Synchronization Transformations for Parallel Computing

Corresponding Author:

PEDRO C. DINIZ

Information Sciences Institute

University of Southern California

4676 Admiralty Way, Suite 1001

Marina del Rey, CA, 90292-6695

Phone: (310) 822-1511 x457

Fax: (310) 823-6714

E-mail: pedro@isi.edu

Synchronization Transformations for Parallel Computing

PEDRO C. DINIZ

Information Sciences Institute

University of Southern California

and

MARTIN C. RINARD

Laboratory for Computer Science

Massachusetts Institute of Technology

This research was supported in part by an Alfred P. Sloan Research Fellowship.

As parallel machines become part of the mainstream computing environment, compilers will need to apply synchronization optimizations to deliver efficient parallel software. This paper describes a new framework for synchronization optimizations and a new set of transformations for programs that implement critical sections using mutual exclusion locks. These transformations allow the compiler to move constructs that acquire and release locks both within and between procedures and to eliminate acquire and release constructs.

The paper also presents a new synchronization algorithm, lock elimination, for reducing synchronization overhead. This optimization locates computations that repeatedly acquire and release the same lock, then uses the transformations to obtain equivalent computations that acquire and release the lock only once. Experimental results from a parallelizing compiler for object-based programs illustrate the practical utility of this optimization. For three benchmark programs, the optimization dramatically reduces the number of times the computations acquire and release locks, which significantly reduces the amount of time processors spend acquiring and releasing locks. For one of the three benchmarks, the optimization always significantly improves the overall performance. Depending on the number of processors executing the computation, the optimized version runs between 1.03 and 1.30 times faster than the unoptimized version. For one of the other benchmarks, the optimized version runs between 1.07 and 1.02 times faster than the unoptimized version, with a mean of 1.06 times faster. For the final benchmark, the optimization does not improve the overall performance.

Categories and Subject Descriptors: D.3.4 [**Compilers**]: Parallelizing Compilers; D.4.1 [**Synchronization and Mutual Exclusion**]: Synchronization Transformations

General Terms: Parallelizing Compilers, Compilers, Parallel Computing, Synchronization Transformations

1. INTRODUCTION

The characteristics of future computational environments ensure that parallel computing will play an increasingly important role in many areas of computer science. As small-scale shared-memory multiprocessors become a commodity source of computation, customers will demand the efficient parallel software required to fully exploit the parallel hardware. The growth of the World-Wide Web will provide a new distributed computing environment with unprecedented computational power and functionality. Parallel computing will continue to play a crucial role in delivering maximum performance for scientific and engineering computations. The increasing use of multiple threads as an effective program construction technique (used, for example, in user interface systems and multi-threaded servers [Hauser et al. 1993; Cardelli and Pike 1985; Reppy 1992]) demonstrates that parallelism is not just for performance — it can also increase the expressive power of a language.

Efficient synchronization is one of the fundamental requirements of effective parallel computing. The tasks in fine-grain parallel computations, for example, need fast synchronization for efficient control of their frequent interactions. Efficient synchronization also promotes the development of reliable parallel software because it allows programmers to structure programs as a set of synchronized operations on fine-grain objects. This development methodology helps programmers overcome the challenging problems (nondeterministic behavior, deadlock, etc.) that complicate the development of parallel software.

Given the central role that efficient synchronization plays in parallel computing, we expect that future compilers will apply a wide range of synchronization optimizations. This paper takes a first step towards that goal by presenting a new transformation framework and set of specific transformations for programs that contain synchronization operations. It also describes a novel synchronization optimization algorithm called *lock elimination*. This optimization is designed for programs that use mutual exclusion locks to implement critical sections. Lock elimination drives down the locking overhead by coalescing multiple critical sections that acquire and release the same lock multiple times into a single critical section that acquires and releases the lock only once. This algorithm provides a concrete example of how the transformations enable meaningful optimizations.

Finally, this paper presents experimental results that demonstrate the practical utility of lock elimination. These experimental results come from a compiler that automatically parallelizes object-based programs written in a subset of serial C++. Because this compiler uses a new analysis technique called commutativity analysis [Rinard and Diniz 1996], it automatically inserts synchronization constructs into the generated parallel code to make operations execute atomically. The significant performance improvements that synchronization optimizations deliver in this context illustrates their importance in achieving good parallel performance.

This paper makes the following contributions:

—It presents a new set of basic synchronization transformations. These synchronization transformations provide the solid foundation that a compiler requires to effectively apply synchronization optimizations.

—It presents a novel optimization algorithm, lock elimination, that a parallelizing compiler can use to reduce the synchronization overhead.

—It presents experimental results that characterize the performance impact of applying the lock elimination optimization in a parallelizing compiler for object-based programs. These results show that the optimization has a substantial impact on the performance of three benchmark programs.

2. THE MODEL OF COMPUTATION

The framework is designed for programs that execute a sequence of parallel and serial phases. Each parallel phase consists of a set of parallel threads that access shared data objects.

Parallel and serial phases are delimited by concurrency generation constructs, such as **parbegin/parend** or **doall** or even **spawn**. These high-level constructs can either be conveyed by an explicity parallel language or inserted explicitly by a parallelizing compiler. The actual mechanism is irrelevant for the purposes of the transformation and analysis algorithms described in this article. The basic assumption is that the analysis algorithm must have a clear definition of what the code of each thread is. To that respect, parallel constructs such as **doall** define syntactically the boundaries of the threads. Threads that have nested and possibly irregular forms of concurrency present a challenge to accurately define their code boundaries.

The framework presented here is orthogonal to the problem of defining thread boundaries. A compiler can identify the boundaries of threads by tracing the possible code executed at the concurrency generation site. While in some more ”regular” parallel constructs, such as **doall** where the creation and termination of threads is clearly defined, other constructs such as **spawn** pose a serious challenge for a compiler to automatically recognize the boundaries of the thread code.

Multiple parallel threads in the program use critial sections to atomically access one or more pieces of shared data. Programs implement critical sections by acquiring a mutual exclusion lock at the beginning of the critical section (using the **acquire** construct), then releasing the lock at the end of the section (using the **release** construct). Threads are therefore not allowed to synchronize through direct shared memory accesses.

In practice we expect programmers to mentally associate mutual exclusion locks with data; each critical section would then acquire and release the lock associated with the manipulated data. The transformations and framework are independent of any such association, however.

Each of the mutual exlcusion locks the program manipulates is denoted by a symbolic lock expression. Two lock expressions denoted the same mutual exclusion lock iff they evaluate to the same lock. In many, but not all, of the cases the analysis algorithms will be able to distinguish between two locks being manipulated in concurrent thread by analysis of the corresponding lock expression. This ability to distinguish any two locks depends on the scope and naming mechanism of the underlying concurrent language. In Java, for example, locking an object occurs at the class level and the locking of the class is implicit in the execution of synchronized methods. In other languages there is a single, and globally accessible, array of mutual exclusion locks that all parallel threads can use.

The synchronization transformation algorithms decribed in this article makes no assumptions over the number of outstanding locks each thread might have and therefore make no assumptions or guarantees about deadlock of the application. The key property that the synchronization algorithms ensure is that if more that one lock is held at any given time then the order in which the locks are acquired and released in the transformed code remains unchanged. For section of the code that do not acquire and release locks the described algorithm might, however, introduce a single mutual exclusion region. Both these key properties ensure that the resulting code has the same properties with respect to deadlock as the original program.

The described algorithm makes no fairness assumptions or guarantees for the transformed code. While fairness might be a critical issues for some application domains for the set of tested applications, progress of at least one thread over fairness, is a sufficient correctness requirement.

3. PROGRAM REPRESENTATION

In this analysis abstraction we represent threads of a given parallel phase, by their interprocedural control flow graph (ICFG) [Reps et al. 1995]. This ICFG consists of the union of the control flow graphs of the individual procedures that make the code segment the thread can execute. Each parallel program may contain several thread creation sites. The number of threads created is, in general, not known statically. Each thread creation site has an ICFG that represents the sequential computations of the threads that are created at that site. Each ICFG therefore represents a sequential computation that may execute in parallel with other sequential computations. The program is represented as the union of the ICFGs for all of the thread creation sites.

In this representation a procedure call is represented by two nodes: a call node and a return node. There is an edge from each call node to the entry node of the invoked procedure and an edge back from the exit of the invoked procedure to the corresponding return node. Each node in the graph has four attributes:

—**Type:** The type of computation that the node performs. Standard types include:

—acquire (acquire a mutual exclusion lock),

—release (release a mutual exclusion lock),

—assignment (set a variable to a new value),

—call (invoke a procedure),

—return (return from a procedure),

—entry (dummy node at the beginning of a procedure),

—exit (dummy node at the end of a procedure),

—if (flow of control), and

—merge (flow of control).

There is also a summary type (described below in Section 4.1) that represents the computation of several nodes. All release, acquire, call, return, assignment and summary nodes have a single incoming and a single outgoing edge. All entry and merge nodes have a single outgoing edge; all exit and if nodes have a single incoming edge.

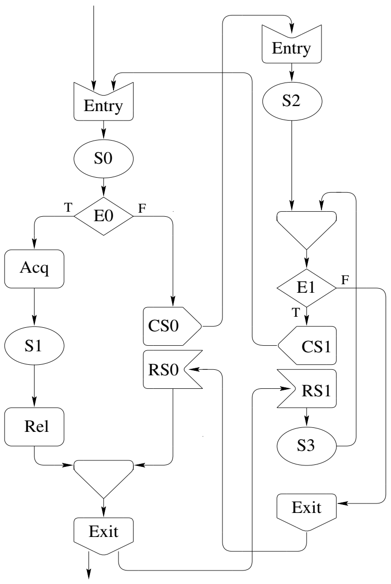
—**Expressions:** One or more expressions representing the computation associated with the node. For example, the expression for an acquire or release node specifies the lock to acquire or release, and a call node has one expression for each parameter.

—**Read Set:** A conservative approximation of the set of variables that the node's computation reads. In general, the compiler may have to use an interprocedural pointer or alias analysis to compute a reasonably precise read set [Emami et al. 1994; Wilson and

Lam 1995; Landi et al. 1993]. In restricted contexts the compiler may be able to use simpler algorithms. Our prototype compiler, for example, is designed for object-based programs. Because these programs use references to objects instead of pointers, it is possible to extract a reasonable read set directly from the expressions in the node. [Rinard and Diniz 1996]

—**Write Set:** A conservative approximation of the set of variables that the node's computation writes.

Figure 1 contains an example ICFG. The different shapes correspond to nodes of different types. To simplify the figure, we omit the expressions, read sets and write sets of the nodes.



Shape	Attributes			
	Type	Expressions	Read Set	Write Set
Acq Rel	Acquire Release	Lock	Yes	No
Entry Exit	Entry Exit	None	No	No
S0	Assignment	Left Hand Side Right Hand Side	Yes	Yes
	Summary	None	Yes	Yes
T E0 F	Conditional	Condition	Yes	No
	Merge	None	No	No
CS0	Call	Arguments	Yes	No
RS0	Return	None	No	No

Fig. 1. ICFG Example

4. TRANSFORMATIONS

We next present the basic program transformations. The lock elimination algorithm, described in Section 5, uses these basic transformations in a coherent way to transform the original program, as to exhibits a lower synchronization overhead.

4.1 Abstraction Transformations

Because the synchronization transformations deal primarily with the movement and manipulation of synchronization nodes, it is appropriate for the compiler to use an abstract, simplified representation of the actual computation in the ICFG. The compiler can therefore apply several transformations that replace concrete representations of computation with more abstract representations. The end result is a simpler ICFG, which improves the performance and functionality of the synchronization optimization algorithms. The transformations are as follows:

—**Node Abstraction:** A connected set of assignment, if or summary nodes with a single incoming edge and a single outgoing edge is replaced by a single summary node. Figure 2 presents this transformation.

—**Procedure Abstraction:** The invocation of a procedure that consists only of assignment, if or summary nodes is replaced with a single node summarizing the execution of the procedure. Figure 3 presents this transformation.[1]

In both cases the read set and write set of the new summary node are, respectively, the union of the read sets and the union of the write sets of the set of summarized nodes. The compiler can apply these transformations both at the beginning of the optimization phase before any other transformations, and during intermediate steps of the optimization phase as they become enabled.

These transformation can have a significant impact on the size of the ICFG program representation and consequently on the run-time and memory complexity of the synchronization transformation algorithms. For the set of three applications codes described in the experimental results section (Section 6) node and procedure abstraction is able to reduce the size of the ICFG from as low as 25% to a maximum of 92% with an average of 57% over all of the parallel phases of these applications.

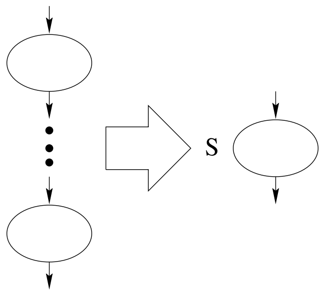


Fig. 2. Node Abstraction Transformation

[1]It is straightforward to extend this transformation to abstract sets of mutually recursive procedures.

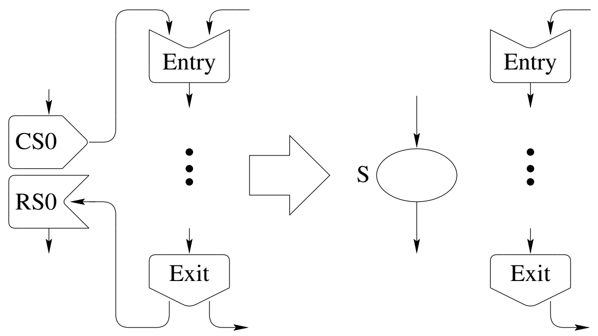


Fig. 3. Procedure Abstraction Transformation

4.2 Lock Cancellation

If a computation releases a lock, then immediately reacquires the same lock, it is possible to reduce the lock overhead by eliminating the adjacent release and acquire. A similar situation occurs when the computation acquires, then immediately releases a lock. The conditional lock cancellation transformations in Figures 4 and 5 start with two adjacent release and acquire nodes and introduce a new if node that tests if the nodes acquire and release the same lock. If so, the transformed ICFG simply skips the acquire and release.

The compiler may be able to detect statically that the acquire and release manipulate the same lock. This is clearly the case, for example, if the expressions in the acquire and release are the same. In this case the compiler can simply eliminate the two nodes as illustrated in Figure 6. In effect, these transformations combine one of the conditional lock cancellation transformations with dead code elimination.

It is not usually the case that the original ICFG contains adjacent acquire and release

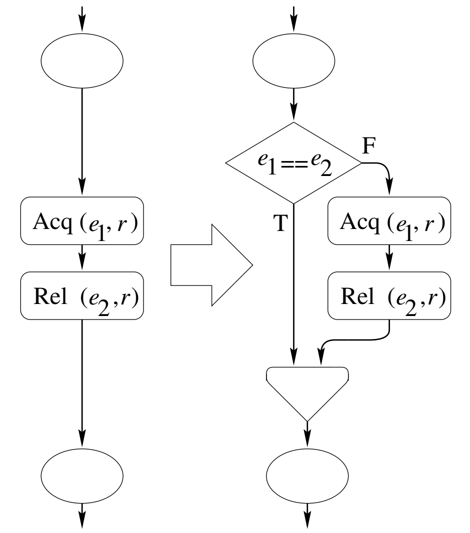


Fig. 4. Conditional Lock Cancellation Transformation

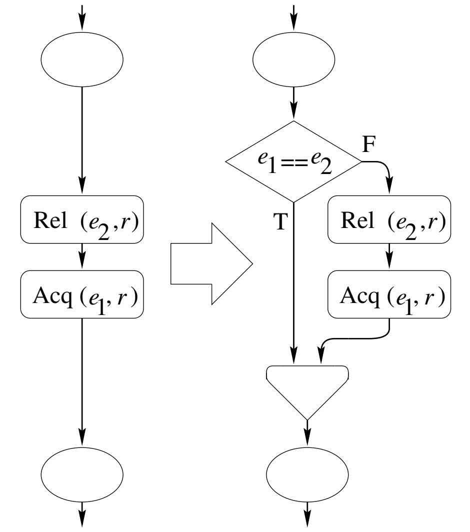


Fig. 5. Conditional Lock Cancellation Transformation

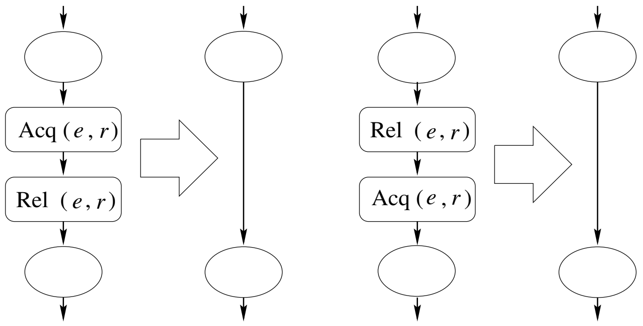


Fig. 6. Lock Cancellation Transformations

nodes that manipulate the same lock. Our experimental results indicate, however, that when combined with the lock movement transformations described below in Section 4.3, the lock cancellation transformations can significantly reduce the number of executed acquires and releases.

4.3 Lock Movement

The lock movement transformations move an acquire or release node across an adjacent node. There are two dual transformations — one for acquire nodes and one for release nodes. Figure 7 presents the transformation that moves an acquire node A against the flow of control across an adjacent node N. The transformation introduces new acquire nodes before N, removes the original acquire node A, and introduces new release nodes on all of the edges out of N except the one that led to A. In effect, the transformed code moves N into the critical section that started with A in the original code. The release nodes ensure that the original successor of A is the only successor of N that is in the newly enlarged critical section.

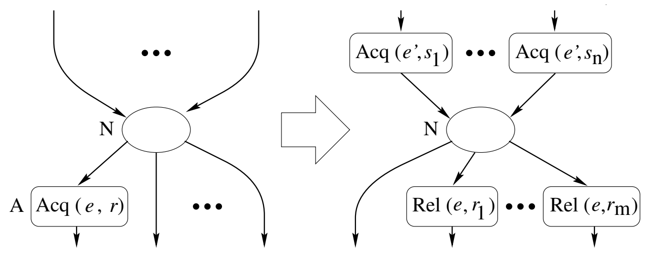


Fig. 7. Acquire Lock Movement Transformation

To ensure that the new acquire nodes do not change the order in which the program acquires and releases locks, possibly introducing deadlock, we require that, in the original code, N can never be part of a critical section. We formalize this requirement by introducing the concept of an acquire-exposed node.

DEFINITION 1. *A node is* acquire-exposed *if there may exist a path to that node from an acquire node and the path does not go through a release node that releases the lock that the acquire node acquires.*

It is illegal to move an acquire node A across an acquire-exposed node N.

The original acquire node A has an expression e and read set r. The new acquire nodes have expression e' and read sets $s_1, \ldots, s_n$; the new release nodes have expression e and read sets $r_1, \ldots, r_m$. The expression e' may differ from e because it is evaluated before N executes rather than after N executes. The expression manipulations required to transform

e to e' may involve performing variable substitutions to undo the parameter bindings when moving a node out of a procedure and replacing variables with corresponding expressions when moving across an assignment node. If there are multiple edges out of N, the compiler must also ensure that the expressions in the new nodes always denote a valid lock.

The read sets in the new acquire nodes may differ from the original read set because they summarize how a new expression, e', reads variables. Even if the new expressions are identical, the new read sets may differ from the original read set and from each other because the expressions may be evaluated in different contexts.

In some cases the compiler may be unable to apply the transformation because it cannot generate the new expression or read sets. This may happen, for example, if N may write variables that A reads. There are also potential complications if the transformation would move an acquire out of a procedure into a new naming context. Appendix A specifies the cases that the compiler must handle.

Figure 8 presents the lock release transformation that moves a release node R with the flow of control across an adjacent node N. This transformation is the dual of the lock acquire transformation. In effect, the lock release transformation moves N into the critical section that originally ended with R. As for the acquire lock movement transformation, we require that N not be acquire-exposed. If there are multiple edges into N, the compiler must verify that all of the expressions in the new nodes always denote a valid lock. As for the acquire lock movement transformation, the compiler may be unable to apply the release transformation because it can not generate the new expression or read sets. Appendix A specifies the cases that the compiler must handle.

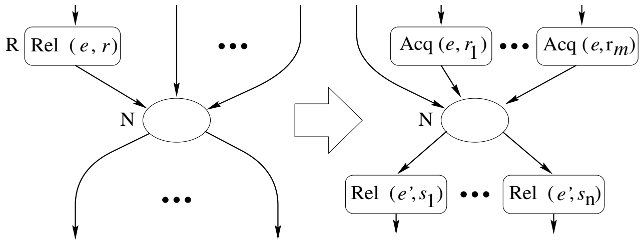


Fig. 8. Release Lock Movement Transformation

In principle, both transformations are reversible. When used in the other direction (moving acquires with the flow of control and releases against the flow of control), they have the effect of reducing the size of the critical section. It is therefore possible to use the transformations to minimize the sizes of the critical sections, which may increase the amount of available parallelism in the program. There is one additional requirement: to ensure that the computation in the original critical sections still executes atomically, the transformations must not move a node out of a critical section if the node reads a variable that other parallel threads may write or writes a variable that other parallel threads may access.

5. LOCK ELIMINATION

The goal of the lock elimination algorithm is to reduce the number of times the computation releases and acquires locks. The basic idea is to identify a computation that contains multiple critical sections that acquire and release the same lock, then transform the computation so that it contains one large critical section that acquires and releases the lock only once. Because the transformed computation acquires and releases the lock fewer times, it generates less lock overhead.

Given a region over which to eliminate synchronization constructs, the algorithm uses the lock movement transformation to increase the sizes of critical regions that acquire and release the same lock until they are adjacent in the ICFG. It then uses the lock cancellation transformation to eliminate adjacent release and acquire nodes. The end result is a single larger critical section that encompasses all of the computation between the start of the first critical section and the end of the last critical section.

5.1 False Exclusion

An overly aggressive lock elimination algorithm may introduce *false exclusion*. False exclusion may occur when a processor holds a lock during an extended period of computation that was originally part of no critical section. If another processor attempts to execute a critical section that uses the same lock, it must wait for the first processor to release the lock even though the first processor is not executing a computation that needs to be in a critical section. False exclusion may therefore reduce the performance by decreasing the amount of available concurrency.

The amount of false exclusion in a given parallel execution depends on information such as the dynamic interleaving of the parallel tasks and the relative execution times of pieces of the ICFG. This information is, in general, unavailable at compile time. For the same computation it may even be different in different executions. The lock elimination algorithm therefore addresses the issue of false exclusion using a heuristic false exclusion policy. This policy limits the subgraphs of the ICFG to which the algorithm applies the lock elimination algorithm. The policy considers each procedure in turn to determine if it should apply the algorithm to the subgraph of the ICFG containing that procedure and all procedures that it (transitively) invokes. It only applies the algorithm to a subgraph if the subgraph satisfies the false exclusion policy. The current compiler supports four different policies:

—**Original:** Never apply the transformation — always use the default placement of acquire and release constructs. In the default placement for our current compiler, each operation that updates an object acquires and releases that object's lock.

—**Bounded:** Apply the transformation only if the new critical region will contain no cycles in the call graph. The idea is to limit the severity of any false exclusion by limiting the dynamic size of the critical region.

—**Aggressive:** Always apply the transformation unless the transformation would serialize the entire computation of the parallel phase. In the current implementation the compiler checks if the application of the transformations completely serializes the computation of a parallel phase. If the compiler is unable to verify this condition, it applies the transformation.

—**Greedy:** Always apply the transformation whenever possible.

5.2 The Lock Elimination Algorithm

The basic idea behind the lock elimination algorithm is to find an acquire node and a release node, find a path in the ICFG along which they can move until they are adjacent, then use lock cancellation transformations to eliminate them. A by-product of the sequence of transformations is a set of new acquire and release nodes introduced on edges that lead into and out of the path. The algorithm performs the following steps:

—**Apply False Exclusion Algorithm:** The algorithm performs a depth-first traversal of the call graph. At each node of the call graph, the algorithm considers the subgraph reachable from that node in the call graph. If this subgraph satisfies the false exclusion policy, the algorithm invokes the lock elimination algorithm on the procedure corresponding to that node.

—**Reachability Tree:** The lock elimination algorithm chooses a release node and an acquire node, then computes the *reachability tree* for each node. The reachability tree contains the set of edges to which the algorithm can move the acquire or release node using the lock movement transformations. It also contains the new expressions and read sets computed in the lock movement transformations. Figure 9 contains an example reachability tree for an acquire node. In this figure the edges in the reachability tree are shaded. Figure 10 contains an example reachability tree for a release node. These figures omit the expressions and read sets in the reachability trees.

—**Reachability Tree Intersection:** Given two reachability trees, the algorithm next checks if they intersect and have the same expression at an edge where they intersect. If so, it is possible to move the acquire and release nodes to be adjacent in the ICFG. Note that it may be possible to move a release node and an acquire node to be adjacent even though neither node's reachability tree reaches the other node. This may occur, for example, if it is necessary to move the nodes out of invoked procedures into a common caller procedure. Figure 11 contains an example of this situation. It identifies the intersection edge using a thick line. In some cases the intersection of the reacheability tree might not be a single edge, but be a whole segment of the ICFG, or even disjoint segments of the ICFG. A possible concearn is the choice of which edge to select as the intersection edge.

The key property of the lock movement transformation is that any compensating nodes placed along edges in the set of intersection edges, will have a reachability tree that is a strict subset of the reachability tree of the moved synchronization node. This means that, which and in which order, the algorithm picks an intersection edge is irrelevant. This key property suggest an obvious algorithmic improvement. For a given pair of acquire and release nodes, the algorithm can trace the set of acquire and release nodes and avoid recomputing from skratch the reachability trees for these newly introduced synchronization constructs as the algorithm can use the uncovered portions of the reachability tree computed for the synchronization nodes being moved.

This argument of picking any edges in the rechability tree intersection trivially extends to disjoint reacheability trees.

—**Movement Paths:** If the trees intersect, the algorithm follows the edges in the reachability trees to obtain paths from the release and acquire nodes to the intersection edge. The acquire and release can move along these paths to become adjacent. Figure 12 shows the movement paths in our example.

—**Transformation:** To apply the transformation, the algorithm eliminates the original acquire and release nodes, then introduces new acquire and release nodes into edges that lead into and out of the two movement paths. In effect, the algorithm applies all of the lock movement and cancellation transformations in one step to move all of the nodes in the path into the enlarged critical section. Figure 13 presents the transformed ICFG in our example. It identifies the new acquire and release nodes using thick boxes.

—**Repetition:** The algorithm repeatedly eliminates acquire and release nodes until there are no such nodes whose reachability trees intersect. Figure 14 shows the final transformed ICFG.

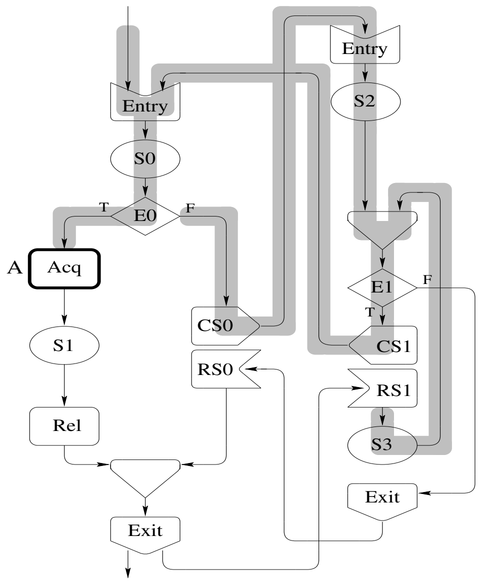


Fig. 9. Reachability Tree for Acquire Node A

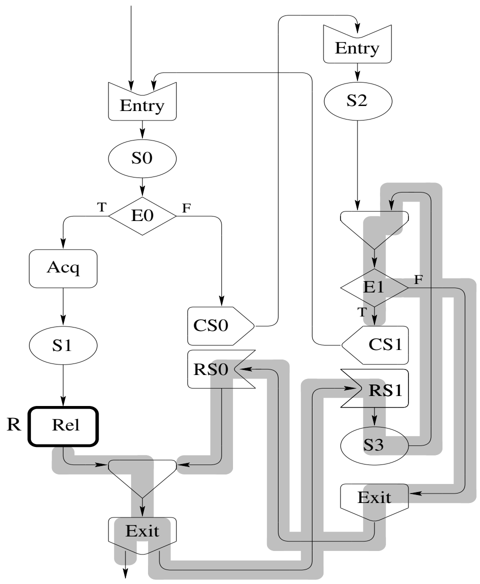


Fig. 10. Reachability Tree for Release Node R

Appendix B contains a precise specification of the lock elimination algorithm. We next discuss a few of its properties, including preserving the deadlock properties of the original program and the algorithm's algorithmic complexity.

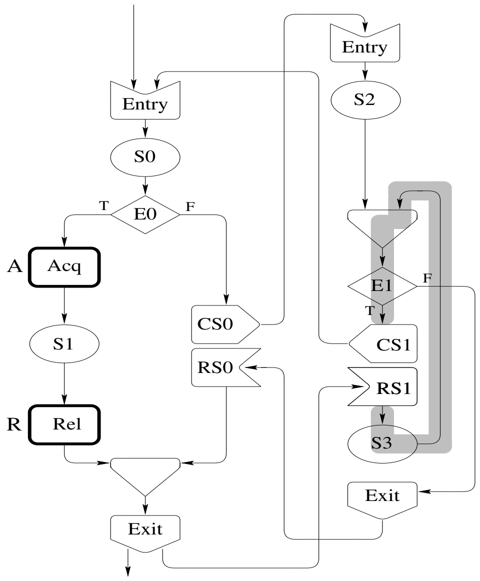


Fig. 11. Intersection of Reachability Trees for Acquire Node A and Release Node R

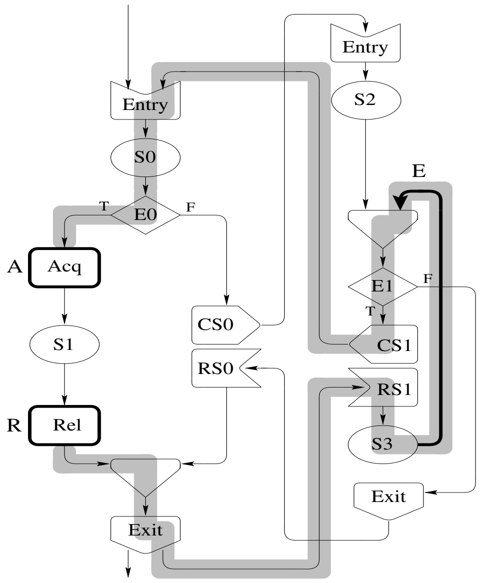


Fig. 12. Movement Paths and Intersection Edge E for Acquire Node A and Release Node R

5.3 Correctness and Synchronization Deadlock

The lock elimination algorithm has the same effect as performing multiple lock movement and lock cancellation transformations. Neither of these two transformations introduces deadlock. The key property is that these transformations do not move acquire or release nodes nodes past acquire-exposed nodes. What this means is that the algorithm never changes the order in which the original code acquires multiple locks, although it may create a critical section at a place in the execution of the code where previously there was none. Because the order in which the code acquire and releases multiple locks remain unchanged, and the algorithm creates at most a single critical regions in a section of the code that have

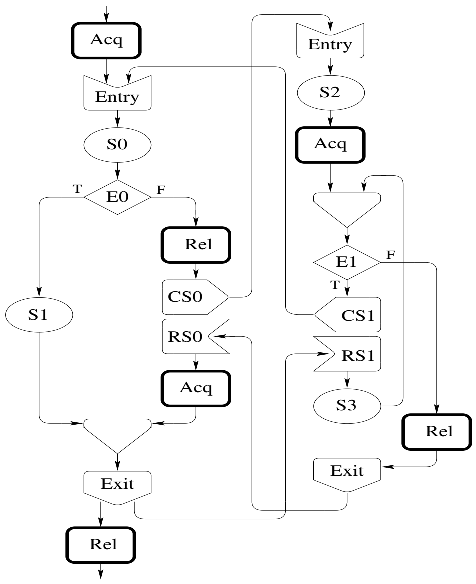


Fig. 13. Result of a Single Lock Elimination Step

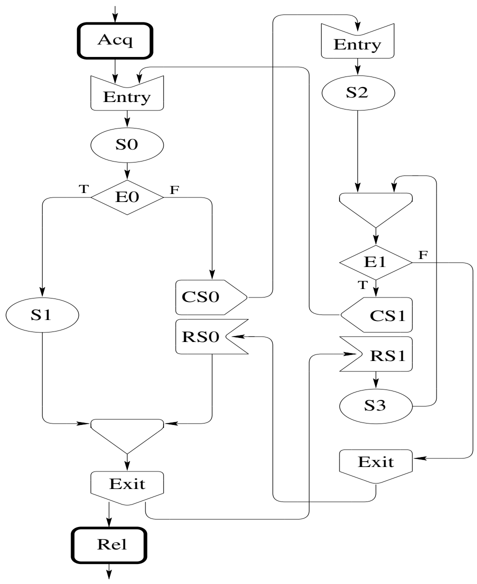


Fig. 14. Final Result of Lock Elimination Algorithm

none. The lock elimination algorithm, therefore, does not modify the deadlock behaviour of the original program. If the original program is deadlock free so is the transformed program.

5.4 Termination

We first address the termination property of the lock elimination algorithm. Because the lock movement transformations introduce new acquire and release nodes, it may not be completely obvious that the algorithm always terminates. Intuitively, the termination argument is that the transformations always increase the sizes of the critical sections. Clearly, the critical sections can get no bigger than the entire ICFG. Therefore, the algorithm must terminate.

The presence of cycles, or irreducible control flow, in the ICFG does not present any termination problems for the algorithm. The key observation is that when the algorithm moves a synchronization node A over another node N it inserts compensating nodes in all of the edges other than the edges that connects the node N to the synchronization node before that lock movement. For the algorithm not to terminate, it would mean that there exists synchronization free path in the rechability tree of the node A from the synchronization node A to itself. This is clearly impossible. The node immediately adjacent to the synchronization node (either along the flow of control for the case of an acquire, or against the flow of control for the case of a release) would then be an acquire-exposed node. Because rechability trees do not include acquire-exposed nodes, no such synchronization-free cycle can exist.

5.5 Time Complexity

We now addres the algorithmic time complexity of the lock elimination algorithm. To simplify the discussion we and avoid pathological algorithmic cases we normalize the ICFG by verifying and possibly applying two minor ICFG transformations. These two transformations reduce the number of acquire and release pairs the algorithm needs to check. These simplification account for some trivial cases where acquire and release nodes are already adjacent in the ICFG.

The first simplification is to eliminate any adjacent acquire and release nodes with the exact same lock expression. If the acquire node proceeds the release node it defines a clearly empty critical section. If the release node preceeds the acquire node, because they use the same lock, we can merge the two critical sections. The compiler can simply scan all of the edges of the ICFG and test for any two adjacent acquire and release nodes with the same expression before testing all pairs of acquire and release nodes for a possible lock elimination. The second simplification is to test an acquire or release node for a possible lock elimination only if it is connected to a non-synchronization node.

In this normalized form of the ICFG the upper bound on the running time of the lock elimination algorithm is $O(ne^2 \max(e, nC))$, where e is the number of edges in the ICFG, n is the number of nodes and C is the complexity of computing the new expression and read sets for a single lock movement.

The first observation is that each lock elimination transformation inserts at least one non-synchronization node into a critical section and takes no nodes out of critical sections, i.e., the critical sections are always expanded. The algorithm therefore performs at most $n + 1$ iterations. We next consider the amount of work done per iteration. The algorithm considers at most 2 synchronization nodes per edge of the original ICFG in the pairs of tested acquire and release nodes. The total number of tested pairs is therefore at most $O(e^2)$. For each pair the algorithm performs at most $O(\max(e, nC))$ work. To construct the reachability trees, the algorithm must visit at most $O(e)$ edges, computing the new expression and read sets at most $O(n)$ times. It is possible to fold the reachability tree intersection into the reachability tree construction. The movement path and transformations can also be computed and performed in $O(e)$ time. Each iteration therefore takes $O(\max(e, nC))$ time, and the total running time is $O(ne^2 \max(e, nC))$. We expect that, in practice, the number of synchronization nodes will be small relative to the number of edges and nodes in the ICFG, and the running time will be substantially faster than this upper sound might suggest.

Because of the symbolic variable substitution, it is possible for the computations of the new expressions and read sets to generate expressions that are exponentially larger than the original expressions in the graph. We expect that, in practice, few programs will contain lock expressions that elicit this behavior.

6. EXPERIMENTAL RESULTS

We now describe the context and methodology used to quantitatively evaluate the performance of the lock elimination algorithm described in this article.

6.1 Parallelizing Compiler

We have implemented a lock elimination algorithm in the context of a parallelizing compiler for serial object-based programs. The compiler uses commutativity analysis [Rinard and Diniz 1996] to extract the concurrency in the program. It views the computation as consisting of a sequence of operations on objects, then analyzes the program to determine if operations commute (two operations commute if they generate the same result regardless of the order in which they execute). If all of the operations in a given computation commute, the compiler can automatically generate code that executes the operations in parallel. The generated code uses mutual exclusion locks to ensure that operations execute atomically. Even though the parallel execution may change the order in which the operations are performed relative to the serial computation (which may violate the data dependences), the fact that all operations commute guarantees that all parallel executions generate the same final result as the serial execution.

The compiler exploits the structure present in the object-based programming paradigm to use a significantly simplified lock elimination algorithm. In this paradigm, each operation accesses only the local variables, the parameters and the object that it operates on. The compiler also exploits this control to simplify the data structures used in the implementation of the lock elimination algorithm — the implemented algorithm operates on the call graph and control flow graph for each procedure rather than on an explicit enlarged ICFG.

The compiler also controls the association of locks with data and the placement of the acquire and release constructs in operations. If a given operation (i.e., the execution of a methods) is executed concurrently the compiler augments the class declaration for the class that method is declared in, using a compiler defined symbolic name **mutex**. Objects of that class automatically have a lock instance variable the code can access by the symbolic name **this.mutex** or simply **mutex**. Each of the parallel methods invokes the operation **mutex.acquire()** to get exclusive access to the receiver object. To release the receiver object lock, a method invokes the **mutex.release()** operation. To ensure that the generated parallel code is deadlock free, the compiler generates code whose threads holds at most a single lock at a time. To achieve this the compiler augments the parallel methods by inserting synchronization constructs around the sections that exclusively update the receiver object and do not invoke any other operations. A concurrent method must therefore release the receiver object lock and will never hold a lock while invoking another operation.

6.2 Methodology

We report performance results for three automatically parallelized scientific applications: the Barnes-Hut hierarchical N-body solver [Barnes and Hut 1986], the Water code [Singh et al. 1992] and the String code [Harris et al. 1990]. Barnes-Hut simulates the trajectories of a set of interacting bodies under Newtonian forces; it consists of approximately 1500 lines of C++ code. Water simulates the interaction of water molecules in the liquid state; it consists of approximately 1850 lines of C++ code. String constructs a two-dimensional discrete velocity model of the geological medium between two oil wells; it consists of approximately 2050 lines of C++ code. The performance of the serial C++ versions of

Barnes-Hut and Water is slightly better than the performance of highly optimized parallel C versions from the SPLASH-2 benchmark set [Singh et al. 1992] running on one processor. The performance of the serial C++ version of String is approximately 1% slower than the original C version.

The compiler currently supports all four false exclusion policies described in Section 5.1. We generated four instrumented versions of each application; each version uses a different false exclusion policy. We evaluated the performance of each version by running it on a 16-processor Stanford DASH machine [Lenoski 1992].

6.3 Barnes-Hut

We evaluate the overhead of each false exclusion policy by running the three automatically parallelized versions on one processor and comparing the execution times with the execution time of the sequential program. The performance results in Table I show that the lock elimination algorithm has a significant impact on the overall performance. Without lock elimination, the original parallel version runs approximately two times slower than the serial version. Lock elimination with the Bounded policy reduces the lock overhead, and the Aggressive and Greedy policies virtually eliminate it. For this application, using the Greedy false exclusion policy the compiler generates the same parallel code as using the Aggressive policy. The table presents the number of executed acquire and release pairs for each of the different versions; these numbers correlate closely with the execution times.

The table also presents the number of static acquire and release constructs for each version; this is the number of these constructs in the code.

Version	Execution Time	Execution Time Overhead	Acquire/Release Pairs Executed	Acquire/Release Pairs Static
Serial	147.8	—	—	—
Original	217.2	46.9%	15,471,682	3
Bounded	191.7	29.7%	7,744,033	3
Aggressive Greedy	149.9	1.4%	49,152	2

Table I. Locking Overhead for Barnes-Hut (16384 bodies) on a Single Processor

Table II presents the execution times for the different parallel versions running on a variety of processors; Figure 15 presents the corresponding speedup curves. The speedups are calculated relative to the serial version of the code, which executes with no lock or parallelization overhead.[2] All versions scale well, which indicates that the compiler was able to effectively parallelize the application. Although the absolute performance varies with the false exclusion policy, the performance of the different parallel versions scales at approximately the same rate. This indicates that the lock elimination algorithm introduced no significant false exclusion.

[2]The speedup is the running time of the serial version divided by the running time of the parallel version. The speedup curves plot the speedup as a function of the number of processors executing the parallel version.

Version	Processors					
	1	2	4	8	12	16
Serial	147.8	—	—	—	—	—
Original	217.2	111.6	56.59	32.61	20.76	15.64
Bounded	191.7	97.25	49.22	26.98	19.62	15.12
Aggressive Greedy	149.9	76.30	37.81	21.88	15.57	12.87

Table II. Execution Times for Barnes-Hut (16384 bodies) (seconds)

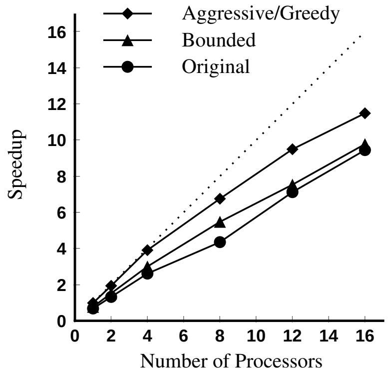


Fig. 15. Speedups for Barnes-Hut (16384 bodies)

6.4 Water

Table III presents the execution statistics for the single processor runs of Water. For this application, using the Aggressive false exclusion policy the compiler generates the same parallel code as using the Bounded policy. With no lock elimination, the synchronization overhead is 16% over the original serial version. Lock elimination with the Bounded, Aggressive and Greedy policies drives the overhead down substantially. As expected, the number of executed acquire and release constructs is correlated with the execution times.

Version	Execution Time	Execution Time Overhead	Acquire/Release Pairs Executed	Acquire/Release Pairs Static
Serial	159.5	—	—	—
Original	184.4	16%	4,200,448	3
Bounded Aggressive	175.8	10%	2,099,200	5
Greedy	165.3	4%	1,577,980	5

Table III. Locking Overhead for Water (512 molecules) on a Single Processor

Table IV presents the execution times for the different parallel versions running on a variety of processors; Figure 16 presents the corresponding speedup curves. The Original, Bounded and Aggressive versions initially perform well (the speedup over the sequential C++ version at eight processors is close to six). But both versions fail to scale beyond twelve processors. The Greedy version fails to scale well at all.

Version	Processors					
	1	2	4	8	12	16
Serial	159.5	—	—	—	—	—
Original	184.4	94.60	47.51	28.39	22.06	19.87
Bounded Aggressive	175.8	88.36	44.28	26.42	21.06	19.50
Greedy	165.3	115.2	88.45	79.18	75.16	73.54

Table IV. Execution Times for Water (512 molecules) (seconds)

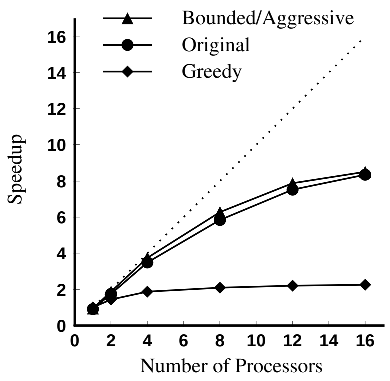


Fig. 16. Speedups for Water (512 molecules)

We instrumented the parallel code to determine the source of the performance loss. Figure 17 presents the *contention proportion*, which is the proportion of the time that processors spend waiting to acquire a lock held by another processor.[3] This figure clearly shows that lock contention is the primary cause of performance loss for this application, and that the Greedy false exclusion policy generates enough false exclusion to severely degrade the performance.

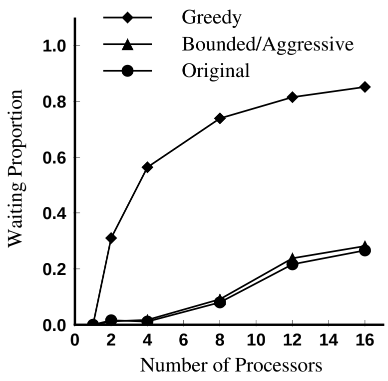


Fig. 17. Contention Proportion for Water (512 molecules)

6.5 String

Table V presents the execution statistics for the single processor runs of String. All experimental results are for the Big Well input data set. For this application, using the Bounded

[3] More precisely, the contention proportion is the sum over all processors of the amount of time that each processor spends waiting to acquire a lock held by another processor divided by the execution time times the number of processors executing the computation.

false exclusion policy the compiler generates the same parallel code as using the Original policy. With no lock elimination, the synchronization overhead is 18% over the original serial version. Lock elimination with the Aggressive and Greedy policies reduce the overhead to 6%. As expected, the number of executed acquire and release constructs is correlated with the execution times.

Version	Execution Time	Execution Time Overhead	Acquire/Release Pairs Executed	Acquire/Release Pairs Static
Serial	2208.9	—	—	—
Original Bounded Aggressive	2599.0	18%	30,286,506	1
Greedy	2337.7	6%	2,313	2

Table V. Locking Overhead for String (*big* well) on a Single Processor

Version	Processors 1	2	4	8	12	16
Serial	2208.9	—	—	—	—	—
Original Bounded Aggressive	2599.0	1289.4	646.7	331.9	223.9	172.3
Greedy	2337.7	2313.5	2231.9	2244.3	2254.8	2260.9

Table VI. Execution Times for String (seconds)

Table VI presents the execution times for the different parallel versions running on a variety of processors; Figure 18 presents the corresponding speedup curves. The Original, Bounded and Aggressive versions perform very well, attaining a speedup of more than 12 on 16 processors. The Greedy version fails to scale at all: the Greedy false exclusion policy serializes the entire computation.

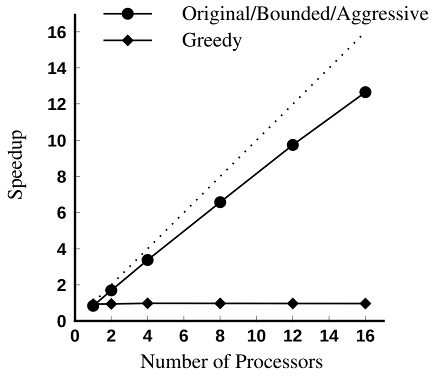


Fig. 18. Speedups for String (*big* well)

Figure 19 presents the contention proportion for this application. This figure shows that lock contention is the primary cause of performance loss for this application, and that it generates enough false exclusion to severely degrade the performance.

6.6 Discussion

Our experimental results indicate that the applications are originally locking objects at a very fine granularity generating a significant locking overhead. This is a natural consequence of a programming style that focuses on a layered set of data abstractions as pro-

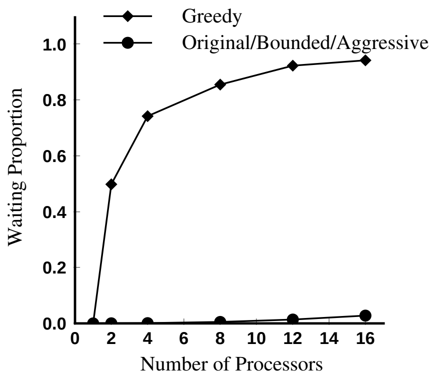


Fig. 19. Contention Proportion for String (*big* well)

moted, for example by object-oriented languages. For applications with these characteristics, and where a sequence of very fine grain oparations updated the same data object, lock elimination will be able to enlarge the mutual exclusion region and significantly reduce the locking overhead of the application.

In addition our experimental results indicate that different false exclusion policies have a significant impact in the resulting parallel program performance. For all of the benchmark applications, however, selecting the Aggressive false exclusion policy yields the best performance.

In general, we expect different applications to have distinct best false exclusion policies. Moreover, different parallel phases of the same application, might exhibit distinct best false exclusion policies. A production system could therefore choose a default policy, but allow the programmer to override the default to obtain better performance.

Alternatively, a production system could use run-time performance information to select the correct false exclusion policy for each of the parallel phases of the application. We envision sophisticated production system automatically instrumenting and using feedback mechanisms to dynamically bind program transformations [Diniz and Rinard 1997]. Using such feedback mechanisms, production system could automatically, and without programmer assistance, generate code that dynamically chooses the best false exclusion policy.

7. RELATED WORK

The closest related work is our own previous research on techniques to reduce lock overhead in automatically parallelized object-based programs [Diniz and Rinard 1996]. This research used a monolithic algorithm that depends heavily on the restrictions of the object-based programming paradigm and the fact that the compiler controls the placement of the acquire and release constructs. The algorithm is formulated as a set of conditions on the call graph. If the call graph meets the conditions, the compiler can omit the automatic insertion of synchronization constructs into some of the procedures. There is no clearly identified set of transformations, the algorithms are only capable of increasing the sizes of the critical sections, and they work only at the granularity of entire procedures.

This paper, on the other hand, describes a general set of transformations for programs that use mutual exclusion locks to implement critical sections. This flexible set of transformations enables the movement and cancellation of acquire and release constructs both within and across procedures.

The lock elimination algorithm in this paper is formulated as a reachability problem in the ICFG rather than as a set of conditions on a call graph. While the two algorithms yield identical results in the context of our prototype compiler, formulating the problem as a reachability problem means that the new algorithm inherits all of the advantages of the basic transformations. In particular, the new algorithm is more flexible and applies to explicitly parallel programs that already contain synchronization constructs. It is possible to apply the optimization both within and across procedures, rather than only at the granularity of procedures. The formulation also removes the dependence on the compiler's ability to control the placement of the synchronization constructs.

Plevyak, Zhang and Chien have developed a similar synchronization optimization technique, *access region expansion*, for concurrent object-oriented programs [Plevyak et al. 1995]. Because access region expansion is designed to reduce the overhead in sequential executions of such programs, it does not address the trade off between lock overhead and waiting overhead. The goal is simply to minimize the lock overhead.

7.1 Parallel Loop Optimizations

Other synchronization optimization research has focused almost exclusively on parallel loops in scientific computations [Midkiff and Padua 1987]. The natural implementation of a parallel loop requires two synchronization constructs: an initiation construct to start all processors executing loop iterations, and a barrier construct at the end of the loop. The majority of synchronization optimization research has concentrated on removing barriers or converting barrier synchronization constructs to more efficient synchronization constructs such as counters [Tseng 1995]. Several researchers have also explored optimizations geared towards exploiting more fine-grained concurrency available within loops [Cytron 1986]. These optimizations automatically insert one-way synchronization constructs such as post and wait to implement loop-carried data dependences.

The transformations and algorithms presented in this paper address a different problem. They are designed to optimize mutual exclusion synchronization, not barrier synchronization or post/wait synchronization. We believe, however, that would be possible and worthwhile to combine both classes of optimizations into a single unified synchronization optimization framework.

7.2 Analysis of Explicitly Parallel Programs

The transformations presented in this paper operate on explicitly parallel programs. Other researchers have investigated the issues associated with performing standard serial compiler analyses and optimizations in the presence of explicit concurrency [Chow and Harrison III 1992; Midkiff and Padua 1990]. Our research is orthogonal to this research in the sense that it focuses on optimization opportunities that appear only in explicitly parallel programs rather than on the significant challenges associated with applying standard optimizations to parallel programs.

7.3 Concurrent Constraint Programming

The lock movement transformations are reminiscent of transformations from the field of concurrent constraint programming that propagate *tell* and *ask* constructs through the program [Saraswat et al. 1991]. The goal is to make tells and corresponding asks adjacent in the program. This adjacency enables an optimization that removes the ask construct. A difference is the asymmetry of asks and tells: the optimization that eliminates the ask leaves the tell in place. The lock cancellation transformation, of course, eliminates both the acquire and the release.

7.4 Efficient Synchronization Algorithms

Other researchers have addressed the issue of synchronization overhead reduction. This work has concentrated on the development of more efficient implementations of synchronization primitives using various protocols and waiting mechanisms [Goodman et al. 1989; Lim and Agarwal 1994].

The research presented in this paper is orthogonal to and synergistic with this work. Lock elimination reduces the lock overhead by reducing the frequency with which the generated parallel code acquires and releases locks, not by providing a more efficient implementation of the locking constructs.

8. CONCLUSION

As parallel computing becomes part of the mainstream computing environment, compilers will need to apply synchronization optimizations to deliver efficient parallel software. This paper describes a new framework for synchronization optimizations, a new set of transformations for programs that implement critical sections using mutual exclusion locks, and a new synchronization optimization algorithm for reducing synchronization overhead in such programs. Experimental results from a parallelizing compiler for object-based programs illustrate the practical utility of this optimization. For three benchmark programs, the optimization dramatically reduces the number of times the computations acquire and release locks, which significantly reduces the amount of time processors spend acquiring and releasing locks. For one of the three benchmarks, the optimization always significantly improves the overall performance. Depending on the number of processors executing the computation, the optimized version runs between 1.03 and 1.30 times faster than the unoptimized version. For one of the other benchmarks, the optimized version runs between 1.07 and 1.02 times faster than the unoptimized version, with a mean of 1.06 times faster. For the final benchmark, the optimization does not improve the overall performance.

REFERENCES

BARNES, J. AND HUT, P. 1986. A hierarchical O(NlogN) force calculation algorithm. *Nature 324,* 4 (Dec.), 446–449.

CARDELLI, L. AND PIKE, R. 1985. Squeak: a language for communicating with mice. In *SIGGRAPH*. ACM Press, New York, NY.

CHOW, J. AND HARRISON III, W. 1992. Compile-time analysis of parallel programs that share memory. In *Proceedings of the Nineteenth Annual ACM Symposium on the Principles of Programming Languages*. ACM Press, New York, NY, 130–141.

CYTRON, R. 1986. Doacross: Beyond vectorization for multiprocessors. In *Proceedings of the 1986 International Conference on Parallel Processing*. IEEE Computer Society Press, Los Alamitos, CA.

DINIZ, P. AND RINARD, M. 1996. Lock coarsening: Eliminating lock overhead in automatically parallelized object-based programs. In *In Proceedings of the Nineth Workshop on Languages and Compilers for Parallel Computing*. Vol. 1239. Springer-Verlag, Berlin, 285–299. Published as Lecture Notes in Computer Science, Vol 1239.

DINIZ, P. AND RINARD, M. 1997. Dynamic feedback: An effective technique for adaptive computing. In *Proceedings of the ACM SIGPLAN '97 Conference on Program Language Design and Implementation*. ACM Press, New York, NY, 71–84.

EMAMI, M., GHIYA, R., AND HENDREN, L. 1994. Context-sensitive interprocedural points-to analysis in the presence of function pointer. In *Proceedings of the ACM SIGPLAN '94 Conference on Program Language Design and Implementation*. ACM Press, New York, NY, 242–256.

GOODMAN, J., VERNON, M., AND WOEST, P. 1989. Efficient synchronization primitives for large-scale cache-coherent multiprocessors. In *Proceedings of the Third International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM Press, New York, NY, 64–75.

HARRIS, J., LAZARATOS, S., AND MICHELENA, R. 1990. Tomographic string inversion. In *60th Annual International Meeting, Society of Exploration and Geophysics, Extended Abstracts*. 82–85.

HAUSER, C., JACOBI, C., THEIMER, M., WELCH, B., AND WEISER, M. 1993. Using threads in interactive systems: A case study. In *Proceedings of the Fourteenth ACM Symposium on Operating Systems Principles*. Ashville, NC.

LANDI, W., RYDER, B., AND ZHANG, S. 1993. Interprocedural modification side effect analysis with pointer aliasing. In *Proceedings of the ACM SIGPLAN '93 Conference on Program Language Design and Implementation*. ACM Press, New York, NY, 56–67.

LENOSKI, D. 1992. The design and analysis of DASH: A scalable directory-based multiprocessor. Ph.D. thesis, Stanford, CA.

LIM, B.-H. AND AGARWAL, A. 1994. Reactive synchronization algorithms for multiprocessors. In *Proceedings of the Sixth International Conference on Architectural Support for Programming Languages and Operating Systems*. ACM Press, New York, NY, 25–37.

MIDKIFF, S. AND PADUA, D. 1987. Compiler algorithms for synchronization. *IEEE Transactions on Computers 36,* 12 (Dec.), 1485–1495.

MIDKIFF, S. AND PADUA, D. 1990. Issues in the optimization of parallel programs. In *Proceedings of the 1990 International Conference on Parallel Processing*. IEEE Computer Society Press, Los Alamitos, CA, II–105–113.

PLEVYAK, J., ZHANG, X., AND CHIEN, A. 1995. Obtaining sequential efficiency for concurrent object-oriented languages. In *Proceedings of the Twenty-second Annual ACM Symposium on the Principles of Programming Languages*. ACM Press, New York, NY.

REPPY, J. 1992. Higher–order concurrency. Ph.D. thesis, Dept. of Computer Science, Cornell University.

REPS, T., HOROWITZ, S., AND SAGIV, M. 1995. Precise interprocedural dataflow analysis via graph reachability. In *Proceedings of the Twenty-second Annual ACM Symposium on the Principles of Programming Languages*. ACM Press, New York, NY, 49–61.

RINARD, M. AND DINIZ, P. 1996. Commutativity analysis: A new framework for parallelizing compilers. In *Proceedings of the ACM SIGPLAN '96 Conference on Program Language Design and Implementation*. ACM Press, New York, NY, 54–67.

SARASWAT, V., RINARD, M., AND PANANGADEN, P. 1991. Semantic foundations of concurrent constraint programming. In *Proceedings of the Eighteenth Annual ACM Symposium on the Principles of Programming Languages*. ACM Press, New York, NY, 333–352.

SINGH, J., WEBER, W., AND GUPTA, A. 1992. SPLASH: Stanford parallel applications for shared memory. *Computer Architecture News 20,* 1 (March), 5–44.

TSENG, C. 1995. Compiler optimizations for eliminating barrier synchronization. In *Proceedings of the Fifth ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*. ACM Press, New York, NY, 144–155.

WILSON, R. AND LAM, M. 1995. Efficient context-sensitive pointer analysis for C programs. In *SIGPLAN '95 Conference on Programming Language Design and Implementation,* SIGPLAN Notices 30(6). ACM Press, New York, NY, 1–12.

A. LOCK MOVEMENT ALGORITHMS

The lock movement algorithms are given an expression e, read set r and node N with read set r_N and write set w_N. They compute the new expression e' and read sets $r_1, \ldots, r_m, s_1, \ldots, s_n$ required to move a synchronization node with expression e and read set r across N. There are several potential complications:

—The compiler may be unable to generate the new expression e'. This may happen, for example, if N writes variables in the read set r. In this case the compiler cannot apply the transformation.

—If the algorithm moves an acquire node over a node with multiple outgoing edges or a release node over a node with multiple incoming edges, the new expressions may be evaluated in different contexts than the original expression. The compiler must therefore ensure that the evaluation of the expressions in the new contexts does not generate an error and that the new nodes always acquire or release a valid lock.

In general, the compiler may have to use an interprocedural pointer or alias analysis to verify that these conditions hold [Emami et al. 1994; Wilson and Lam 1995; Landi et al. 1993]. In restricted contexts, the compiler may be able to use simpler algorithms.

Our prototype compiler, for example, is designed for object-based programs. These programs structure the computation as a sequence of *operations* on *objects*. Each object contains a lock. Within an operation on an object, the expression that denotes the object's lock always denotes a valid lock.

—Whenever the new expressions may be evaluated in a different context than the original expression, the new read sets must reflect the behavior of the expressions in the new contexts. In general, the compiler may have to use an interprocedural pointer or alias analysis to compute the new read sets. The structure of the object-based paradigm allows the compiler to use a simpler approach. Expressions in object-based programs contain only local variables, parameters, and references to instance variables of objects. For a given expression, the read set is simply the set of variables that appear in the expression. The read set therefore depends only on the expression, not the context in which it appears.

Here are the cases that the acquire lock movement algorithm must handle:

—type(N) = entry. (The acquire node moves out of a callee towards the caller and can no longer access the local variables of the callee.) e must contain no local variables (but may contain parameters); $e' = e$ and $s_1, \ldots, s_n = r$. The compiler may be able to generate more precise read sets for $s_1, \ldots, s_n$.

—type(N) = exit. (The acquire node moves into a callee from a caller.) $e' = e$. The compiler must verify that e always denotes a valid lock in all of the new contexts. It must also generate new read sets $s_1, r_1, \ldots, r_m$ to reflect the variables that e may read in the new contexts. For object-based programs, $s_1, r_1, \ldots, r_m = r$.

—type(N) = merge (The acquire node moves up into several different flow of control paths.) $e' = e$ and $s_1, \ldots, s_n = r$. The compiler may be able to generate more precise read sets for $s_1, \ldots, s_n$.

—type(N) = call. (The acquire node moves into the caller of the procedure that it just moved out of.) $e' = e$ with the expressions in the call node that denote the values of the actual parameters substituted in for the corresponding formal parameters in e. If e contains any formal parameters, $s_1 = r \cup r_N$, otherwise $s_1 = r$. The compiler may be able to generate a more precise read set s_1.

—type(N) = return. (The acquire node moves out of a caller towards a callee). In the new context the acquire node will no longer have access to the local naming environment of the caller.) Expression e must contain no local variables or parameters; $e' = e$ and $s_1 = r$.

—type(N) = summary. If $w_N \cap r = \emptyset$, (if the summary writes no variables that the release reads) $e' = e$ and $s_1 = r$. Otherwise the transformation can not be applied.

—type(N) = assignment. There are several cases:

—$w_N \cap r = \emptyset$. (N writes no variables that the acquire node may read.) $e' = e$ and

$s_1 = r$.

—The assignment is of the form $v = exp$, where v is a local variable. Expression e must not dereference a pointer variable that may point to v; $e' = e$ with exp substituted in for v in e. If expression e contains at least one occurrence of v, $s_1 = r \cup r_N$, otherwise $s_1 = r$. The compiler may be able to generate a more precise read set s_1.

—Otherwise the transformation can not be applied.

—type(N) = if. (The acquire node moves across a branch.) The compiler must verify that expression e always denotes a valid lock in all of the new contexts. It must also generate new read sets $s_1, r_1, \ldots, r_m$ to reflect the variables that e may read in the new contexts. For object-based programs, $s_1, r_1, \ldots, r_m = r$.

—type(N) = acquire or release. The transformation can not be applied.

The release lock movement algorithm must handle the following cases:

—type(N) = entry. (The release node is moving into a procedure from the caller of the procedure.) $e' = e$. The compiler must verify that expression e always denotes a valid lock in all of the new contexts. It must also generate new read sets $r_1, \ldots, r_m, s_1$ to reflect the variables that e may read in the new contexts. For object-based programs, $r_1, \ldots, r_m, s_1 = r$.

—type(N) = exit. (The release node is moving out of a callee towards the caller, and can no longer access the local variables of the callee.) Expression e must contain no local variables (but may contain parameters); $e' = e$ and $s_1, \ldots, s_n = r$. The compiler may be able to generate more precise read sets for $s_1, \ldots, s_n$.

—type(N) = merge. $e' = e$. The compiler must verify that expression e always denotes a valid lock in all of the new contexts. It must also generate new read sets $r_1, \ldots, r_m, s_1$ to reflect the variables that e may read in the new contexts. For object-based programs, $r_1, \ldots, r_m, s_1 = r$.

—type(N) = call. (The release node is moving out of a caller towards a callee and no longer has access to the local variables or parameters of the caller.) e must contain no local variables or parameters; e' = e and $s_1 = r$.

—type(N) = return. (The release node is moving into a caller from a callee.) First find the call node that corresponds to N. None of the nodes in the invoked procedure or any procedures that it directly or indirectly invokes may write any of the variables in the call node's read set. $e' = e$ with the expressions in the call node that denote the values of the actual parameters substituted in for the corresponding formal parameters in e. If expression e contains any formal parameters, $s_1 = r \cup r_N$, otherwise $s_1 = r$. The compiler may be able to generate more precise read sets.

—type(N) = assignment or type(N) = summary. If $w_N \cap r = \emptyset$, (if the assignment or summary writes no variables that the release reads) $e' = e$ and $s_1 = r$. Otherwise the transformation can not be applied.

—type(N) = if. $e' = e$ and $s_1, \ldots, s_n = r$. The compiler may be able to generate more precise read sets for $s_1, \ldots, s_n$.

—type(N) = acquire or release. The transformation can not be applied.

B. LOCK ELIMINATION ALGORITHM

The lock elimination algorithm uses the following primitives.

—`invokedProcedures`(p) : the set of procedures directly or indirectly invoked by procedure p.

—`procedure`(N) : the procedure that the ICFG node N is in.

—`type`(N) : the type of the ICFG node N.

—`predecessor`(N) : the predecessor of N in the ICFG. Only valid for nodes with one predecessor.

—`successor`(N) : the successor of N in the ICFG. Only valid for nodes with one successor.

—`predecessors`(N) : the set of predecessors of N in the ICFG.

—`successors`(N) : the set of successors of N in the ICFG.

—`insertNode`$(\langle N_{from}, N_{to}\rangle, t, \langle e, r\rangle)$: insert a new node into the ICFG whose type is t, expression is e and read set is r. There is an edge from N_{from} to the new node and an edge from the new node to N_{to}. Remove the edge $\langle N_{from}, N_{to}\rangle$ from the ICFG.

—$\texttt{removeNode}(N)$: remove node N from the ICFG. Make all predecessors of N predecessors of the successor of N and all successors of N successors of the predecessor of N. Only valid for nodes with one successor and one predecessor.

—$\langle e', r_1, \ldots, r_m, s_1, \ldots, s_n \rangle = \texttt{acquireTransform}(N, e, r)$: computes the new expression e' and read sets r_1, ..., r_m, s_1, ...,s_n that result from moving an acquire node with expression e and read set r across node N. If the transformation cannot be applied, $e' = \epsilon$.

—$\langle e', r_1, \ldots, r_m, s_1, \ldots, s_n \rangle = \texttt{releaseTransform}(N, e, r)$: computes the new expression e' and read sets $r_1, \ldots, r_m, s_1, \ldots, s_n$ that result from moving a release node with expression e and read set r across node N. If the transformation cannot be applied, $e' = \epsilon$.

```
// lockElimination(p) applies the lock elimination algorithm to the procedure p.
lockElimination(p){
  do {
    ps = invokedProcedures(p);
    ns = {N : procedure(N) ∈ ps};
    ns_acq = {N ∈ ns : type(N) = acquire};
    ns_rel = {N ∈ ns : type(N) = release};
    applied = false;
    for all ⟨N_acq, N_rel⟩ ∈ {ns_acq × ns_rel}
      if(attemptTransform(N_acq, N_rel, ns))
        applied = true;
        break;
  } while (applied = true);
}
```

Fig. 20. Lock Elimination Algorithm

```
// attemptTransform(N_acq, N_rel, ns) attempts to propagate and cancel N_acq and N_rel.
// To implement the false exclusion policy, the transformation must be confined to the set of nodes ns.
attemptTransform(N_acq, N_rel, ns){
  // Step 1. Compute the reachability trees, expressions and read sets for the acquire and release nodes.
  ⟨parent_acq, edges_acq, ed_acq, visted_acq⟩ =
    acquireTraverse(predecessor(N_acq), N_acq, exp(N_acq), read(N_acq), ∅, ∅, ∅, ∅, ns);
  ⟨parent_rel, edges_rel, ed_rel, visted_rel⟩ =
    releaseTraverse(N_rel, successor(N_rel), exp(N_rel), read(N_rel), ∅, ∅, ed_acq, ∅, ns);
  // Step 2. Check if the two reachability trees intersect and if the acquire and release
  // manipulate the same lock.
  if(∃ ⟨N_from, N_to⟩ ∈ edges_acq ∩ edges_rel : ed_acq(⟨N_from, N_to⟩) = ed_rel(⟨N_from, N_to⟩)) {
    choose any ⟨N_from, N_to⟩ ∈ edges_acq ∩ edges_rel;
    // Step 3. Find the path in the reachability trees from the release node to the acquire node.
    path = computePath(N_to, N_acq, parent_acq) ∪ computePath(N_from, N_rel, parent_rel);
    // Step 4. Find the edges coming into the path and the edges going out of the path.
    // The algorithm will insert new acquire nodes on all of the incoming edges and new
    // release nodes on all of the outgoing edges.
    acq = ⋃_{N∈path} (predecessors(N) − (path ∪ {N_rel})) × {N};
    rel = ⋃_{N∈path} {N} × (successors(N) − (path ∪ {N_acq}));
    // Step 5. Insert new acquire and release nodes.
    for all ⟨N_from, N_to⟩ ∈ acq do
      insertNode(⟨N_from, N_to⟩, acquire, ed_rel(⟨N_from, N_to⟩));
    for all ⟨N_from, N_to⟩ ∈ rel do
      insertNode(⟨N_from, N_to⟩, release, ed_rel(⟨N_from, N_to⟩));
    // Step 6. Remove original acquire and release nodes.
    removeNode(N_acq);
    removeNode(N_rel);
    return true;
  } else {
    return false;
  }
}
```

Fig. 21. `attemptTransform` Algorithm

```
// computePath computes the path from the node N in parent back to N_final.
computePath(N, N_final, parent){
  if(N = N_final)
    return ∅
  else
    return {N} ∪ computePath(parent(N), N_final, parent);
}
```

Fig. 22. `computePath` Algorithm

```
// acquireTraverse computes the reachability tree for an acquire node. At each step
// it computes the result of moving the acquire node across another node in the ICFG.
// The meanings of the variables are as follows:
// ⟨N_from, N_to⟩: edge that the traversal reached in the last step.
//    The traversal will next try to move the acquire node back across N_from.
// e and r: expression and read set that result from propagating the acquire node through
//    the ICFG to the edge ⟨N_from, N_to⟩.
// parent: partial function from ICFG nodes to ICFG nodes.
//    It records the path back to the acquire node.
// edges: set of ICFG edges reachable by propagating the acquire node through the ICFG
//    against the flow of control.
// ed: partial function from ICFG edges to tuples of expressions and read sets.
//    For each edge it records the expression and read set that would result from propagating
//    the acquire node back through the ICFG to that edge.
// visited: set of ICFG nodes already visited by the traversal.
// ns: set of ICFG nodes. To implement the false exclusion policy, the tree must stay within this set.
acquireTraverse(N_from, N_to, e, r, parent, edges, ed, visited, ns){
  edges = edges ∪ {⟨N_from, N_to⟩};
  ed = ed[⟨N_from, N_to⟩ ↦ ⟨e, r⟩];
  if(N_from ∉ visited and N_from ∈ ns) {
    visited = visited ∪ {N_from};
    parent = parent[N_from ↦ N_to];
    // Compute the new expression and read set that result from moving the acquire across N_from.
    ⟨e′, r_1, ..., r_m, s_1, ..., s_n⟩ = acquireTransform(N_from, e, r);
    if(e′ ≠ ε) {
      // Record the expression and read set for any new release nodes.
      for all N_i ∈ successors(N_from) − {N_to} do
        ed = ed[⟨N_from, N_i⟩ ↦ ⟨e, r_i⟩];
      for all N_i ∈ predecessors(N_from) do
        ⟨parent, edges, ed, visited⟩ =
          acquireTraverse(N_from, N_i, e′, s_i, parent, edges, ed, visited, ns);
    }
  }
  return ⟨parent, edges, ed, visited⟩;
}
// releaseTraverse computes the reachability tree for a
// release node. Essentially the dual of acquireTraverse.
releaseTraverse(N_from, N_to, e, r, parent, edges, ed, visited, ns){
  edges = edges ∪ {⟨N_from, N_to⟩};
  ed = ed[⟨N_from, N_to⟩ ↦ ⟨e, r⟩];
  if(N_to ∉ visited and N_to ∈ ns) {
    visited = visited ∪ {N_to};
    parent = parent[N_to ↦ N_from];
    ⟨e′, r_1, ..., r_m, s_1, ..., s_n⟩ = releaseTransform(N_to, e, r);
    if(e′ ≠ ε) {
      for all N_i ∈ predecessors(N_to) − {N_from} do
        ed = ed[⟨N_i, N_to⟩ ↦ ⟨e, r_i⟩];
      for all N_i ∈ successors(N_to) do
        ⟨tree, ed, visited⟩ =
          releaseTraverse(N_to, N_i, e′, s_i, parent, edges, ed, visited, ns);
    }
  }
  return ⟨parent, edges, ed, visited⟩;
}
```

Fig. 23. `acquireTraverse` and `releaseTraverse` Algorithms