Frey, Cathy May, Derek Williams, and Hung Le. 2013. Robust architectural support for transactional memory in the POWER architecture. In *International Symposium on Computer Architecture (ISCA)*. DOI:http://dx.doi.org/10.1145/2485922.2485942

Heming Cui, Jiri Simsa, Y. Lin, Hao Li, Ben Blum, Xinan Xu, Junfeng Yang, Garth Gibson, and Randal Bryant. 2013. Parrot: A practical runtime for deterministic, stable, and reliable threads. In *ACM Symposium on Operating Systems Principles (SOSP)*. DOI:http://dx.doi.org/10.1145/2517349.2522735

Luke Dalessandro, Michael F. Spear, and Michael L. Scott. 2010. NOrec: Streamlining STM by Abolishing Ownership Records. In *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*. DOI:http://dx.doi.org/10.1145/1693453.1693464

Joseph Devietti, Brandon Lucia, Luis Ceze, and Mark Oskin. 2009. DMP: Deterministic Shared Memory Multiprocessing. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. DOI:http://dx.doi.org/10.1145/1508244.1508255

Dave Dice, Ori Shalev, and Nir Shavit. 2006. Transactional locking II. In *International Symposium on Distributed Computing (DISC)*. DOI:http://dx.doi.org/10.1007/11864219_14

Free Software Foundation. 2014. Transactional memory in GCC. gcc.gnu.org/wiki/TransactionalMemory. (2014).

M. A. Gonzalez-Mesa, Eladio Gutierrez, Emilio L. Zapata, and Oscar Plata. 2014. Effective Transactional Memory Execution Management for Improved Concurrency. *ACM Transactions on Architecture and Code Optimization (TACO)* 11, 3 (2014). DOI:http://dx.doi.org/10.1145/2633048

Rachid Guerraoui and Michal Kapalka. 2008. On the Correctness of Transactional Memory. In *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*. DOI:http://dx.doi.org/10.1145/1345206.1345233

Rachid Guerraoui, Michal Kapalka, and Jan Vitek. 2007. STMBench7: A Benchmark for Software Transactional Memory. In *ACM European Conference on Computer Systems (EuroSys)*. DOI:http://dx.doi.org/10.1145/1272996.1273029

Maurice Herlihy and J. Eliot B. Moss. 1993. Transactional memory: Architectural support for lock-free data structures. In *International Symposium on Computer Architecture (ISCA)*. DOI:http://dx.doi.org/10.1145/165123.165164

H. Kung and John Robinson. 1981. On optimistic methods for concurrency control. *ACM Transactions on Database Systems (TODS)* 6, 2 (1981).

Yossi Lev, Mark Moir, and Dan Nussbaum. 2007. PhTM: Phased transactional memory. In *ACM SIGPLAN Workshop on Transactional Computing (TRANSACT)*.

Tongping Liu, Charlie Curtsinger, and Emery Berger. 2011. DThreads: Efficient deterministic multithreading. In *ACM Symposium on Operating Systems Principles (SOSP)*. DOI:http://dx.doi.org/10.1145/2043556.2043587

Kai Lu, Xu Zhou, Tom Bergan, and Xiaoping Wang. 2014. Efficient Deterministic Multithreading Without Global Barriers. In *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*. DOI:http://dx.doi.org/10.1145/2555243.2555252

Shan Lu, Soyeon Park, Eunsoo Seo, and Yuanyuan Zhou. 2008. Learning from Mistakes: A Comprehensive Study on Real World Concurrency Bug Characteristics. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. DOI:http://dx.doi.org/10.1145/1346281.1346323

Chí Cao Minh, JaeWoong Chung, Christos Kozyrakis, and Kunle Olukotun. 2008. STAMP: Stanford transactional applications for multi-processing. In *IEEE International Symposium on Workload Characterization (IISWC)*. DOI:http://dx.doi.org/10.1109/IISWC.2008.4636089

Marek Olszewski, Jason Ansel, and Saman Amarasinghe. 2009. Kendo: Efficient deterministic multithreading in software. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. DOI:http://dx.doi.org/10.1145/1508244.1508256

Kaushik Ravichandran, Ada Gavrilovska, and Santosh Pande. 2014. DeSTM: Harnessing Determinism in STMs for Application Development. In *International Conference on Parallel Architectures and Compilation Techniques (PACT)*. DOI:http://dx.doi.org/10.1145/2628071.2628094

Fred B. Schneider. 1990. Implementing Fault-tolerant Services Using the State Machine Approach: A Tutorial. *ACM Computing Surveys (CSUR)* 22, 4 (1990). DOI:http://dx.doi.org/10.1145/98163.98167

Nir Shavit and Dan Touitou. 1997. Software transactional memory. *Distributed Computing* 10, 2 (1997). DOI:http://dx.doi.org/10.1007/s004460050028

Vesna Smiljkovic, Srdan Stipic, Christof Fetzer, Osman Ünsal, Adrián Cristal, and Mateo Valero. 2014. DeTrans: Deterministic and Parallel Execution of Transactions. In *International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*. DOI:http://dx.doi.org/10.1109/SBAC-PAD.2014.20

M.F. Spear, M. Silverman, L. Dalessandro, M.M. Michael, and M.L. Scott. 2008. Implementing and exploiting inevitability in software transactional memory. In *International Conference on Parallel Processing (ICPP)*. DOI:http://dx.doi.org/10.1109/ICPP.2008.55

Michael F Spear, Kirk Kelsey, Tongxin Bai, Luke Dalessandro, Michael L Scott, Chen Ding, and Peng Wu. 2009. Fastpath speculative parallelization. In *Workshop on Languages and Compilers for Parallel Computing*. DOI:http://dx.doi.org/10.1007/978-3-642-13374-9_23

Alexander Thomson, Thaddeus Diamond, Shu-Chun Weng, Kun Ren, Philip Shao, and Daniel Abadi. 2012. Calvin: Fast distributed transactions for partitioned database systems. In *ACM International Conference on Management of Data (SIGMOD)*. DOI:http://dx.doi.org/10.1145/2213836.2213838

Christoph von Praun, Luis Ceze, and Calin Caşcaval. 2007. Implicit Parallelism with Ordered Transactions. In *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*. DOI:http://dx.doi.org/10.1145/1229428.1229443

J. Wamhoff, Christof Fetzer, Pascal Felber, Etienne Rivière, and Gilles Muller. 2013. FastLane: Improving performance of software transactional memory for low thread counts. In *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*. DOI:http://dx.doi.org/10.1145/2442516.2442528

Adam Welc, Bratin Saha, and A. Adl-Tabatabai. 2008. Irrevocable transactions and their applications. In *ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*. DOI:http://dx.doi.org/10.1145/1378533.1378584

Richard Yoo, Christopher Hughes, Konrad Lai, and Ravi Rajwar. 2013. Performance evaluation of Intel® transactional synchronization extensions for high-performance computing. In *International Conference for High Performance Computing Networking, Storage, and Analysis (SC)*. DOI:http://dx.doi.org/10.1145/2503210.2503232

Received May 2016; revised November 2016; accepted November 2016