Lerna: Parallelizing Dependent Loops Using Speculation

MOHAMED M. SAAD, Alexandria University, Egypt
ROBERTO PALMIERI, Lehigh University, USA
BINOY RAVINDRAN, Virginia Tech, USA

We present Lerna, an end-to-end tool that automatically and transparently detects and extracts parallelism from data dependent sequential loops. Lerna uses speculation combined with a set of techniques including code profiling, dependency analysis, instrumentation, and adaptive execution. Speculation is needed to avoid conservative actions and detect actual conflicts. Lerna targets applications that are hard-to-parallelize due to data dependency. Our experimental study involves the parallelization of 13 applications with data dependencies. Results on a 24-core machine show an average of 2.7x speedup for micro-benchmarks and 2.5x for the macro-benchmarks.

CCS Concepts: • **Theory of computation → Concurrency**; • **Computing methodologies → Parallel computing methodologies**; **Concurrent algorithms**.

Additional Key Words and Phrases: Code Parallelization, LLVM, Transactions

ACM Reference Format:
Mohamed M. Saad, Roberto Palmieri, and Binoy Ravindran. 2019. Lerna: Parallelizing Dependent Loops Using Speculation. *ACM Trans. Storage* 1, 1, Article 1 (January 2019), 24 pages. https://doi.org/10.1145/3310368

1 INTRODUCTION

Sequential code parallelization is a widely studied research field (e.g., [1, 28, 32, 68, 70]) that aims at extracting parallelism from sequential (often legacy) applications; it has gained particular traction in the last decade given the diffusion of multicore architectures as commodity hardware (offering affordable parallelism). Techniques for parallelizing sequential code are classified as manual, semi-automatic, and automatic. The classification indicates the amount of effort needed to rewrite/annotate the original application as well as the level of knowledge required on the codebase.

In this paper we focus on the automatic class, where the programmer is kept out of the parallelization process. We target sequential applications whose source code is no longer actively maintained, for these would benefit most from an automatic solution. In this class, effective solutions have been proposed in the past, but most assume that (or are well-behaved when) the application itself has no data dependencies, or dependencies can be identified [25, 29, 41] and handled prior the parallel execution [30, 68]. In practice, this supposes the possibility of identifying regions of the code that have no data dependencies [38, 59] through static analysis or that can be activated in parallel after having properly partitioned the dataset [43, 65, 68].

Static analysis of code is less effective if the application contains sections that could be activated in parallel but enclose computation that may affect their execution flow and accessed memory

Authors' addresses: Mohamed M. Saad, Alexandria University, Alexandria, 21526, Egypt, msaad@alexu.edu.eg; Roberto Palmieri, Lehigh University, Bethlehem, PA, 18015, USA, palmieri@lehigh.edu; Binoy Ravindran, Virginia Tech, Blacksburg, VA, 24061, USA, binoy@vt.edu.

1553-3077/2019/1-ART1 $15.00
https://doi.org/10.1145/3310368

locations. This uncertainty leads the parallelization process to take the conservative decision of executing those sections serially, nullifying any possible gain. For convenience, we say that an application has *non-trivial* data dependencies if accessed data cannot be partitioned according to the threads' access pattern, and therefore the application execution flow cannot be disjoint.

In this paper we follow the direction proposed by Mehrara *et al.* with the development of STMLite in [46]. STMLite consists of *speculating* over those sections to capture the actual data dependencies at run time, thus if the current execution does not exhibit data dependencies, parallelism will be exploited. We go beyond the original intuition in [46] presenting *Lerna*, an integrated software tool that parallelizes sequential applications with non-trivial data dependencies, automatically. The main difference between Lerna and STMLite (as well as many other existing solutions) is that, in the latter, programmer is required to annotate programs to be parallelized, Lerna deduces it with a combination of code profiling and refactoring. With Lerna, data-dependent applications can be parallelized and their performance increased thanks to the exploitation of parallelism on multicores.

In a nutshell, Lerna is a system that works with the source code's intermediate representation, compiled using *LLVM* [42], and produces ready-to-run parallel code. Its parallelization process is preceded by a profiling phase that discovers blocks of code that are prone to be parallelized (i.e., loops). After that, the refactoring step uses the Transactional Memory (TM) [32] abstraction to mark each possible parallel task. Such TM-style transactions are then automatically instrumented by Lerna to make the parallel execution equivalent to the serial execution. This is how Lerna handles the parallelization of data dependent tasks.

Despite the high-level goal, deploying the above idea leads application performance to be slower than the sequential, non-instrumented execution without fine-grain optimizations. As an example, a blind loop parallelization entails wrapping the whole body of the loop within a transaction. By doing so, we either generate an excessive amount of conflicts on variables depending on the actual iteration count or the level of instrumentation produced to guarantee a correct parallel execution becomes high. Also, variables that have never been modified should not be transactionally accessed; and local processing should be taken out from the transaction execution to alleviate the cost of abort/retry. Lerna's transactifier pass provides all of these. It instruments a small subset of code instructions, which is enough to preserve correctness, and optimizes the processing by a mix of static optimizations and runtime tuning.

We evaluated Lerna's performance using a set of 13 applications including micro-benchmarks, STAMP [10], a suite of applications, and a subset of the PARSEC [8] benchmark. Lerna is on average 2.7× faster than the sequential version using micro-benchmarks (with a peak of 3.9×), and 2.5× faster considering macro-benchmarks (with a top speedup of one order of magnitude reached with STAMP). These results have been collected on a 24-core machine.

Lerna is the first end-to-end automatic parallelization tool that exploits TM. Its main contribution is on the design and development of a solution that integrates novel (e.g., the ordered TM algorithms) and existing (e.g., the static analysis) techniques in order to serve the goal of parallelizing sequential applications with non-trivial data dependencies.

2 RELATED WORK

Automatic parallelization has been extensively studied in the past. The papers in [15, 23] overview some of the most important contributions in the area.

Optimistic concurrency techniques, such as Thread-Level Speculation and Transactional Memory, have been proposed as a means for extracting parallelism from legacy code. Both techniques split an application into sections and run them speculatively on parallel threads. A thread may buffer its state or expose it. Eventually, the executed code becomes safe and it can proceed as if it was

executed sequentially. Otherwise, the code's changes are reverted, and the execution is restarted. Some efforts combined TLS and TM through a unified model [7, 53, 54] to get the best of the two techniques.

Parallelization using thread-level speculation (TLS) has been studied using hardware [14, 31, 40, 66] and software [13, 20, 44, 46, 55]. It was originally proposed by Rauchwerger *et al.* [55] for parallelizing loops with independent data access – primarily arrays. The common characteristics of TLS implementations are that they largely focus on loops as a unit of parallelization, they mostly rely on hardware support or changes to the cache coherence protocols, and the size of parallel sections is usually small.

Regarding code parallelization and TM, Edler von Koch *et al.* [24] proposed an epoch-based speculative execution of parallel traces using hardware transactional memory (HTM). Parallel sections are identified at runtime based on binary code. The conservative nature of the design does not allow the full exploitation of all cores. Besides, relying only on runtime support for parallelization introduces a non-negligible overhead to the framework. Similarly, DeVuyst *et al.* [21] uses HTM to optimistically run parallel sections, which are detected using special hardware.

STMLite [46], shares the same sweet-spot we aim for; namely, applications with non-partitionable accesses and data dependencies. STMLite provides a low-overhead access by eliminating the need for locks and constructing a read-set; instead, it uses signatures to represent accessed addresses. A central transactional manager orchestrates the in-order commit process with the ability of having concurrent commits. In contrast with Lerna, it requires user interventions to support the parallelization and it has centralized components forming possible performance bottlenecks.

Sambamba [67] showed that static optimization at compile-time does not exploit all possible parallelism. It relies on user input for defining parallel sections. Gonzalez *et al.* [28] proposed a user API for defining parallel sections and the ordering semantics. Based on user input, STM is used to handle concurrent sections. In contrast, Lerna does not require special hardware, it is fully automated with an optional user interaction, and it improves the parallel processing itself with specific pattern-dependent (e.g., loop) optimization.

The study in [69] classified applications into: sequential, optimistically parallel, or truly parallel, and tasks into: ordered (speculative iterations of loop), and unordered (critical sections). It introduces a TM model that captures data and inter-dependencies. The study showed important per-application [6, 10, 12, 50] features as the size of read and write sets, dependency density, and size of parallel sections.

Most of the methodologies, tools and languages for parallelizing programs target scientific and data parallel computation applications, where the actual data sharing is very limited and the data-set is precisely analyzed by the compiler and partitioned so that the parallel computation is possible. Examples of those approaches include [34, 44, 49, 57]. ASC [70] is a system that automatically predicts the evolution of a program and whether the program produces jobs that have partitioned accesses. It does that by leveraging speculation and a fast-forwarding technique that shares cached values among threads running subsequent jobs. Lerna does not require the programmer and offers innovations effective when the application exposes data dependencies with non-partitionable access patterns.

The concept of providing memory transactions with ordering constraints has been also explored in design of TxOS+ [39], an operating system that implements system transactions. In TxOS+, the ordering constraint applies to the processing of incoming requests that access a shared state that is kept consistent using State Machine replication.

Lerna builds upon an initial concept named HydraVM [62]. HydraVM relies on a java virtual machine and uses transactions for parallelization. Unlike Lerna: HydraVM reconstructs the code at runtime through recompilation and reloading class definition. The extensive instrumentation for

establishing a relation between basic blocks and their accessed memory addresses limits its usage to small size applications and prevents the achievement of high performance. Lerna overcomes all the above limitations.

3 LERNA

3.1 General Architecture and Workflow

Lerna splits the code of loops into parallel *jobs*. For each job, we create a synthetic method that: *i)* contains the code of the job; *ii)* receives variables accessed by the job as input parameters; *iii)* and returns the *exit* point of the job (i.e., the point where the loop breaks). Synthetic methods are executed by separate threads as memory transactions, and our TM library is used for managing their contention. While executing, each transaction operates on a private copy of the accessed memory. Upon a *successful* completion of the transaction, all modified variables are exposed to the memory.

We define a *successful execution* of a job as an execution that satisfies the following two conditions:

1) it is reachable, meaning it is not preceded by a job that terminates early (e.g., using *break* instruction); and
2) it does not cause a memory conflict with any other job having an older chronological order.

Any execution of Lerna's parallel program is made of a sequence of jobs committed after a successful execution.

Lerna's core components are the following:

- an *automated software tool* that performs a set of transformations and analysis steps (*passes*) that run on the LLVM intermediate representation of the application code, and produces a refactored, multi-threaded version of the program;
- a *runtime library* that is linked dynamically to the generated program, and is responsible for the following: organizing the transactional execution of dispatched jobs so that the original program order (i.e., the chronological order) is preserved, selecting the most effective number of worker threads according to the actual deployment and the feedback collected from the online execution, scheduling jobs to threads based on threads' characteristics (e.g., stack size, priority), and performing memory and computational housekeeping.

Figure 1 shows the architecture and the workflow of Lerna. Lerna relies on LLVM, thus it does not require the application to be written in one specific programming language. In this paper we focus on the fully automated process without considering any programmer intervention; however, although automated, Lerna's design does not preclude the programmer from providing hints that can be leveraged to make the refactoring process more effective, which will be discussed separately in Section 6.

Lerna's workflow includes the following three steps in this order.

(1) *Code Profiling*. In the first step, our software tool executes the original application by activating our own profiler that collects some important parameters used later by the Static Analysis.
(2) The goal of the Static Analysis is to produce a multi-threaded (also called reconstructed) version of the input program. This process follows the passes below.
    - *Dictionary Pass.* It scans the input program to provide a list of the functions of the byte-code (or *bitcode* as in LLVM) that we can analyze to determine how to transform them. By default, any call to an external function is flagged as *unsafe*. This information is important because transactions cannot contain unsafe calls, such as I/O system calls.

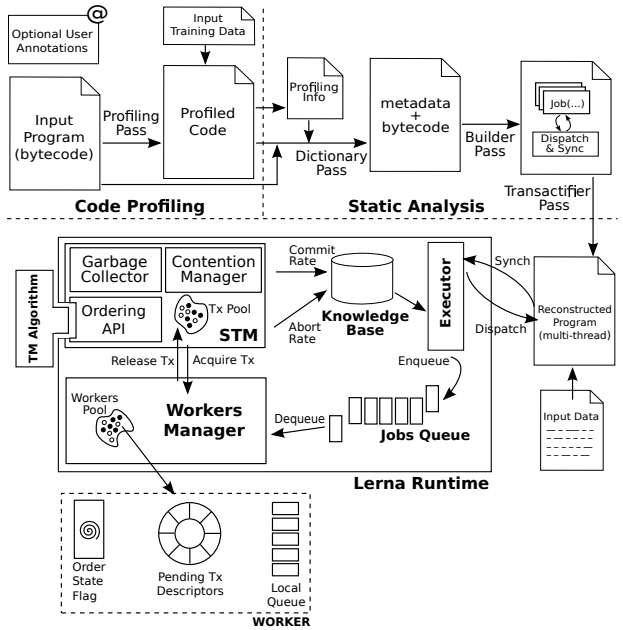


Fig. 1. Lerna's Architecture and Workflow

- *Builder Pass.* It detects the code eligible for parallelization, it transforms this code into a callable synthetic method, and it defines the transaction's boundaries, meaning the transaction's begin and end.
- *Transactifier Pass.* It applies the alias analysis [16] and detects depended memory operations in order to reduce the number of transactional reads and writes. It also provides the instrumentation of memory operations invoked within the body of a transaction by wrapping them into transactional calls for read, write, and allocate.

(3) Once the Static Analysis is complete, the reconstructed version of the program is linked to the application through the Lerna runtime library, which is mainly composed of the following three components:

- *Executor.* It dispatches the parallel jobs and provides the exit of the last job to the program. To exploit parallelism, the executor dispatches multiple jobs at-a-time by grouping them as a batch. Once a batch is complete, the executor simply waits for the result of this batch. Not all the jobs are enclosed in a single batch, thus the executor could need to dispatch more jobs after the completion of the previous batch. If no more job should be dispatched, the executor finalizes the execution of the parallel section.
- *Workers Manager.* It extracts jobs from a batch and it delivers ready-to-run transactions at available worker threads.
- *STM.* It provides the handlers for transactional accesses performed by executing jobs. In case a conflict is detected, it also behaves as a contention manager by aborting the conflicting transactions with the higher chronological order, this way the original program's order is respected. Lastly, it handles the garbage collection of the memory allocated by a transaction after it completes.

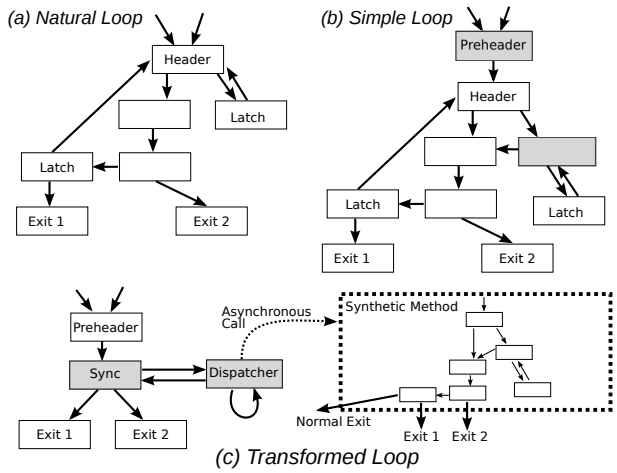


Fig. 2. Natural, Simple and Transformed Loop

The runtime library makes use of two additional components: the *jobs queue*, which stores the (batch of) dispatched jobs until they are executed; and the *knowledge base*, which maintains the feedback collected from the execution in order to enable the adaptive behavior.

3.2 Code Profiling

Lerna uses the code profiling technique for identifying hotspot sections of the original code, namely those most visited during the execution. For example, it would not be effective to parallelize a for-loop with only two iterations. To do that, we consider the program as a set of *basic blocks*, where each is a sequence of non-branching instructions that ends either with a branch instruction, conditional or non-conditional, or a return. Figure 2 shows an example of such a representation.

Our goal is to identify the context, frequency and reachability of each basic block. To determine that information, we profile the input program by instrumenting its LLVM byte-code at the boundaries of any basic blocks to detect whenever a basic block is reached. This code modification does not affect the behavior of the original program. We call this version of the modified program *profiled byte-code*.

3.3 Program Reconstruction

Here we illustrate in detail the transformation from sequential code to parallel made during the static analysis phase. The LLVM intermediate representation is in the Static Single Assignment (SSA) form. With SSA, each variable is defined before it is used, and it is assigned exactly once. Therefore, any use of such a variable has one definition, which simplifies the program analysis [51].

Dictionary Pass. In the dictionary pass, a full byte-code scan is performed to determine the list of accessible code (i.e., the dictionary) and, as a consequence, the external calls. Any call to an external function that is not included in the input program prevents the enclosing basic block from being included in the parallel code. However, the user can override this rule by providing a list of *safe* external calls. An external call is defined as *safe* if: *i)* it is revocable; *ii)* it does not affect the state of the program; and *iii)* it is thread safe. Examples of safe calls are stateless random generators.

Builder Pass. This pass is one of the core steps made by the refactoring process because it takes the code to transform along with the output of the profiling phase and makes it parallel by

matching the outcome of the dictionary pass. In fact, if the profiler highlights an often invoked basic block that contains calls not in the dictionary, then that block cannot be parallelized.

The actual operation of building the parallel code takes place after the following two transformations.

- *Loop Simplification analysis.* A *natural loop* has one entry block *header* and one or more back edges (*latches*) leading to the header. The predecessor blocks for the loop header are called *pre-header* blocks. We say that a basic block α dominates another basic block β if every path in the code β go through α. The *body* of the loop is the set of basic blocks that are dominated by its header, and reachable from its latches. The *exits* are basic blocks that jump to a basic block that is not included in the loop body. A *simple loop* is a natural loop with a single pre-header and single latch; and its index (if exists) starts from zero and increments by one. Loop simplification puts the loop into its simple form.
  Examples of a natural and a simple loop are reported in Figures 2 (a) and (b), respectively. In Figure 2 (a), the loop header has two types of predecessors, external basic blocks from outside the loop and body latches. Changing this loop in its simple form requires adding *i)* a single pre-header and changing the external predecessors to jump to the pre-header; and *ii)* an intermediate basic block to isolate the second latch from the header.
- *Induction Variable analysis.* An *induction* variable is a variable of a loop whose value changes by a fixed amount every iteration or is a linear function of another induction variable. Affine (linear) memory accesses are commonly used in loops (e.g., arrays, recurrences). The index of the loop, if any, is often an induction variable, and the loop can contain more than one induction variable. The *induction variable substitution* [52] is a transformation to rewrite any induction variable in the loop as a function of its index (i.e., closed form). It starts by detecting the candidate induction variables; it then sorts them topologically and creates a closed symbolic form for each. Finally, it substitutes their occurrences with the corresponding symbolic form.

As a part of our transformation, a loop is simplified, and its induction variable is transformed into its canonical form where it starts from zero and is incremented by one. A simple loop with multiple induction variables is a very good candidate for parallelization. However, induction variables introduce dependencies between iterations, which does not maximize parallelism. To solve this problem, the value of such induction variables is calculated as a function of the index loop prior to executing the loop body, and it is sent to the synthetic method as a runtime parameter.

Next, we extract the body of the loop as a synthetic method. The return value of this method is a numeric value representing the exit that should be used. Also, addresses of all accessed variables are passed as parameters.

The loop body is replaced by two basic blocks: *Dispatcher* and *Sync*. In *Dispatcher*, we prepare the arguments for the synthetic method, calculate the value of the loop index and invoke an API of our library, named *lerna_dispatch*, providing it with the address of the synthetic method and the list of the just-computed arguments. Each call to that API adds a job to our internal jobs queue, but it does not start the actual job execution. The *Dispatcher* keeps dispatching jobs until our API decides to stop. When this happens, the control passes to the *Sync* block. *Sync* immediately blocks the main thread and waits for the completion of the current jobs. Figure 2 (c) shows the control flow diagram for the loop before and after transformation.

Regarding the exit of a job, we define two types of exits: *normal exit* and *breaks*. A normal exit occurs when a job reaches the loop latch at the end of its execution. In this case, the execution should go to the header and the next job should be dispatched. If there are no more dispatched jobs to execute and the last one returned a normal exit, then the *Dispatcher* will invoke more jobs.

```
    Tx1                Tx2
Y[1] := 0          Y[2] := 0
Z[1] := X[1]       Z[2] := X[2]
```

(a) Symmetric Transactions

```
    Tx1                Tx2
Y[1] := 0
                   X[1] := Y[1]
Z[1] := X[1]
```

(b) Non-symmetric Transactions

Fig. 3. Symmetric vs Normal Transactions

On the other hand, when the job exit is a break, then the execution needs to leave the loop body, and hence ignore all later jobs. For example, assume a loop with N iterations. If the Dispatcher invokes B jobs before moving to the Sync, then $\lceil N/B \rceil$ is the maximum number of transitions that can happen between Dispatcher and Sync. In summary, the Builder Pass turns the execution into the job-driven model, which can exploit parallelism.

Transactifier Pass. After turning the byte-code into executable jobs, we employ additional passes to encapsulate jobs into transactions. Each synthetic method is demarcated by *tx_begin* and *tx_end*, and any memory operation (i.e., load, stores or allocation) within the synthetic method is replaced by the corresponding transactional handler.

It is quite common that memory reads are numerous, and outnumber writes, thus it would be highly beneficial to minimize those performed transactionally. That is because, the read-set maintenance and the validation performed at commit time, which iterates over the read-set to preserve transaction correctness, is the primary source of overhead.

In the case of code parallelization, for which Lerna is designed, all parallel transactions have the characteristic that the code to be executed is the same. In fact, when Lerna parallelizes a loop, any application code executed after the loop is postponed until the parallelization of the loop itself terminates. Therefore, there cannot be any transaction executing code that does not belong to the body of the parallelized loop. We name such transactions as *symmetric*. The transactifier pass makes use of this unique characteristic of having symmetric transactions by relaxing the need to support TM strong atomicity [3], and by eliminating unnecessary transactional reads as explained below. This improves performance because non-transactional memory reads are even three times faster than transactional reads [11, 64].

Clearly, local addresses defined within the scope of the loop are not required to be accessed transactionally. Global addresses allow iterations to share information, and thus they need to be accessed transactionally. We perform the *global alias analysis* as a part of our transactifier pass to exclude some of the loads to shared addresses from the instrumentation process. To reduce the number of transactional reads, we apply the global alias analysis between all loads and stores in the transaction body. A load operation that will never alias with any store operation does not need to be read transactionally. When a memory address is always loaded and never written in *any path* of the symmetric transaction code, then the load does not need to be performed transactionally. Note that this optimization can be applied only because our transactions are symmetric. Figure 3a shows an example of symmetric transaction while in Figure 3b we include an example of general transactions, which are inherently not symmetric.

Sometimes the induction variable substitution cannot produce a closed form of the loop index, if it exists. For example, if a variable is incremented (or decremented) based on any arbitrary condition. If the address value is used only after the parallelized loop, then it is eligible for the

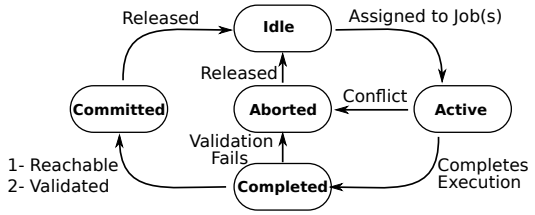


Fig. 4. Transaction States

Read-Modify-Write [58] optimization. With it, the increments (or decrements) are postponed at the transaction commit time. The modified locations are not part of the read-set, therefore, transactions do not conflict on these updates (see the Transactional Increment optimization in the next Section).

Transactions may contain calls to functions. As these functions may manipulate memory locations, they must be handled. If a function can be inline, e.g., it does not contain unpredictable escape branches, Lerna does that; otherwise we create a transactional version of the function called within a transaction. In that case, instead of calling the original function, we call its transactional version. Inlined functions are preferable because they permit the detection of dependent loops, or of dependencies between variables.

Finally, to avoid overhead in the presence of single-threaded computation or a single job executed at a time, we also keep a non-transactional version of synthetic methods.

3.4 Transactional Execution

The atomicity of transactions is mandatory, as it guarantees the consistency of the refactored code when run in parallel. However, if no additional care is taken, transactions commit in any order (i.e., the fastest commits first), and that could revert the chronological order of the program, which must be preserved to ensure application correctness because Lerna's parallelization process is entirely automated.

A transaction in general (and not in Lerna) is allowed to commit whenever it finishes. This property is desirable to increase thread utilization and avoid fruitless stalls, but it can easily lead to erroneous computation where: *i)* transactions make memory modifications as a result of the computation of an unreachable iteration; or *ii)* transactions executing iterations with lower indexes read future values from committed transactions. If materialized, these scenarios clearly violate the logic of the original sequential program.

Motivated by that, we propose an ordered transactional execution model based on the original program's chronological order. Lerna's engine for executing transactions works as follows. Each transaction has an *age* identifier indicating its chronological order in the program. Transactions can be in five states: *idle*, *active*, *completed*, *committed*, and *aborted*. Figure 4 shows these states and the transitions between them.

A transaction is idle because it is still in the transactional pool waiting to be attached to a job to dispatch. A transaction becomes active when it is attached to a thread and starts its execution. When a transaction finishes the execution, it becomes completed. That means that the transaction is ready to commit, and it completed its execution without conflicting with any other transaction; otherwise it would abort and restart with the same age. Note that, a transaction in the complete state still holds its lock(s) until its commit. Finally, the transaction is committed when it becomes reachable from its predecessor transaction. Decoupling completed and committed states, permits threads to process next transactions.

Preserving Commit Order. Lerna requires an enhanced TM design for enforcing a specific commit order (i.e., lower-age transactions must not observe the changes made by higher-age transactions). We identified the following requirements needed to support ordering:

Supervised Commit. Threads are not allowed to commit once they complete their execution. Instead, there must be a single stakeholder at-a-time that performs transaction commits, namely the *committer*. It is not necessary to have a dedicated committer because worker threads can take over this role according to their age. For example, the thread executing the transaction with the lowest age could be the committer, and thus it is allowed to commit. While a thread is committing, other threads can proceed by executing next transactions speculatively, or wait until the commit completes. Allowing threads to proceed with their execution is risky because it can increase the contention probability given that the life of an uncommitted transaction enlarges (e.g., the holding time of their locks increases); therefore, this speculation must be limited by a certain (tunable) threshold. Upon a successful commit, the committer role is delegated to the subsequent thread with lowest age. This strategy allows only one thread to commit its transaction(s) at a time.

An alternative approach is to use a single committer (as in [46]) to monitor the completed transactions, and to permit non-conflicting threads to commit in parallel by inspecting their read- and write-set. Although this strategy allows for concurrent commits, the performance is bounded by the committer execution time.

Age-based Contention Management. Algorithms should implement a contention manager that favors transactions with lower age because they have to commit earlier and unblock waiting subsequent transactions.

Lerna currently integrates four TM implementations with different designs: NOrec [19], which executed commit phases serially without requiring any ownership record; TL2 [22], which allows parallel commit phases at the cost of maintaining an external data structure for storing meta-data associated with the transactional objects; *UndoLog* [26] with visible readers, which uses encounter time versioning and locking for accessed objects and maintains a list of accessors transactions; and STMLite [46], which replaces the need for locking objects and maintaining a read-set with the use of signatures. STMLite is the only TM library designed for supporting loop parallelization.

Among TM algorithms, NOrec has some interesting characteristics which nominate it as the best match for our framework. This is because NOrec offers low memory access overhead with a constant amount of global meta-data. Unlike most STM algorithms, NOrec does not associate ownership records (e.g., locks or version number) with accessed addresses; instead, it employs a value-based validation technique during commit. It permits a single committing writer at a time, which matches the need of Lerna's concurrency control. Our modified version of NOrec decides the next transaction to commit according to the chronological order (i.e., age). Recently, in [61] the design of new TM algorithms that fit Lerna's requirement has been presented. Although Lerna does not currently integrate these algorithms, its architecture is flexible and enables integration with alternative STM implementations for code parallelization.

High-priority Transactions. A transaction performs a read-set validation at commit time to ensure that its read-set has not been overwritten by any other committed transaction. Let Tx_n be a transaction that has just started its execution, and let Tx_{n-1} be its immediate predecessor (i.e., Tx_{n-1} and Tx_n process consecutive iterations of a loop). If Tx_{n-1} has been committed before that Tx_n performs its first transactional read, then we can avoid the read-set validation of Tx_n when it commits because Tx_n is now the highest priority transaction at this time, thus no other transaction can commit its changes to the memory. We do that by flagging Tx_n as *high-priority* transaction. A transaction is high-priority if: *i)* it is the first and thus does not have a predecessor; *ii)* it is a retried

```
for(int i=0;i<100;i++){
    ...
    if(some_condition)
        counter++;
    ...
}
```

(a) Conditional increments

```
while(proceed) {
    ...
    counter++;
    ...
}
```

(b) No induction variable

Fig. 5. Conditional Counters

transaction of the single committer thread; *iii)* there is a sequence of transactions with consecutive age running on the same thread.

Transactional Increment. Figure 5 illustrates a common situation, which is the presence of a *counter* in the parallelized code. Loops with counters hamper achieving high parallelism because they create data dependencies between iterations, even non-consecutive iterations, hence producing a large amount of conflicts if not specifically handled. The induction variable substitution cannot produce a closed form (function) of the loop index (if it exists). If a variable is incremented (or decremented) based on any arbitrary condition and its value is used only after the loop completes the whole execution, then it is eligible for the *Transactional Increment* optimization.

In addition to the classical transactional read and write (tx_read and tx_write), we propose a new transactional primitive, the *transactional increment*, to enable the parallelization of loops with irreducible counters. This type of counter can be detected by Lerna during its code transformation process. Within the transactional code, a store S_t is eligible for our optimization if it aliases only with one load L_d, and it writes a value that is based on the return value of L_d. The load, change, and store operations are replaced with a single call to $tx\_increment$, which receives the address and the value to increment. We propose two ways to implement the $tx\_increment$ API:

- Using an atomic increment to the variable, and storing the address to the transaction's meta-data. The atomic operation preserves data consistency; however, it affects the shared memory before the transaction commits. To address this issue, aborted transactions compensate all accessed counters by performing the same increment but with the inverse value.
- By storing the increments into thread-local meta-data. At the end of each *Sync* operation, threads coordinate with each other to expose the aggregated per-thread increments of the counter. This method is appropriate for floating point variables, which cannot be updated atomically on commodity hardware.

Using this approach, transactions will not conflict on this address, and the correct value of the counter will be in memory after the completion of the loop.

3.5 Transaction Correctness and Sandboxing

Assessing correctness for our ordered STM implementations is trivial since the age-based contention management makes sure that transactions are validated and committed in a predefined order. Any execution that is not equivalent to the sequential (single-threaded) processing of the same set of transactions in the given predefined order will not successfully pass the validation and therefore will be aborted.

```
c = min;
while(i < max){
  i++;
  c = c + 5;
  //local processing
  if(i < j)
    k = k + c;
}
```

(a) Loop with data dependency

```
c = min;
while(i < max){
  atomic{
   TX_WRITE(i, TX_READ(i) + 1);
   TX_WRITE(c, TX_READ(c) + 5);
   //local processing
   if(TX_READ(i) < TX_READ(j))
     TX_WRITE(k, TX_READ(k)
       + TX_READ(c));
  }
}
```

(b) Loop with atomic body

```
c = min;
while(i < max){
  i++;
  parallel(i){
   c = min + i*5;
   //local processing
   atomic{
   if(i < j)
     TX_INCREMENT(k, c);
   }
  }
}
```

(c) Loop with parallelized body

Fig. 6. Lerna's Loop Transformation: from Sequential to Parallel.

However, there are important challenges related with the speculative execution of iterations that will never be committed due to a prior divergency in execution flow (e.g., a previous iteration issued a break statement). In this case, some iterations might speculatively process code that in a sequential execution will never be scheduled. Although these executions will not be ultimately committed, during the speculative computation they might process instructions with incorrect input, which can lead to erroneous states (e.g., division by zero). If not handled, errors might propagate to the invoking software components, which can hamper correct functionality of Lerna.

Lerna solves this problem by implementing a sandboxing environment that protects the speculative transactional execution. Timeouts are used to reclaim threads execution in case a transaction does not announce its completion after a certain amount of (tunable) time. In addition to that, OS signals are armed so that errors are caught by the transaction processing engine in Lerna with the result of reclaiming worker thread's execution. An implementation of sandboxing for STM can be found in [18].

4 SUMMARY OF CODE TRANSFORMATION

In Figure 6 we report a simple example of the entire code transformation used by Lerna to convert a sequential loop with dependency (Figure 6a), in a loop with a body that can be safely executed in parallel (Figure 6c). In the original sequential code we have programming patterns that are commonly used in developing sequential applications, such as variable increments, local processing that does not access shared variables, and counters. The same patterns would not be used by the programmer if the code was meant to run in parallel without data sharing. Through its transformation, Lerna catches those patterns and applies optimizations known in parallel computing to avoid needless sequential execution of iterations.

In Figure 6b we show the translation of the original code applying the transactional memory abstraction. The resulting code is extremely inefficient because, although transactions guarantee that any dependency between iterations is handled by executing the involved iterations sequentially, the presence of local processing inside the transactions or increment operations, either stretches the transactional execution resulting in higher abort rates, or induces unnecessary dependencies.

Lerna detects the possibility to refactor the update on c, as well as the increment on k. These simple modifications allows for removing the local processing from the transactional execution, which significantly speeds up performance and reduces the probability for this transaction to abort. The outcome of all the described transformation can be seen in Figure 6c.

5 DISCUSSION ABOUT CODE PARALLELIZATION USING HARDWARE TRANSACTIONAL MEMORY

Intel has recently introduced Haswell [56], the first processor family with hardware transactional memory support (HTM) [35]. HTM has the potential to significantly improve the transaction execution time in Lerna because HTM eliminates the overhead of transactional loads and stores. However, it introduces restrictions on the transaction execution in terms of memory footprint and the overall progress. For this reason, the current implementation of HTM is categorized as best-effort. These restrictions impose additional challenges whose solutions are not trivial. In this section we propose a possible direction to exploit HTM to execute transactions according to a predefined order.

5.1 Background on Intel HTM Implementation

Intel implemented HTM by providing programmers three new hardware instructions: `XBEGIN`, `XEND`, and `XABORT`. `XBEGIN` and `XEND` define the transaction start and end, respectively. `XABORT` intentionally aborts the executing transaction. HTM implements read-set and write-set by maintaining memory locations accessed transactionally into the processor cache (i.e., L1 for writes and L1 and L2 for reads). Only at the commit time the cache is flushed to main memory, which enables other threads to reach the newly written memory locations. Thanks to this approach, transactions are effectively executing atomically.

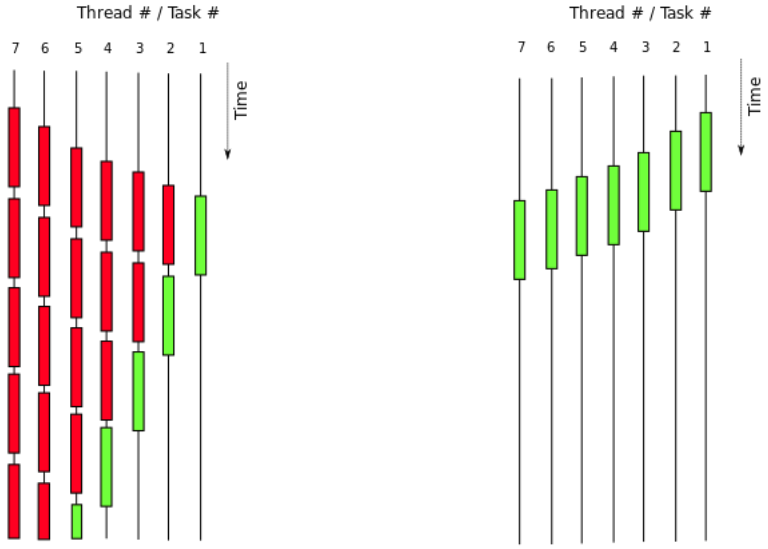
As a consequence of the above design, any invalidation triggered by some other thread and targeting some cache line accessed by an executing transaction triggers the transaction itself to abort. Also, conflict granularity is the cache line, therefore programmers should pay additional care to the application memory layout. Another side effect of using the L1 cache as transactional buffer is that, when no cache line is available in L1 (i.e., an eviction is needed), the executing transaction is forced to abort.

Roughly, a hardware transaction is aborted anytime an interrupt is received by the processor. On the one hand, this design allows for protection in case the transactional execution manifests undesirable behaviors (e.g., infinite loop); on the other hand, this design cannot provide a guaranteed forward progress for transactions [47], namely transactions might not be able to commit in hardware. To overcome the absence of progress, a fallback software path is used after the transaction is repeatedly aborted. However, since a transaction can be running in the fallback software path, it is required to synchronize its execution against all transactions executing in hardware for correctness [9, 17, 45, 47].

Although approaches like PartHTM [47] propose software solutions to overcome the best-effort nature of HTM; the fundamental problem, when applied to code parallelization, is that any shared metadata access is interpreted by the hardware transactional execution as a conflict that must be handled by aborting one transaction, even though no real shared data has been accessed. With such a constraint, enforcing a predefined order and retaining high performance is extremely challenging.

5.2 Challenges of Enforcing Commit Order Using HTM

Ordering transactions imposes exchanging information for coordinating threads on their commit order, which is a source of abort since HTM cannot differentiate between conflicts on data or metadata. To illustrate this, let us assume there is a shared variable that stores the `next-to-commit` transaction age. This variable will be checked when a transaction attempts to commit, and updated if the age is equals to its commit order. Let T_i be a transaction that is about to commit. If $i =$ `next-to-commit` then it will commit successfully. However, when $i \neq$ `next-to-commit` then the



(a) Worst Case Scenario (b) Best Case Scenario

Fig. 7. HTM executions using the `next-to-commit` technique.

address of the `next-to-commit` shared variable is now loaded in the processor cache, hence, it is effectively considered part of the transaction's read-set.

When the transaction with an age equals to `next-to-commit` commits, the `next-to-commit` variable will be updated. According to the *requestor wins* contention policy of HTM, transaction T_i (as well as all other transactions that read `next-to-commit` meanwhile) has to be aborted. In summary, an algorithm that uses a shared variable such as the `next-to-commit` to identify the commit order must abort (and retry) the transaction immediately when its age is not equals to `next-to-commit`. The STM implementations presented in Section 3.4 cannot be simply ported into the new HTM model because their performance will be severely affected.

Figure 7 illustrates the worst and the best case scenarios for a HTM-based parallelization algorithm that relies on the `next-to-commit` variable. Recall that reading the `next-to-commit` occurs only at the end of the transaction.

5.3 Ticketing-based Commit Order using HTM

We propose a direction to guarantee a specific commit order in HTM called *ticketing commit order* (or T-CO). This technique reduces the number of aborts due to the enforcing a commit order in HTM. T-CO leverages time as a dimension for transferring the ordering information across threads. In T-CO, each transaction is associated with a list of tickets. A ticket holds a boolean value: true (1) or false (0). To avoid cache sharing, tickets are allocated over different cache lines. Tickets are all initialized to false except the case of the tickets associated with the transaction with the lowest age, which are initialized with the true value since the transaction with the lowest age does not have any transaction that should commit before it.

When the transaction completes its transactional execution using HTM, it traverses its tickets in a deterministic order. The traversal involves a delay before moving to the next ticket in line. The goal is to look for one of them with a value equals to true. Upon finding a ticket with a true value, the transaction stops traversing and commits. If a ticket is found with a value equals to false, then

we say that the ticket has been *consumed*. If all the tickets were inspected and none of them was found true, transaction aborts and restarts.

Since the transaction with the absolute minimum age has all tickets with values equal to true, it commits as soon as it finishes the transactional execution. After its commit, the transaction iterates *in a reverse order* over the tickets of its *immediate successor* transaction, with the same delay.

This strategy has a twofold benefit. First, it allows a committed transaction to notify its successor to terminate its hardware execution without aborting it, as long as the higher age transaction does not consume all its tickets. Tickets that were set by the committed transaction do not affect the active transaction because they are not added yet to the transaction read-set yet. On the other hand, the committed transaction will not be affected by the memory operation that write 1 to tickets since this process is done outside the hardware transaction. Second, if a higher age transaction conflicts with a lower age transaction, then it will enter the ticket burning phase, which will delay it from a possible repeated abort and restart.

In the evaluation study of this paper Lerna has been evaluated without the assumption of the existence of hardware transactional memory implemented by the underlying computing machine. The integration, optimization, and evaluation of T-CO is left as future work but a preliminary evaluation of a STM implementation with T-CO is available in [60].

6 ADAPTIVE RUNTIME AND OPTIMIZATION

The Adaptive Optimization System (AOS) [5] is an architecture that allows for online feedback-directed optimizations. In Lerna, we apply the AOS to optimize the runtime environment by tuning some important parameters (e.g., the batch size, the number of workers) and by dynamically refining sections of code already parallelized statically according to the characteristics of the actual application execution.

The Workers Manager (Figure 1) is the component responsible for executing jobs. Jobs are evenly distributed over workers. Each worker keeps a local queue of its slice of dispatched jobs and a circular buffer of completed transactions' descriptors. It is in charge of executing transactions and keeping them in the completed state once they finish. As stated before, after the completion of a transaction, the worker can speculatively begin the next transaction. However, to avoid unmanaged behaviors, the number of speculative jobs is limited by the size of its circular buffer. Its size is crucial as it controls the transaction lifetime. A larger buffer allows a worker to execute more transactions, but it also increases the transaction lifetime, and thus the conflict probability.

For the non-dedicated committer algorithms, the ordering is managed by a per-worker flag called *state flag*. This flag is read by the current worker, but is modified by its predecessor worker. After completing the execution of each job, the worker checks its local state flag to determine if it is permitted to commit or proceed to the next transaction. If there are no more jobs to execute, or the transactions buffer is full, the worker spins on its state flag. Upon successful commit, the worker resets its flag and notifies its successor to commit its completed transactions. Finally, if one of the jobs has a break condition (i.e., not the *normal exit*) the workers manager stops other workers by setting their flags to a special value. This approach maximizes the cache locality as threads operate on their own transactions and access thread-local data structures, which also reduces bus contention.

Batch Size. The static analysis does not always provide information about the number of iterations; hence, we cannot accurately determine the best size for batching jobs. A large batch may cause many aborts due to unreachable jobs, meaning jobs that should not be executed given that the execution terminated in some prior job. However, small batches increase the number of iterations between the *dispatcher* and the *executor* and therefore the number of pauses to perform due to Sync. The current implementation, evaluated in Section 7, uses an exponentially increasing

batch size, meaning we add an exponentially increasing number of jobs to a batch until a predefined threshold that depends upon the number of executors in the system. Once a loop is executed, we record the last batch size used so that, if the execution goes back and calls the same loop, we do not need to perform again the initial tuning.

Jobs Tiling and Partitioning. Here we discuss an optimization, named *jobs tiling*, that allows the association of multiple jobs to a single transaction. Without tiling, a worker thread is assigned with a single job to execution. The transaction processing engine then executes this job as a transaction and, once completed, becomes ready for the next job to be assigned. In the presence of tiling, multiple jobs (e.g., multiple subsequent loop iterations) are combined in a single transaction to be executed by a single worker thread. Increasing the number of jobs in a single transaction allows for assigning enough computation to threads, which outweighs the cost of transactional management (e.g., initialization of metadata). However, abusing of tiling increases the size of read and write sets, which might degrade performance due to higher probability of abort. Tiling is a runtime technique; we tune it by taking into account the number of instructions per job, and the commit rate of past executions using the *knowledge base*.

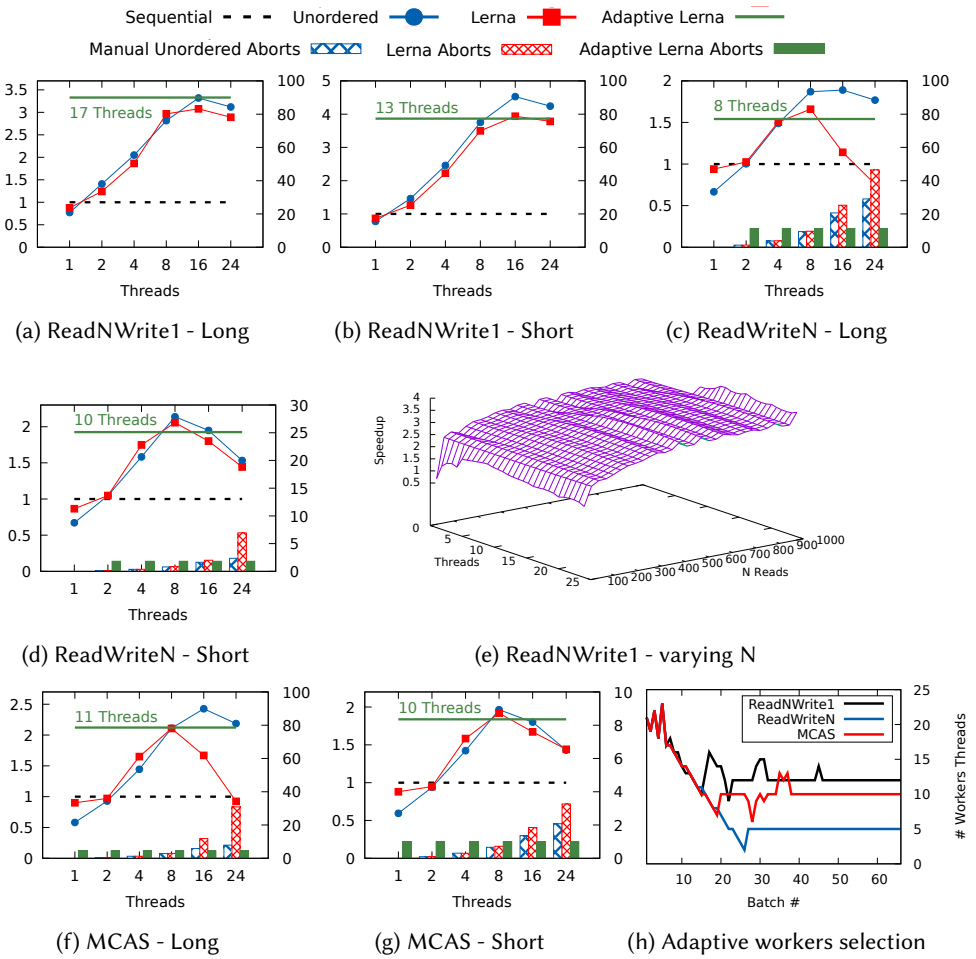
A similar known technique is *loop unrolling* [4], in which a loop is rewritten at compile time as a repeated sequence of its iteration code. Lerna employs static unrolling and runtime tiling according to the loop size. Figure 10 shows the impact in performance of tiling.

In contrast to tiling, a job may perform a considerable amount of non-transactional work. In this case, enclosing the whole job within the transaction boundaries makes the abort operation very costly. Instead, the transactifier pass checks the basic blocks with transactional operations and finds the nearest *common dominator* basic block for all of them. Given that, the transaction start (*tx_begin*) is moved to the common dominator block, and *tx_end* is placed at each *exit* basic block that is dominated by the common dominator. That way, the job is partitioned into non-transactional work, which is now moved out of the transaction scope, and the transaction itself, so that aborts become less costly.

Workers Selection. Figure 1 shows how the *workers manager* module handles the concurrent executions. The number of worker threads in the pool is not fixed during the execution, and it can be changed by the *executor* module. The number of workers affects directly the transactional conflict probability. The smaller the number of concurrent workers, the lower the conflict probability. However, optimistically increasing the number of workers can increase the overall parallelism (thus performance). In practice, at the end of the execution of a batch of jobs, we calculate the throughput and we record it into the *knowledge base* along with the commit rate, tiles and the number of workers involved. We apply a greedy strategy to find an effective number of workers by matching with the obtained throughput (Figure 8h in the evaluation section contrasts the variation of number of workers with batch size).

Finally, in some situations such as high contention or very small transactions, it is better to use a single worker. For this reason, if it is decided by our heuristic, then we use the non-transactional version as a fast path of the synthetic method to avoid the unnecessary overhead.

Read-Only Methods. The alias analysis technique (Section 3.3) helps in detecting dependencies between loads and stores; however, in some situations [16] it results in conservative decisions, which limit parallelization. It is non-trivial for the static analysis to detect aliases throughout nested calls. To assist the alias analysis, we try to inline the called functions within the transactional context. Nevertheless, it is common in many programs to find a function that does only loads of immutable variables (e.g., reading memory input). Marking that as read-only can significantly reduce the number of transactional reads, as we will be able to use the non-transactional version of the function.



(a) ReadNWrite1 - Long (b) ReadNWrite1 - Short (c) ReadWriteN - Long

(d) ReadWriteN - Short (e) ReadNWrite1 - varying N

(f) MCAS - Long (g) MCAS - Short (h) Adaptive workers selection

Fig. 8. Micro-benchmarks. In (a) to (f), left y-axis shows the speedup; right y-axis is the percentage of aborts.

7 EVALUATION

In this section we evaluate Lerna and measure the effect of the key performance parameters (e.g., job size, worker count, tiling) on the overall performance. Our evaluation involves 13 applications grouped into micro- and macro-benchmarks.

We compare the speedup of Lerna over the sequential, and the manual (*manual unordered* in the plots) transactional version of the code, when available. Note that the latter is provided directly by the benchmark (e.g., applications in micro-benchmarks and STAMP are already developed using transactions), thus it can leverage optimizations, such as the out-of-order completion, that cannot be caught by Lerna automatically (this is why we name it manual). Lerna's performance goal is twofold: providing a substantial speedup over the sequential code, and being as close as possible to the manual transactional version. It is worth noting that the reported performance of Lerna have been produced without any programmer intervention. The process is all automated.

The testbed consists of a server equipped with 2 AMD Opteron 6168 processors, each with 12-cores running at 1.9GHz. The memory available is 12GB and the cache sizes are 512KB for the

L2 and 12MB for the L3. On this machine, the overall refactoring process, from the results of the profiling phase to the generation of the binary, takes ~10s for simple applications and ~40s for the more complex ones.

7.1 Micro-benchmarks

We consider micro-benchmarks [2] to evaluate the effect of different workload characteristics, such as the amount of transactional operations per job and its length, and the read/write ratio. Figure 8 reports the speedup over sequential code by varying the number of used threads. We report two versions of Lerna: one adaptive, where the most effective number of workers is selected at runtime and we report the ending setting in the plot, and one with a fixed number of workers. We also reported the percentage of aborted transactions (right y-axis). To improve the clarity of the presentation, in the plots we report the best results achieved with the different TM algorithms (often NOrec). Each experiment includes running half a million transactions. For each micro benchmark, we configure two types of transactions: short and long. The short type has a random number of transactional accesses between 10 and 20; the long transactions simply produce more transactional accesses (i.e., a random between 30 and 60).

As a general comment, Lerna is very close to the manual transactional version. Unlike shown, the adaptive version of Lerna would never be slower than the single-threaded execution because, as fallback path, it would set the number of workers as one. The slow-down for the single thread is related to the fact that the thread adaptation is disabled for the competitor labeled “Lerna". Our adaptive version gains on average 2.7× over the original code and it is effective because it finds (or is close to) the configuration where the top performance is reached.

In *ReadNWrite1Bench* (Figures 8a, 8b, 8e), transactions read 1k locations and write 1 location. Thus, the transaction write-set is very small, and hence it implies a fast commit of a lazy TM as ours. The abort rate is low, and the transaction length is proportional to the read-set size. With long transactions, Lerna performs closer to the manual unordered; however, when transactions become smaller, the ordering overhead slightly outweighs the benefit of more parallel threads. In Figure 8e we vary the number of read locations and we report the achieved speedup over sequential varying the number of threads.

In *ReadWriteN* (Figures 8c and 8d), each transaction reads N locations, and then writes to another N locations. The large transaction write-set introduces a delay at commit time and increases the number of aborts. Both Lerna and manual unordered incur performance degradation at high numbers of threads due to the high abort rate (up to 50%). The commit phase of long transactions for Lerna forces some (ready to commit) workers to wait for their predecessor, thus degrading the overall performance. For that, the adaptive worker selection helps Lerna avoid this degradation.

MCASBench performs a multi-word compare and swap by reading and then writing N consecutive locations. Similar to ReadWriteN, the write-set is large, but the abort probability is lower than before because each pair of read and write acts on the same location. Figures 8f and 8g illustrate the impact of increasing workers with long and short transactions. Interestingly, unlike the manual unordered, Lerna performs better at single thread because it uses the fast path version of the jobs (non-transactional) to avoid any overhead.

Figure 8h shows the adaptive selection of the number of workers while varying the size of the batch. In the x-axis of the plot there is time represented by the number of batches executed by the transaction processing engine. The procedure starts by trying different worker counts within a fixed window (i.e., 7), then it picks the best according to the calculated throughput. Changing the worker count shifts the window looking for a more effective setting.

7.2 The STAMP Benchmark

STAMP [10] is a suite with eight applications covering different domains (Yada and Bayes have been excluded because of their non-deterministic behaviors). Figure 9 shows the speedup of Lerna's transformed code over the sequential code, and against the manual transactional version of the applications, which exploits unordered commits. Here the automatic tiling optimization is disabled.

Kmeans, a clustering algorithm, iterates over a set of points and associates them to clusters. The main computation is in finding the nearest point, while shared data updates occur at the end of each iteration. Using job partitioning, Lerna achieves 21× (Low contention) and 7× (High contention) speedup over the sequential code, using NOrec. Under high contention, NOrec is 3× slower compared to the manual unordered transactional version (more data conflicts and stalling overhead); however, they are very close in the low contention scenario. TL2 and STMLite suffer from false conflicts, which limits their scalability.

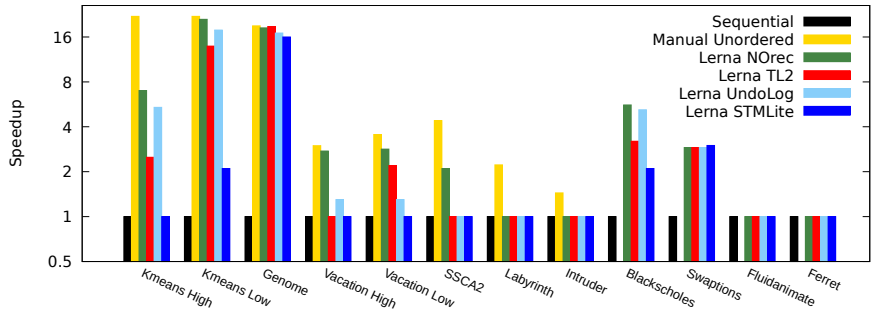


Fig. 9. STAMP & PARSEC Benchmarks Speedup. (y-axis in log-scale)

Genome, a gene sequencing program, reconstructs the gene sequence from segments of a larger gene. It uses a shared hash-table to organize the segments and eliminate duplicates. Lerna has 16-19× speedup over sequential. Genome conducts a large number of read-only transactions (*exists* operation), a friendly behavior for implemented algorithms. TL2 is just 10% slower than the manual competitor.

Vacation is a travel reservation system where the workload consists of clients reservation. This application emulated an OLTP workload. Lerna improves the performance by 2.8× faster than sequential, and it is very close to the manual.

SSCA2 is a multi-graph kernel application. The core of the kernel uses a shared graph structure updated at each iteration. The transformed kernel outperforms the original by 2.1× using NOrec,

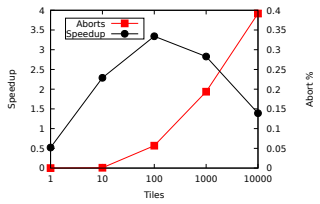


Fig. 10. Effect of Tiling using 8 workers and Genome.

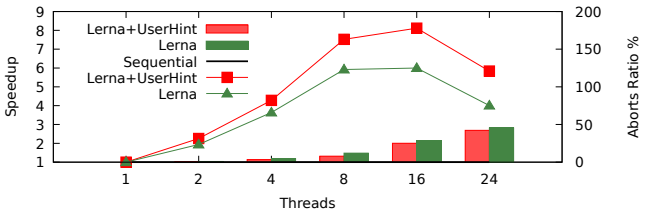


Fig. 11. Kmeans performance.

while dropping the in-order commit allows up to 4.4×. It is worth noting that NOrec is the only algorithm that manages to achieve speedup because it tolerates high contention and is not affected by false sharing as it deploys a value-based validation.

Lerna exhibits no speedup using *Labyrinth* and *Intruder* because, from the analysis of the application code, they use an internal shared queue for storing the processed elements and they access it at the beginning of each iteration to dispatch (i.e., a single contention point). While our jobs execute as a single transaction, the manual transactional version creates multiple transactions per iteration. The first iteration handles just the queue synchronization, while others do the processing. Adverse behaviors like this are discussed in later.

As explained in Section 6, selecting the number of jobs per each transaction (jobs tiling) is crucial for performance. Figure 10 shows the speedup and abort rate with changing the number of jobs per transaction from 1 to 10000 using the Genome benchmark. Although the abort rate decreases when reducing the number of jobs per transaction, it does not achieve the best speedup. The reason is that the overhead for setting up transactions nullifies the gain of executing small jobs. For this reason, we dynamically set the job tiling according to the job size and the gathered throughput.

The manual tuning further assists Lerna for improving the code analysis and eliminating avoidable overheads. An evidence of this is reported in Figure 11 where we show the speedup of Kmeans High against different numbers of workers using two variants of the transformed code: the first is the normal automatic transformation, and the second leverages user hints about memory locations that can be accessed safely (see Section 3.3). These results show that Lerna's framework can be deployed even more effectively if the programmer knows aspects of the original code.

7.3 The PARSEC Benchmark

PARSEC [8] is a benchmark suite for shared memory chip-multiprocessors architectures. For these applications, the manual unordered version is not included because PARSEC does not provide a transactional version of the code.

The Black-Scholes equation [37] is a differential equation that describes how the value of an option changes as the price of the underlying asset changes. This benchmark calculates Black-Scholes equation for different inputs. Iterations are relatively short, which generates many jobs in Lerna's transformed code. However, jobs can be tiled (Section 6). The speedup achieved is 5.6×. Figure 12 shows the speedup with different configurations of the loop unrolling.

Swaptions benchmark contains routines to compute various security prices using Heath-Jarrow-Morton (HJM) [33]. Swaptions employs Monte Carlo (MC) simulation to compute prices. The workload produced by this application provide similar speedup over sequential with all TM algorithms.

The following two applications have some workload characteristic that disallow Lerna to produce an effective parallel code. *Fluidanimate* [48] is an application performing physics simulations (about incompressible fluids) to animate arbitrary fluid motion by using a particle-based approach. The main computation is spent on computing particle densities and forces, which involves six levels of loops nesting updating a shared array structure. However, iterations updates a global shared matrix of particles, which makes every concurrent transaction conflict with its preceding transactions.

Ferret is a toolkit used for content-based similarity search. The benchmark workload is a set of queries for image similarity search. Similar to *Labyrinth* and *Intruder*, Ferret uses a shared queue to process its queries; which represents a single contention point and prevents any speedup with Lerna.

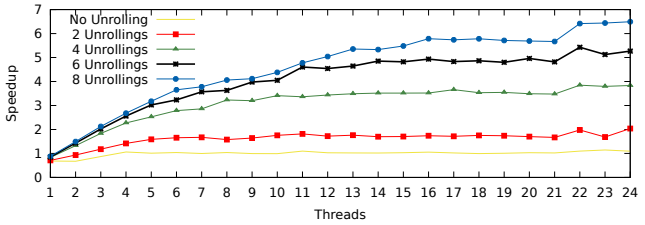


Fig. 12. Impact of unrolling using Black-Scholes.

7.4 Lerna Limitations

As confirmed by our evaluation study, there are scenarios where Lerna encounters obstacles that cannot be overcome due to the lack of semantics. Examples include data structure operations, loops with few iterations because the actual application parallelization degree is limited, there is an irreducible global access at the beginning of each loop iteration, and the workload is heavily unbalanced across iterations.

8 CONCLUSION AND FUTURE WORK

We presented Lerna, an end-to-end automated system that combines a tool and a runtime library to extract parallelism from sequential applications with data dependencies, efficiently. Lerna overcomes the pessimism of the static analysis of the code by exploiting speculation.

There is a number of interesting future directions to increase the effectiveness of Lerna further.

- The integration with HTM has the potential to increase performance significantly because reduces the instrumentation overhead. T-CO is just an initial step; other HTM implementations have the suspend operation to pause transactional work [27, 36]. Through suspend, hardware transactions can wrap the coordination logic to enforce the in-order commit therefore avoiding spurious aborts due to metadata access.
- Transaction processing has a strong semantics that forces an execution to be aborted every time there is a memory conflicts. As investigated earlier in [63], conflict detection can be improved by exploiting application semantics in order to reduce abort rate. Examples of leveraging application semantics include tracking shared memory locations that are only accessed conditionally (e.g., through a `if-condition`; it is safe to ignore a conflict on those memory locations as long as the result of the condition holds after a modification of the memory location value.

ACKNOWLEDGMENTS

A conference version of this work was published at 2018 ACM SYSTOR. The authors would like to thank SYSTOR's shepherd Prof. Margo Seltzer and the anonymous reviewers of SYSTOR and ACM ToS for their insightful reviews, feedback, and guidance.

This work, developed as part of the HydraVM project at Virginia Tech, is supported in part by Air Force Office of Scientific Research under grants FA9550-14-1-0187 and FA9550-16-1-0371. The authors gratefully acknowledge the highly insightful feedback from scientists at the US Naval Surface Warfare Center Dahlgren Division in developing the HydraVM project.

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-17-1-0367.

REFERENCES

[1] [n. d.]. Intel Parallel Studio. https://software.intel.com/\en-us/intel-parallel-studio-xe.

[2] [n. d.]. RSTM: The University of Rochester STM. www.cs.rochester.edu/research/synchronization/\rstm/.

[3] Martín Abadi, Tim Harris, and Mojtaba Mehrara. 2009. Transactional memory with strong atomicity using off-the-shelf memory protection hardware. In *ACM Sigplan Notices*, Vol. 44. ACM, 185–196.

[4] Alfred V Aho, Jeffrey D Ullman, et al. 1977. *Principles of compiler design*. Addision-Wesley Pub. Co.

[5] Matthew Arnold, Stephen Fink, David Grove, Michael Hind, and Peter F. Sweeney. 2000. Adaptive optimization in the Jalapeno JVM. In *Proceedings of the 15th ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications (OOPSLA '00)*. ACM, New York, NY, USA, 47–65. https://doi.org/10.1145/353171.353175

[6] David A Bader and Kamesh Madduri. 2005. Design and implementation of the HPCS graph analysis benchmark on symmetric multiprocessors. In *High Performance Computing–HiPC 2005*. Springer, 465–476.

[7] Joao Barreto, Aleksandar Dragojevic, Paulo Ferreira, Ricardo Filipe, and Rachid Guerraoui. 2012. Unifying thread-level speculation and transactional memory. In *Proceedings of the 13th International Middleware Conference*. Springer-Verlag New York, Inc., 187–207.

[8] Christian Bienia, Sanjeev Kumar, Jaswinder Pal Singh, and Kai Li. 2008. The PARSEC Benchmark Suite: Characterization and Architectural Implications. In *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques (PACT '08)*. ACM, New York, NY, USA, 72–81. https://doi.org/10.1145/1454115.1454128

[9] Irina Calciu, Tatiana Shpeisman, Gilles Pokam, and Maurice Herlihy. 2014. Improved single global lock fallback for best-effort hardware transactional memory. In *9th Workshop on Transactional Computing (TRANSACT '14)*. Available: http://transact2014.cse.lehigh.edu/.

[10] Chi Cao Minh, JaeWoong Chung, Christos Kozyrakis, and Kunle Olukotun. 2008. STAMP: Stanford Transactional Applications for Multi-Processing. In *IISWC '08: Proceedings of The IEEE International Symposium on Workload Characterization*.

[11] Chi Cao Minh, Martin Trautmann, JaeWoong Chung, Austen McDonald, Nathan Bronson, Jared Casper, Christos Kozyrakis, and Kunle Olukotun. 2007. An Effective Hybrid Transactional Memory System with Strong Isolation Guarantees. In *Proceedings of the 34th Annual International Symposium on Computer Architecture*.

[12] B Chan. 2002. The UMT Benchmark Code. *Lawrence Livermore National Laboratory, Livermore, CA* (2002).

[13] Michael Chen and Kunle Olukotun. 2003. TEST: a tracer for extracting speculative threads. In *Code Generation and Optimization, 2003. CGO 2003. International Symposium on*. IEEE, 301–312.

[14] Michael K Chen and Kunle Olukotun. 2003. The Jrpm system for dynamically parallelizing Java programs. In *Computer Architecture, 2003. Proceedings. 30th Annual International Symposium on*. IEEE, 434–445.

[15] Doreen Y Cheng. 1993. A survey of parallel programming languages and tools. *Computer Sciences Corporation, NASA Ames Research Center, Report RND-93-005 March* (1993).

[16] Rezaul A Chowdhury, Peter Djeu, Brendon Cahoon, James H Burrill, and Kathryn S McKinley. 2004. The limits of alias analysis for scalar optimizations. In *Compiler Construction*. Springer, 24–38.

[17] Luke Dalessandro, François Carouge, Sean White, Yossi Lev, Mark Moir, Michael L. Scott, and Michael F. Spear. 2011. Hybrid NOrec: A Case Study in the Effectiveness of Best Effort Hardware Transactional Memory. In *Proceedings of the Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS XVI)*. ACM, New York, NY, USA, 39–52. https://doi.org/10.1145/1950365.1950373

[18] Luke Dalessandro and Michael L. Scott. 2012. Sandboxing transactional memory. In *International Conference on Parallel Architectures and Compilation Techniques, PACT '12, Minneapolis, MN, USA - September 19 - 23, 2012*, Pen-Chung Yew, Sangyeun Cho, Luiz DeRose, and David J. Lilja (Eds.). ACM, 171–180. https://doi.org/10.1145/2370816.2370843

[19] Luke Dalessandro, Michael F Spear, and Michael L Scott. 2010. NOrec: streamlining STM by abolishing ownership records. In *ACM Sigplan Notices*, Vol. 45. ACM, 67–78.

[20] Francis Dang, Hao Yu, and Lawrence Rauchwerger. 2001. The R-LRPD test: Speculative parallelization of partially parallel loops. In *Parallel and Distributed Processing Symposium., Proceedings International, IPDPS 2002*. IEEE, 10–pp.

[21] Matthew DeVuyst, Dean M Tullsen, and Seon Wook Kim. 2011. Runtime parallelization of legacy code on a transactional memory system. In *Proceedings of the 6th International Conference on High Performance and Embedded Architectures and Compilers*. ACM, 127–136.

[22] Dave Dice, Ori Shalev, and Nir Shavit. 2006. Transactional Locking II. In *In Proc. of the 20th Intl. Symp. on Distributed Computing*.

[23] Nicholas DiPasquale, T Way, and V Gehlot. 2005. Comparative survey of approaches to automatic parallelization. *MASPLAS'05* (2005).

[24] Tobias JK Edler von Koch and Björn Franke. 2013. Limits of region-based dynamic binary parallelization. In *ACM SIGPLAN Notices*, Vol. 48. ACM, 13–22.

[25] Paul Feautrier. 1992. Some efficient solutions to the affine scheduling problem. I. One-dimensional time. *International journal of parallel programming* 21, 5 (1992), 313–347.

[26] Pascal Felber, Christof Fetzer, and Torvald Riegel. 2008. Dynamic performance tuning of word-based software transactional memory. In *Proceedings of the 13th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPOPP 2008, Salt Lake City, UT, USA, February 20-23, 2008*, Siddhartha Chatterjee and Michael L. Scott (Eds.). ACM, 237–246.

[27] Pascal Felber, Shady Issa, Alexander Matveev, and Paolo Romano. 2016. Hardware read-write lock elision. In *Proceedings of the Eleventh European Conference on Computer Systems, EuroSys 2016, London, United Kingdom, April 18-21, 2016*, Cristian Cadar, Peter R. Pietzuch, Kimberly Keeton, and Rodrigo Rodrigues (Eds.). ACM, 34:1–34:15. https://doi.org/10.1145/2901318.2901346

[28] MA 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), 24.

[29] Tobias Grosser, Hongbin Zheng, Raghesh Aloor, Andreas Simbürger, Armin Größlinger, and Louis-Noël Pouchet. 2011. Polly-Polyhedral optimization in LLVM. In *Proceedings of the First International Workshop on Polyhedral Compilation Techniques (IMPACT)*, Vol. 2011.

[30] Manish Gupta, Sayak Mukhopadhyay, and Navin Sinha. 2000. Automatic parallelization of recursive procedures. *International Journal of Parallel Programming* 28, 6 (2000), 537–562.

[31] Lance Hammond, Mark Willey, and Kunle Olukotun. 1998. Data Speculation Support for a Chip Multiprocessor. *SIGOPS Oper. Syst. Rev.* 32, 5 (Oct. 1998), 58–69.

[32] Tim Harris, James Larus, and Ravi Rajwar. 2010. Transactional Memory, 2nd edition. *Synthesis Lectures on Computer Architecture* 5, 1 (2010), 1–263. https://doi.org/10.2200/S00272ED1V01Y201006CAC011

[33] David Heath, Robert Jarrow, and Andrew Morton. 1992. Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica: Journal of the Econometric Society* (1992), 77–105.

[34] Shan Shan Huang, Amir Hormati, David F. Bacon, and Rodric M. Rabbah. 2008. Liquid Metal: Object-Oriented Programming Across the Hardware/Software Boundary. In *ECOOP 2008 - Object-Oriented Programming, 22nd European Conference, Paphos, Cyprus, July 7-11, 2008, Proceedings*. 76–103.

[35] R Intel. 2012. Architecture Instruction Set Extensions Programming Reference. *Intel Corporation, Feb* (2012).

[36] Shady Issa, Pascal Felber, Alexander Matveev, and Paolo Romano. 2017. Extending Hardware Transactional Memory Capacity via Rollback-Only Transactions and Suspend/Resume. In *31st International Symposium on Distributed Computing, DISC 2017, October 16-20, 2017, Vienna, Austria (LIPIcs)*, Andréa W. Richa (Ed.), Vol. 91. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 28:1–28:16. https://doi.org/10.4230/LIPIcs.DISC.2017.28

[37] Natanael Karjanto, Binur Yermukanova, and Laila Zhexembay. 2015. Black-Scholes equation. *arXiv preprint arXiv:1504.03074* (2015).

[38] Hironori Kasahara, Motoki Obata, and Kazuhisa Ishizaka. 2001. Automatic coarse grain task parallel processing on smp using openmp. In *Languages and Compilers for Parallel Computing*. Springer, 189–207.

[39] Sangman Kim, Michael Z. Lee, Alan M. Dunn, Owen S. Hofmann, Xuan Wang, Emmett Witchel, and Donald E. Porter. 2012. Improving server applications with system transactions. In *European Conference on Computer Systems, Proceedings of the Seventh EuroSys Conference 2012, EuroSys '12, Bern, Switzerland, April 10-13, 2012*, Pascal Felber, Frank Bellosa, and Herbert Bos (Eds.). ACM, 15–28. https://doi.org/10.1145/2168836.2168839

[40] Venkata Krishnan and Josep Torrellas. 1999. A chip-multiprocessor architecture with speculative multithreading. *Computers, IEEE Transactions on* 48, 9 (1999), 866–880.

[41] Leslie Lamport. 1974. The parallel execution of DO loops. *Commun. ACM* 17, 2 (1974), 83–93.

[42] Chris Lattner and Vikram Adve. 2004. LLVM: A compilation framework for lifelong program analysis & transformation. In *Code Generation and Optimization, 2004. CGO 2004. International Symposium on*. IEEE, 75–86.

[43] Amy W Lim and Monica S Lam. 1997. Maximizing parallelism and minimizing synchronization with affine transforms. In *Proceedings of the 24th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. ACM, 201–214.

[44] Wei Liu, James Tuck, Luis Ceze, Wonsun Ahn, Karin Strauss, Jose Renau, and Josep Torrellas. 2006. POSH: a TLS compiler that exploits program structure. In *Proceedings of the eleventh ACM SIGPLAN symposium on Principles and practice of parallel programming*. ACM, 158–167.

[45] Alexander Matveev and Nir Shavit. 2015. Reduced Hardware NOrec: A Safe and Scalable Hybrid Transactional Memory. In *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM, 59–71.

[46] Mojtaba Mehrara, Jeff Hao, Po-Chun Hsu, and Scott Mahlke. 2009. Parallelizing sequential applications on commodity hardware using a low-cost software transactional memory. In *Proceedings of the 2009 ACM SIGPLAN conference on Programming language design and implementation (PLDI '09)*. ACM, New York, NY, USA, 166–176. https://doi.org/10.1145/1542476.1542495

[47] Mohamed Mohamedin, Roberto Palmieri, Ahmed Hassan, and Binoy Ravindran. 2017. Managing Resource Limitation of Best-Effort HTM. *IEEE Trans. Parallel Distrib. Syst.* 28, 8 (2017), 2299–2313. https://doi.org/10.1109/TPDS.2017.2668415

[48] Matthias Müller, David Charypar, and Markus Gross. 2003. Particle-based Fluid Simulation for Interactive Applications. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '03)*. 154–159.
[49] Stefan C. Müller, Gustavo Alonso, Adam Amara, and André Csillaghy. 2014. Pydron: Semi-Automatic Parallelization for Multi-Core and the Cloud. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*. USENIX Association, Broomfield, CO, 645–659.
[50] AB MySQL. 1995. *MySQL: the world's most popular open source database*. MySQL AB.
[51] Nomair A Naeem and Ondrej Lhoták. 2009. Efficient alias set analysis using SSA form. In *Proceedings of the 2009 international symposium on Memory management*. ACM, 79–88.
[52] William Morton Pottenger. 1995. *Induction variable substitution and reduction recognition in the Polaris parallelizing compiler*. Ph.D. Dissertation. Citeseer.
[53] Arun Raman, Hanjun Kim, Thomas R Mason, Thomas B Jablin, and David I August. 2010. Speculative parallelization using software multi-threaded transactions. In *ACM SIGARCH Computer Architecture News*, Vol. 38. ACM, 65–76.
[54] Ravi Ramaseshan and Frank Mueller. 2008. Toward thread-level speculation for coarse-grained parallelism of regular access patterns. In *Workshop on Programmability Issues for Multi-Core Computers*. 12.
[55] Lawrence Rauchwerger and David A Padua. 1999. The LRPD test: Speculative run-time parallelization of loops with privatization and reduction parallelization. *Parallel and Distributed Systems, IEEE Transactions on* 10, 2 (1999), 160–180.
[56] James Reinders. 2013. Transactional Synchronization in Haswell. http://software.intel.com/en-us/blogs/2012/02/07/transactional-synchronization-in-haswell/.
[57] Christopher J. Rossbach, Yuan Yu, Jon Currey, Jean-Philippe Martin, and Dennis Fetterly. 2013. Dandelion: a compiler and runtime for heterogeneous systems. In *ACM SIGOPS 24th Symposium on Operating Systems Principles, SOSP '13, Farmington, PA, USA, November 3-6, 2013*, Michael Kaminsky and Mike Dahlin (Eds.). ACM, 49–68. https://doi.org/10.1145/2517349.2522715
[58] Wenjia Ruan, Yujie Liu, and Michael Spear. 2015. Transactional read-modify-write without aborts. *ACM Transactions on Architecture and Code Optimization (TACO)* 11, 4 (2015), 63.
[59] Radu Rugina and Martin Rinard. 1999. Automatic parallelization of divide and conquer algorithms. In *ACM SIGPLAN Notices*, Vol. 34. ACM, 72–83.
[60] Mohamed M. Saad. 2016. *Extracting Parallelism from Legacy Sequential Code Using Transactional Memory*. Ph.D. Dissertation. Virginia Tech. https://vtechworks.lib.vt.edu/handle/10919/71861.
[61] Mohamed M. Saad, Masoomeh Javidi Kishi, Shihao Jing, Sandeep Hans, and Roberto Palmieri. 2019. Processing Transactions in a Predefined Order. In *Proceedings of the 24th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming, PPOPP 2019, Washington DC, USA, February 16-20, 2019*. ACM.
[62] Mohamed M. Saad, Mohamed Mohamedin, and Binoy Ravindran. 2012. HydraVM: Extracting Parallelism from Legacy Sequential Code Using STM. In *4th USENIX Workshop on Hot Topics in Parallelism, HotPar'12, Berkeley, CA, USA, June 7-8, 2012*.
[63] Mohamed M. Saad, Roberto Palmieri, Ahmed Hassan, and Binoy Ravindran. 2016. Extending TM Primitives using Low Level Semantics. In *Proceedings of the 28th ACM Symposium on Parallelism in Algorithms and Architectures, SPAA 2016, Asilomar State Beach/Pacific Grove, CA, USA, July 11-13, 2016*, Christian Scheideler and Seth Gilbert (Eds.). ACM, 109–120. https://doi.org/10.1145/2935764.2935794
[64] Bratin Saha, Ali-Reza Adl-Tabatabai, and Quinn Jacobson. 2006. Architectural Support for Software Transactional Memory. In *MICRO 39: Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE Computer Society, Washington, DC, USA, 185–196. https://doi.org/10.1109/MICRO.2006.9
[65] Joel H Saltz, Ravi Mirchandaney, and K Crowley. 1991. The Preprocessed Doacross Loop.. In *ICPP (2)*. 174–179.
[66] J Greggory Steffan, Christopher B Colohan, Antonia Zhai, and Todd C Mowry. 2000. *A scalable approach to thread-level speculation*. Vol. 28. ACM.
[67] Kevin Streit, Clemens Hammacher, Andreas Zeller, and Sebastian Hack. 2013. Sambamba: runtime adaptive parallel execution. In *Proceedings of the 3rd International Workshop on Adaptive Self-Tuning Computing Systems*. ACM, 7.
[68] Hans Vandierendonck, Sean Rul, and Koen De Bosschere. 2010. The Paralax infrastructure: automatic parallelization with a helping hand. In *Proceedings of the 19th international conference on Parallel architectures and compilation techniques*. ACM, 389–400.
[69] Christoph von Praun, Rajesh Bordawekar, and Calin Cascaval. 2008. Modeling optimistic concurrency using quantitative dependence analysis. In *Proceedings of the 13th ACM SIGPLAN Symposium on Principles and practice of parallel programming*. ACM, 185–196.
[70] Amos Waterland, Elaine Angelino, Ryan P. Adams, Jonathan Appavoo, and Margo I. Seltzer. 2014. ASC: automatically scalable computation. In *ASPLOS*, Rajeev Balasubramonian, Al Davis, and Sarita V. Adve (Eds.). ACM, 575–590.

Received February 2007; revised March 2009; accepted June 2009