LABORATORY FOR COMPUTER SCIENCE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MIT/LCS/TM-449



# LAZY TASK CREATION: A TECHNIQUE FOR INCREASING THE GRANULARITY OF PARALLEL PROGRAMS

Eric Mohr
David Kranz
Robert H. Halstead, Jr.



June 1991

545 TECHNOLOGY SQUARE, CAMBRIDGE, MASSACHUSETTS 02139

SECURITY CLASSIFICATION OF THIS PAGE

# REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a. REPORT SECURITY CLASSIFICATION<br>Unclassified | 1b. RESTRICTIVE MARKINGS |
| 2a. SECURITY CLASSIFICATION AUTHORITY | 3. DISTRIBUTION/AVAILABILITY OF REPORT<br>Approved for public release; distribution is unlimited. |
| 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S)<br>MIT/LCS/TM 449 | 5. MONITORING ORGANIZATION REPORT NUMBER(S)<br>N00014-87-K-0825 |
| 6a. NAME OF PERFORMING ORGANIZATION<br>MIT Lab for Computer Science<br>6b. OFFICE SYMBOL (If applicable) | 7a. NAME OF MONITORING ORGANIZATION<br>Office of Naval Research/Dept. of Navy |
| 6c. ADDRESS (City, State, and ZIP Code)<br>545 Technology Square<br>Cambridge, MA 02139 | 7b. ADDRESS (City, State, and ZIP Code)<br>Information Systems Program<br>Arlington, VA 22217 |
| 8a. NAME OF FUNDING/SPONSORING ORGANIZATION<br>DARPA/DOD<br>8b. OFFICE SYMBOL (If applicable) | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER |
| 8c. ADDRESS (City, State, and ZIP Code)<br>1400 Wilson Blvd.<br>Arlington, VA 22217 | 10. SOURCE OF FUNDING NUMBERS<br>PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO / WORK UNIT ACCESSION NO. |

11. TITLE (Include Security Classification)

Lazy Task Creation: A Technique for Increasing the Granularity of Parallel Programs

12. PERSONAL AUTHOR(S)

Eric Mohr, David Kranz, Robert H. Halstead, Jr.

| 13a. TYPE OF REPORT | 13b. TIME COVERED | 14. DATE OF REPORT (Year, Month, Day) | 15. PAGE COUNT |
|---|---|---|---|
| Technical | FROM ____ TO ____ | June 1991 | 22 |

16. SUPPLEMENTARY NOTATION

| 17. COSATI CODES | | |
|---|---|---|
| FIELD | GROUP | SUB-GROUP |
| | | |
| | | |

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

load balancing, parallel Lisp, parallel programming languages, process migration, program partitioning, task management

19. ABSTRACT (Continue on reverse if necessary and identify by block number)


Many parallel algorithms are naturally expressed at a fine level of granularity, often finer than a MIMD parallel system can exploit efficiently. Most builders of parallel systems have looked to either the programmer or a parallelizing compiler to increase the granularity of such algorithms. In this paper we explore a third approach to the granularity problem by analyzing two strategies for combining parallel tasks dynamically at run-time. We reject the simpler *load-based inlining* method, where tasks are combined based on dynamic load level, in favor of the safer and more robust *lazy task creation* method, where tasks are created only retroactively as processing resources become available.

These strategies grew out of work on Mul-T [17], an efficient parallel implementation of Scheme, but could be used with other languages as well. We describe our Mul-T implementations of lazy task creation for two contrasting machines, and present performance statistics which show the method's effectiveness. Lazy task creation allows efficient execution of naturally expressed algorithms of a substantially finer grain than possible with previous parallel Lisp systems.

Earlier versions of this paper appeared as [20] and [21].

**Index terms:** load balancing, parallel Lisp, parallel programming languages, process migration, program partitioning, task management.


| | |
|---|---|
| 20. DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS | 21. ABSTRACT SECURITY CLASSIFICATION<br>Unclassified |
| 22a. NAME OF RESPONSIBLE INDIVIDUAL<br>Carol Nicolora | 22b. TELEPHONE (Include Area Code)<br>(617) 253-5894<br>22c. OFFICE SYMBOL |

# Lazy Task Creation:
# A Technique for Increasing the Granularity of Parallel Programs*

Eric Mohr
Yale University
mohr@cs.yale.edu

David A. Kranz
M.I.T.
kranz@ai.mit.edu

Robert H. Halstead, Jr.
DEC Cambridge Research Lab
halstead@crl.dec.com

## Abstract

Many parallel algorithms are naturally expressed at a fine level of granularity, often finer than a MIMD parallel system can exploit efficiently. Most builders of parallel systems have looked to either the programmer or a parallelizing compiler to increase the granularity of such algorithms. In this paper we explore a third approach to the granularity problem by analyzing two strategies for combining parallel tasks dynamically at run-time. We reject the simpler ***load-based inlining*** method, where tasks are combined based on dynamic load level, in favor of the safer and more robust ***lazy task creation*** method, where tasks are created only retroactively as processing resources become available.

These strategies grew out of work on Mul-T [17], an efficient parallel implementation of Scheme, but could be used with other languages as well. We describe our Mul-T implementations of lazy task creation for two contrasting machines, and present performance statistics which show the method's effectiveness. Lazy task creation allows efficient execution of naturally expressed algorithms of a substantially finer grain than possible with previous parallel Lisp systems.

Earlier versions of this paper appeared as [20] and [21].

**Index terms:** load balancing, parallel Lisp, parallel programming languages, process migration, program partitioning, task management.

## 1 Introduction

There have been numerous proposals for implementations of applicative languages on parallel computers. All have in some way come up against a granularity problem—when a parallel algorithm is written naturally, the resulting program often produces tasks of a finer grain than an implementation can exploit efficiently. Some researchers look to hardware specially designed to handle fine-grained tasks [3, 11], while others have looked for ways to increase task granularity by grouping a number of potentially parallel operations together into a single sequential thread. These latter efforts can be classified by the degree of programmer involvement required to specify parallelism, from parallelizing compilers at one end of the spectrum to language constructs giving the programmer a fine degree of control at the other.

In the most attractive world, the programmer leaves the job of identifying parallel tasks to a parallelizing compiler. To achieve good performance, the compiler must create tasks of sufficient size based on estimating the cost of various pieces of code [8, 16, 25]. But when execution paths are highly data-dependent (as for example with recursive symbolic programs), the cost of a piece of code is often unknown at compile time. If only known costs are used, the tasks produced may still be too fine-grained. And for languages that allow mutation of shared variables it can be quite complex to determine where parallel execution is safe, and opportunities for parallelism may be missed.

At the other end of the spectrum a language can leave granularity decisions up to the programmer, possibly providing tools for building tasks of acceptable granularity such as the ***propositional parameters*** of Qlisp [7, 9, 10]. Such fine control can be necessary in some cases to maximize performance, but there are costs in programmer effort and program clarity. Also, any parameters appearing in the program require experimentation to calibrate; this work may have to be repeated for a different target machine or data set. Or, when the code is run in parallel with other code or on a multiuser machine, a given parameterization may be ineffective because the amount of resources available for that code is unpredictable. Similar problems arise when a parallelizing compiler is parameterized with details of a certain machine.

We've taken an intermediate position in our research on Mul-T [17], a parallel version of Scheme based on the `future` construct of Multilisp [13, 14]. The programmer takes on the burden of identifying *what* can be computed safely in parallel, leaving the decision of exactly

*Abridged version published in IEEE Transactions on Parallel and Distributed Systems, July 1991

*how* the division will take place to the run-time system. In Mul-T that means annotating programs with **future** to identify parallelism without worrying about granularity; the programmer's task is to *expose* parallelism while the system's task is to *limit* parallelism.

In our experience with the mostly functional style common to Scheme programs, a program's parallelism can often be expressed quite easily by adding a small number of **future** forms (which however may yield a large number of concurrent tasks at run time). The effort involved is little more than that required for systems with parallelizing compilers, where the programmer must be sure to code in such a way that parallelism is available.

In order to support this programming style we must deal with questions of efficiency. The Encore Multimax[1] implementation of Mul-T [17], based on the T system's Orbit compiler [18, 19], is proof that the underlying parallel Lisp system can be made efficient enough; we must now figure out how to achieve sufficient task granularity. For this we look to dynamic mechanisms in the run-time system, which have the advantage of avoiding the parameterization problems mentioned earlier. The key to our dynamic strategies for controlling granularity is the fact that that the **future** construct[2] has several correct operational interpretations. The canonical **future** expression

($K$ (**future** $X$))

declares that a child computation $X$ may proceed in parallel with its parent continuation $K$. In the most straightforward interpretation, a child task is created to compute $X$ while the parent task computes $K$. Reversing the task roles is also possible; the parent task can compute $X$ while the child task computes $K$. Finally, and most importantly for fine-grained programs, it is also usually correct for the parent task to compute first $X$ and then $K$, ignoring the **future**. This *inlining* of $X$ by the parent task eliminates the overhead of creating and scheduling a separate task and creating a placeholder to hold its value.[3]

Inlining can mean that a program's *run-time granularity* (the size of tasks actually executed at run time) is significantly greater than its *source granularity* (the size of code within the **future** constructs of the source program). A program will execute efficiently if its average run-time granularity is large compared to the overhead of task creation, providing of course that enough parallelism has been preserved to achieve good load balancing.

The first dynamic strategy we consider is *load-based inlining*. In this strategy, (**future** $X$) means, "If the system is not loaded, make a separate task to evaluate $X$; otherwise inline $X$, evaluating it in the current task." A load threshold $T$ indicates how many tasks must be queued before the system is considered to be loaded. Whenever a call to **future** is encountered, a simple check of task queue length determines whether or not a separate task will be created.

The simple load-based inlining strategy works well on some programs, but its several drawbacks (see Section 3) led us to consider another strategy as well: why not inline every task provisionally, but save enough information so that tasks can be selectively "un-inlined" as processing resources become available? In other words, create tasks lazily. With this *lazy task creation* strategy, ($K$ (**future** $X$)) means "Start evaluating $X$ in the current task, but save enough information so that its continuation $K$ can be moved to a separate task if another processor becomes idle." We say that idle processors *steal* tasks from busy processors; task stealing becomes the primary means of spreading work in the system.

The execution tree of a fine-grained program has an overabundance of potential fork points. Our goal with lazy task creation is to convert a small subset of these to actual forks, maximizing run-time task granularity while preserving parallelism and achieving good load balancing. In the subsequent discussion, this is contrasted with *eager task creation*, where all fork points result in a separate task.

An example will help make these ideas more concrete.

## 2 An Example

As a simple example of the spectrum of possible solutions to the granularity problem, consider the following algorithm (written as a Scheme program) to sum the leaves of a binary tree:

```
(define (sum-tree tree)
  (if (leaf? tree)
      (leaf-value tree)
      (+ (sum-tree (left tree))
         (sum-tree (right tree)))))
```

(where **leaf?**, **leaf-value**, **left**, and **right** define the tree datatype). The natural way to express parallelism in this algorithm is to indicate that the two recursive calls to **sum-tree** can proceed in parallel. In Mul-T we might indicate this by adding one **future**:[4]

[1] Multimax is a trademark of Encore Computer Corporation.

[2] (**future** $X$) returns an object called a *future*, a placeholder for the eventual value of $X$. The placeholder is said to be *unresolved* until $X$'s value becomes available. Any task attempting to use the value of an unresolved future is suspended until the value is available. A *touch* is a use of a value $V$ that will cause a task to be suspended if $V$ is an unresolved future.

[3] Such inlining is not always correct; sometimes it can lead to deadlock as described in Section 3.3.

[4] This strategy for adding **future** relies on + evaluating its operands from left to right; if argument evaluation went from

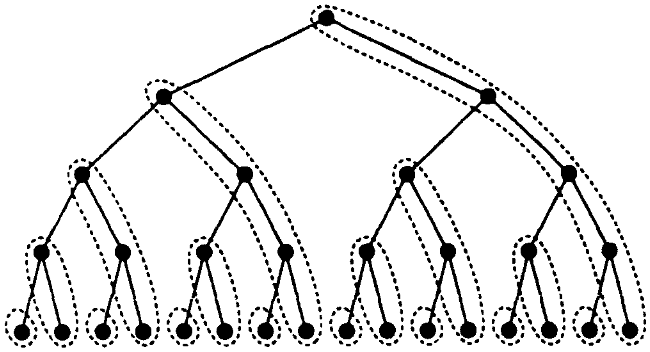


Figure 1: Direct execution of psum-tree.

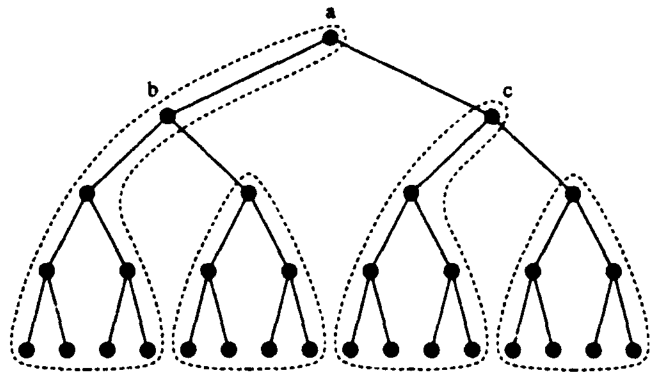



Figure 2: BUSD execution of psum-tree on 4 processors.

```
(define (psum-tree tree)
  (if (leaf? tree)
      (leaf-value tree)
      (+ (future (psum-tree (left tree)))
         (psum-tree (right tree)))))
```

The natural expression of parallelism in this algorithm is rather fine-grained. With eager task creation this program would create $2^d$ tasks to sum a tree of depth $d$; the average number of tree nodes handled by a task would be 2. Figure 1 shows this execution pictorially; each circled subset of tree nodes is handled by a single task. Unless task creation is very cheap, this task breakdown is likely to lead to poor performance.

The ideal task breakdown is one which maximizes the run-time task granularity while maintaining a balanced load. For a divide-and-conquer program like this one, that means expanding the tree breadth-first by spawning tasks until all processors are busy, and then expanding the tree depth-first within the task on each processor. We will refer to this ideal task breakdown as BUSD (Breadth-first Until Saturation, then Depth-first). Figure 2 shows this execution pictorially for a system with 4 processors.

How can we achieve this ideal task breakdown? A parallelizing compiler might be able to increase granularity by unrolling the recursion and eliminating some futures, but in this example we *want* fine-grained tasks at the beginning so as to spread work as quickly as possible (breadth-first). The compiler might possibly produce code to do this as well if supplied with information about available processing resources, but making such a transformation general is a difficult task and would still have the parameterization drawbacks noted earlier.

What if we control task creation explicitly as in Qlisp? In many of Qlisp's parallel constructs the programmer may supply a predicate which, when evaluated at run time, will determine whether or not a separate task is created. (One such predicate, (qemptyp) [10], tests the length of the work queue, achieving the same effect as our load-based inlining.) We might use Qlisp's **spawn** construct (equivalent to **future** with an additional predicate argument) to rewrite **psum-tree**; the style of this program **psum-tree-2** is very similar to an example in [7]:

```
(define (psum-tree-2 tree cutoff-depth)
  (if (leaf? tree)
      (leaf-value tree)
      (+ (spawn (> cutoff-depth 0)
           (psum-tree-2 (left tree)
                        (- cutoff-depth 1)))
         (psum-tree-2 (right tree)
                      (- cutoff-depth 1)))))
```

In this example, **cutoff-depth** specifies a depth beyond which no tasks should be created. The predicate (> cutoff-depth 0) tells **spawn** whether or not to inline the recursive call. A **cutoff-depth** value of 2 would achieve BUSD execution similar to that shown in Figure 2 (actually its mirror image); below level 2 all futures are inlined.

This solution has two problems. First, the code has become more complex by the addition of **cutoff-depth**—it is no longer completely straightforward to tell what this program is doing. Second, the program is now parameterized by the **cutoff-depth** argument, with the associated calibration issues noted previously.

Load-based inlining and lazy task creation are both attempts to approximate the BUSD performance of **psum-tree-2** without sacrificing the clarity of **psum-tree**. In an ideal run of **psum-tree** on a four-processor system with load-based inlining, the first three occurrences of **future** (at nodes $a$, $b$, and $c$ of Figure 2) find that processors are free, and separate tasks

right to left, then (psum-tree (right tree)) would evaluate to completion before (future (psum-tree (left tree))) began, and no parallelism would be realized.

are created (breadth-first). Depending on the value of the load threshold parameter $T$, a few more tasks may be created before the backlog is high enough to cause inlining. But since there is a large surplus of work, most tasks are able to defray the cost of their creation by inlining a substantial subtree (depth-first).

In an ideal run of **psum-tree** with lazy task creation, the future at $a$ (representing the subtree rooted at $b$) is provisionally inlined, but its continuation (representing the subtree rooted at $c$) is immediately stolen by an idle processor. Likewise, the futures at $b$ and $c$ are inlined, but their continuations are stolen by the two remaining idle processors. Now all processors are busy; subsequent futures are all provisionally inlined but no further stealing takes place and each processor winds up executing one of the circled subtrees of Figure 2.

This execution pattern depends on an *oldest-first* stealing policy: when an idle processor steals a task, the oldest available fork point is chosen. In this example the oldest fork point represents the largest available subtree and hence a task of maximal run-time granularity.

We now consider how these idealized execution patterns match up with real-life execution patterns for these methods.

# 3 Dynamic Methods Compared

Load-based inlining has an appealing simplicity and does in fact produce good results for some programs [17], but we have noted several factors which decrease its effectiveness. A major factor is that inlining decisions are irrevocable—once the decision to inline a task has been made there is no way to revoke the decision at a later time, even if it becomes clear at that time that doing so would be beneficial.

The following list summarizes the drawbacks of load-based inlining; the following sections discuss each in turn as a basis for comparing the two dynamic strategies.

1. The programmer must decide when to apply load-based inlining, and at what load threshold $T$.
2. Inlined tasks are not accessible; processors can starve even though many inlined tasks are pending.
3. Deadlock can result if inlining is used on some types of programs.
4. In an implementation with one task queue per processor, load-based inlining creates many more tasks than would be created with an optimal BUSD division.
5. Load-based inlining is ineffective in programs where fine-grained parallelism is expressed through iteration.

## 3.1 Programmer Involvement

Even though load-based inlining is an automatic mechanism it still requires programmer input. Some programs run significantly faster with eager task creation than they do with load-based inlining, so the programmer must identify where load-based inlining should be applied. For example, load balancing is crucial in a coarse-grained program creating relatively few tasks—inlining even a few large tasks can hurt load balancing by lengthening the "tail-off" period when processors are finishing their last tasks. With lazy task creation however, load balancing can't suffer because all inlining decisions are revocable. At worst, all lazily-inlined tasks will have their continuations stolen. But because the cost of stealing a task is comparable to that of creating an eager task[5], performance will not be significantly worse than with eager task creation. Thus lazy task creation can be used safely on such programs without the danger of degrading performance.

With load-based inlining, the programmer must also get involved by supplying a value for the load threshold $T$. Experience has shown that choosing the right value for $T$ is crucial for good performance, but is difficult to do except by experimentation [29]. Since lazy task creation requires no parameterization the programmer is freed of this burden as well.

## 3.2 Irrevocability

The irrevocability of load-based inlining can mean that processors become idle even though the continuations of many inlined tasks have not yet begun to execute. Such problems can be caused by ***bursty task creation*** and ***parent-child welding***. ***Bursty task creation*** refers to the fact that opportunities to create tasks may be distributed unevenly across a program. At the moment when a task is inlined, it may appear that there are plenty of other tasks available to execute, but by the time these tasks finish executing there may be too few opportunities to create more tasks. Consequently, processors may go idle because the continuations of the inlined tasks are not available for execution. This problem never arises with lazy task creation because these continuations are always available for stealing.

***Parent-child welding*** refers to the fact that inlining effectively "welds" together a parent and child task. If an inlined child becomes blocked waiting for a future to resolve (or for some other event), the parent is blocked as

[5] An exceptional case is discussed in Section 4.4.

well and is not available for execution. With lazy task creation, the information kept for each inlined child allows the child to be decoupled if it becomes blocked, allowing the parent to continue.

### 3.3 Deadlock

Perhaps the most serious problem with load-based inlining is that, for some programs, *irrevocable inlining is not a correct optimization*. Irrevocable inlining can lead to deadlock because it imposes a specific sequential evaluation order on tasks whose data dependencies might require a different evaluation order. A simple example appears in [17], where an inlined task waits for a semaphore which its "welded-on" parent will never be able to release. But deadlock is possible even without explicit inter-task synchronization, as shown by the prime-finding program of [21] and [20] (omitted here because of space considerations). If the wrong tasks are inlined a task testing the primality of a number could deadlock trying to access divisor primes which haven't yet been computed by its welded-on parents.

This type of deadlock is not possible with lazy task creation because of the decoupling of blocked tasks mentioned above. Any inlined task can be separated from its parent, so programs that are deadlock-free with eager task creation are also deadlock-free with lazy task creation.

Selective load-based inlining (as is possible in Qlisp) could be used by a sophisticated programmer to ensure that inlining is never performed where it might cause deadlock. However, this solution requires the programmer to accurately recognize all situations where the potential for deadlock exists, and still does not offer the other advantages of lazy task creation.

### 3.4 Too Many Tasks

The behavior of load-based inlining for programs like `psum-tree` has been analyzed by Weening [29, 30]. He assumes, as we do, that each processor maintains its own local task queue and that inlining decisions are based only on the local queue's length. He shows two ways in which the need to maintain at least one task on the local queue leads to non-BUSD execution. First, a lone processor $P$ executing a subtree of height $h$ creates $h$ tasks instead of just one; second, removing a task from $P$'s queue at an inopportune moment (a "transfer") can lead to the creation of $O(h^2)$ tasks. He derives an upper bound of $O(p^2h^4)$ tasks using $p$ processors, and points out that this bound guarantees asymptotically minimal task creation overhead as the problem size grows exponentially in $h$. In our experience, however (see Section 5.3), the overhead of task creation with load-based inlining is significant for problems of substantial size.

The bottom line is that load-based inlining with distributed task queues is unable to achieve oldest-first scheduling; many of the tasks created represent small subtrees. For example, consider what happens when a transfer removes a task from the queue of a processor $P$. The next time $P$ encounters a future call, $P$ will find that its queue is empty and so will create a new task to evaluate the call. But the position of $T$ in the program's call tree is really a matter of chance, determined only by the timing of the transfer operation. Since the majority of potential fork points lie toward the leaves of the tree, $T$ is likely to represent only a small subtree.

It is possible that using one central queue instead of several distributed queues would decrease the number of tasks, but the contention introduced by this alternative would probably be unacceptable and would certainly not be scalable. A much better alternative is the oldest-first scheduling policy of lazy task creation; as can be seen by the task counts in Section 5, lazy task creation results in many fewer tasks than load-based inlining. Tasks created by oldest-first scheduling are able to inline larger subtrees, giving a much better approximation to BUSD execution.

### 3.5 Fine-Grained Iteration

Not all parallel programs have bushy call trees; for example, some programs contain data-level parallelism expressed by iteration over a linear data structure. Unfortunately, neither load-based inlining nor lazy task creation is particularly effective in increasing the runtime granularity of such programs, so poor performance can result when tasks are fine-grained.

With both methods, granularity can only be increased when tasks are able to inline many other tasks. But because the "call tree" of a fine-grained iteration is long and spindly, granularity can be increased only by grouping together adjacent iterations. The simple task stealing methods used in both load-based inlining and lazy task creation are unable to perform this type of grouping (see [20] for further details), resulting in many small tasks.

We have considered several alternatives for handling such programs, involving more complex dynamic methods and/or compiler support. The best solution is not clear at this point, but we will present some ideas at the end of the paper.

## 4 Implementation

We have seen that lazy task creation has several strong advantages over load-based inlining. We now explore the implementation issues to determine whether the overhead of lazy task creation can be acceptably minimized.

Both of our dynamic methods increase efficiency by ignoring selected instances of **future**. But lazy task creation requires maintaining enough information when a **future** is provisionally inlined to allow another processor to steal the **future**'s continuation cleanly. The cost of maintaining this information is *the* critical factor in determining the finest source granularity that can be handled efficiently. The cost is incurred whether a new task is created or not, so a large overhead would overwhelm a fine-grained program. By comparison the cost of actually stealing a task is somewhat less critical; if enough inlining occurs the cost of stealing a task will be small compared to the total amount of work the task ultimately performs.

Still, the cost of stealing a continuation must be kept in the ballpark of the cost of creating an eager future. Stealing a continuation requires splitting an existing stack, which in a conventional stack-based implementation requires the copying of frames from one stack to another. Alternatively, we could use a linked-frame implementation where splitting a stack requires only pointer manipulations. However, care must be taken with such an implementation to ensure that the normal operations of pushing and popping a stack frame have comparable cost with conventional stack operations.

We have pursued both avenues of implementation: a conventional stack-based implementation for the Encore Multimax version of Mul-T as well as a linked-frame implementation for the ALEWIFE multiprocessor. The basic data structures and operations for lazy task creation are common to both implementations however, and are discussed next.

## 4.1 The Lazy Task Queue

Each task maintains a queue of stealable continuations called the *lazy task queue*, shown abstractly in Figure 4.1. When making a *lazy future call* corresponding to an instance of **future** in the source code, a task $T$ first pushes a pointer to the **future**'s continuation onto the lazy task queue. If upon return the continuation has not been stolen by another processor, $T$ dequeues it. We refer to $T$ as the *producer* of lazy tasks; another processor stealing them is called a *consumer*. Consumers remove frames from the head of the lazy task queue while the producer pushes and pops frames from the tail.

Figure 4.1 tells a lazy task creation story for a producer task $P$. 4.1a shows $P$'s stack (growing upward), which contains eight frames. Three of these frames represent continuations to lazy future calls; pointers to these frames have been placed on the lazy task queue. Note that the oldest continuation is at the head (bottom) of the queue while the newest continuation is at the tail (top) of the queue.

At this point a lazy future call occurs, corresponding to the code (**future** $X$), where $X$ denotes an expression to be evaluated. The continuation $K_t$ to this call represents all remaining computation, embodied in Figure 4.1b by the frame labelled $K_t$ and all those below it. As shown, a frame representing $K_t$ has been pushed onto the stack and a pointer to this frame has been added to the tail of the lazy task queue.

As a result of the lazy future call, $P$ begins evaluating $X$ in-line. 4.1c shows what happens if $P$ finishes evaluating $X$ before any stealing occurs—$P$ simply returns to $K_t$ after first popping the lazy task queue (removing the pointer to $K_t$'s top frame from the tail of the queue).

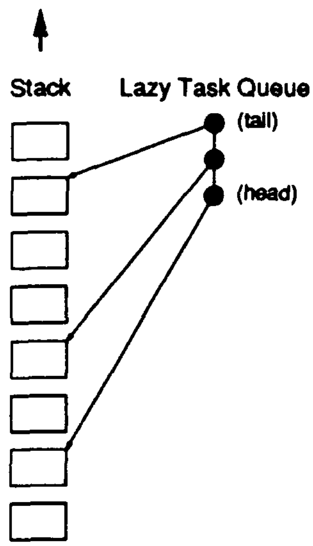
Now an idle consumer $C$ decides to steal a continuation from the head of $P$'s lazy task queue. This continuation $K_h$ was originally created by a lazy future call, say (**future** $Y$). When $P$ made this lazy future call it began evaluating $Y$ in-line, and has not finished doing so at the time of the steal. In order to steal $K_h$, $C$ must change $P$'s stack to appear as though an eager future had been created to compute $Y$. $C$ does this by creating a placeholder and modifying $P$'s stack so that the eventual value of $Y$ will resolve (*i.e.*, supply a value for) the placeholder rather than being passed directly to the continuation $K_h$. $C$ initializes its own stack to contain the frames of the continuation $K_h$ and then "returns" to $K_h$, passing the unresolved placeholder as a value.

Figure 4.1d shows the completed steal operation; it now looks as though an eager future had been created originally, with one processor (the producer $P$) evaluating the child $Y$ and another (the consumer $C$) evaluating the parent $K_h$. Note an important feature of the stealing operation: *the consumer never interrupts the producer.*

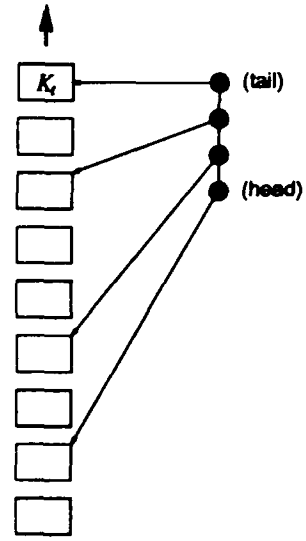
Implementations must take care to guard against two kinds of race conditions to ensure correctness of the stealing operation. First, two consumers may race to steal the same continuation; second, a producer trying to return to a continuation may race with a consumer trying to steal it.

## 4.2 Encore Implementation

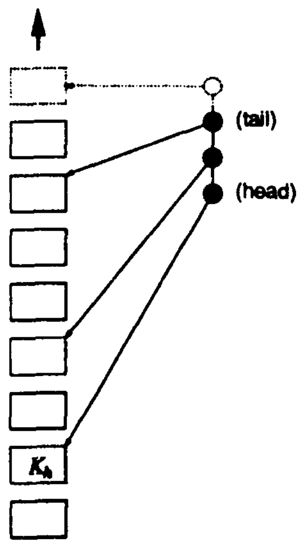
We have implemented lazy task creation in the version of Mul-T running on the Encore Multimax system, a bus-based shared-memory multiprocessor. Our Multimax system has 18 processors; the National Semiconductor 32332 processors used have relatively few



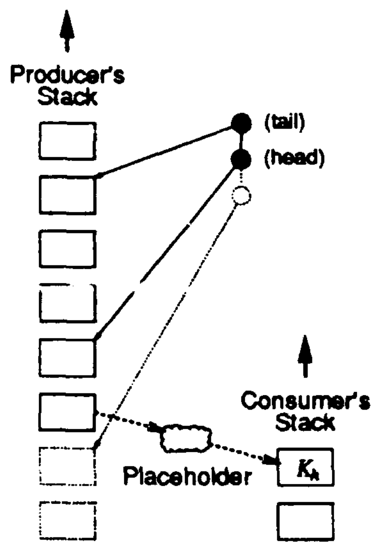

(a) Data structures for lazy task creation.




(b) A lazy future call causes a continuation to be queued.




(c) Returning from a lazy future call causes a continuation to be dequeued.




(d) A continuation is stolen.

Figure 3: Lazy task queue data structures and operations.

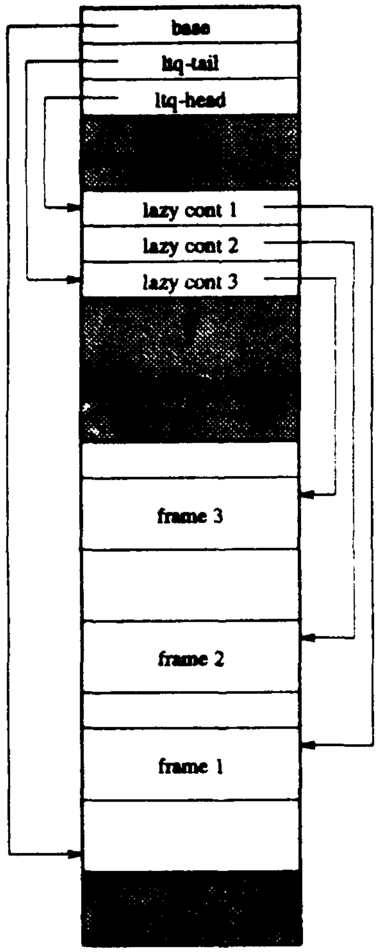



Figure 4: Lazy task queue implemented in conjunction with a conventional stack.

general-purpose registers (8) but fairly powerful memory addressing modes. Synchronization between processors is possible only by using a test-and-set instruction which acquires exclusive access to the bus.

In this implementation stacks are represented conventionally, in contiguous sections of the heap. As seen in Figure 4, the lazy task queue is kept in contiguous memory in the "top" part of a stack. As the producer pushes lazy continuations the queue grows downward while the stack frames grow upward. Stealing continuations effectively shrinks the stack by removing information from both ends (the head of the lazy task queue and the bottom frames of the stack). When a stack overflows (*i.e.*, when the gap between stack frames and lazy task queue gets too small), it may either be repacked to reclaim space created by steal operations or its contents may be copied to a new stack of twice the original size.

To steal from the stack pictured, a consumer first locates the oldest continuation by following the **`ltq-head`** pointer, through the **`lazy cont 1`** pointer, to **`frame 1`**. The consumer then replaces **`frame 1`** in the stack with a continuation directing the producer to resolve a placeholder. Next the consumer copies frames from **`frame 1`** down to the bottom of the live area of the stack (indicated by **`base`**) to a new stack, updating **`base`** and **`ltq-head`** appropriately.

To guard against the race conditions mentioned earlier there is a lock for the entire stack plus a lock for each continuation on the lazy task queue. Only the producer modifies **`ltq-tail`**, and only consumers modify **`ltq-head`** and **`base`**.

### 4.2.1 Lazy Future Call and Return

We now present the lazy task queue operations in somewhat more detail. Figure 5 gives assembler pseudo-code showing how the expression

```
(g (future (f x)))
```

would be compiled in Encore Mul-T with lazy task creation. The lazy future call and return in this example show the crucial lazy task queue operations of enqueuing and dequeuing a lazy continuation.

The first block (**`entry`** and **`call-g`**) shows the compiled code for the lazy future call to **`f`** and its continuation, containing the standard call to **`g`**. **`stack`** is a pointer to the current stack; lazy task queue pointers such as **`ltq-tail`** are referenced via an offset to this pointer.[6]

The code shows that 2 longwords (4 bytes each) are allocated in the lazy task queue area of the stack for each lazy continuation—one for the continuation itself and one for a lock. After storing the continuation pointer **`call-g`** and initializing the lock to 0 we increment the **`ltq-tail`** pointer, which makes the lazy continuation available for stealing. There is no need to test explicitly for overflow of the lazy task queue; the stack overflow check on entry simply tests the size of the empty region between the actual stack (growing upwards) and the lazy task queue (growing downwards).

Before calling **`f`** we push **`return-from-lf-call`** on the stack as the return address. This is a shared, out-of-line routine that serves as the continuation to all lazy future calls. It is shown in the second block of code. Here we see synchronization to guard against interference by a consumer trying to steal the same lazy continuation the producer is trying to return to. The returning producer first acquires the lazy task queue item lock (using the Encore's interlocked test and set instruction), busy-waiting if the lock is currently held by a consumer. Once the lock is acquired the return

[6] This is a slight simplification; in actuality, the current stack is stored in a block of data kept locally by each processor; **`ltq-tail`** is referenced using the double indirection capability of the NS 32332 processor.

```
(lambda (x)
  (g (future (f x))))

entry:
 Standard stack overflow test (3 instructions).
 push-addr  call-g                  # push return address (a.k.a. current continuation) on stack
 move       ltq-tail(stack),r1      # get pointer to tail of lazy task queue
 move       sp,8(r1)                # store pointer to stack continuation in lazy task queue
 move       $0,12(r1)               # initialize lazy task queue item lock
 add        $8,ltq-tail(stack)      # lazy continuation officially enqueued
 push-addr  return-from-lf-call     # call to f will return to return-from-lf-call
 Standard call to unknown procedure f (5 instructions).
call-g:
 Standard continuation code, including call to unknown procedure g (6 instructions).

return-from-lf-call:
 move       ltq-tail(stack),r1      # get pointer to lazy task queue tail
 test&set   4(r1)                   # try to lock tail item of lazy task queue
 br-if-clr  pop-ltq                 # if successful, go pop it
 Busy-wait loop to lock tail item of lazy task queue.
pop-ltq:
 sub        $8,ltq-tail(stack)      # lazy continuation officially dequeued
 adjust-sp  $-4                     # remove return-from-lf-call address from stack
 Standard return (2 instructions).
```

Figure 5: Assembler pseudo-code showing lazy future call and return in the Encore implementation.

address on top of the stack is guaranteed to be valid; in this case it will be either the original value **call-g** or else **resolve-placeholder** if the continuation has been stolen. After dequeuing the tail entry of the lazy task queue we return normally.

If, as is usually the case, the continuation to a lazy future call is known (*i.e.*, unless **future** appears in tail-call position), the code shown in Figure 5 can be streamlined by generating the **return-from-lf-call** code in line. This optimization, which saves 4 instructions (and increases the code size slightly), has not yet been implemented in the current system.

### 4.2.2 Steal Operation

Figure 6 gives the algorithm for stealing a lazy continuation from another processor's lazy task queue. The task to be stolen is chosen by a round-robin search of other processors' lazy task queues. Two locks must be acquired before a continuation is stolen—the producer's stack is locked to avoid races with other consumers and the continuation itself is locked to avoid a race with the producer trying to return to it.

Once a stealable continuation has been chosen and the necessary locks obtained, we replace it in the producer's stack with a continuation to resolve the newly created placeholder, and we update the producer's **base** and **ltq-head** pointers. At this point the producer's stack is in a consistent state, so we unlock the head item of the lazy task queue.[7] Then the bottom of the producer's stack is copied to the consumer's stack (taking care to use the *old* continuation rather than the newly swapped-in one!) and the consumer can begin executing the stolen continuation, passing the placeholder as an argument. The producer (or another processor if further stealing occurs!) will eventually return to our swapped-in continuation, providing a value for the placeholder.

### 4.2.3 Blocking

There is one remaining loose end in this discussion: what happens to the lazy task queue when a task $T$ blocks by touching an unresolved future? It is not sufficient to save the lazy task queue as part of $T$'s state because the queued lazy tasks would become inaccessible. We would then have the same potential deadlock problem that arises with load-based inlining.

[7] The producer's stack is not unlocked at this point because of the possibility of stack overflow—the repacking operation discussed earlier would conflict mightily with a stealer's copying operation.

- Allocate and initialize data structures: a placeholder $P$, a **new task** object $T_2$, and a new stack $S_2$.
- Look for a continuation to steal.
  - Poll other processors to find one whose current stack $S_1$ has a non-empty lazy task queue (*i.e.* **ltq-tail** ≥ **ltq-head**).
  - Try to lock **stack** $S_1$; if it's already locked, skip to next processor.
  - Try to lock head item of $S_1$'s lazy task queue $Q$; if it's already locked, skip to next processor.
- Steal the continuation. In the head item (now locked) of $Q$ is a pointer $CP$ into the stack $S_2$. $CP$ points to a stack frame $C$ representing a stealable continuation. The bottom of the stack (the portion between $CP$ and $S_1$'s **base** pointer) must be copied to the new stack $S_2$.
  - Replace $C$ in $S_1$ with the continuation (**resolve-placeholder** $P$).
  - Update **base** and **ltq-head** pointers in $S_1$.
  - $S_1$ is now in a consistent state; unlock head item of $Q$.
  - Copy bottom portion of $S_1$ into $S_2$.
  - Unlock stack $S_1$.
  - "Return" to top continuation in new stack $S_2$, passing placeholder $P$ as the argument.

Figure 6: Algorithm for steal operation in Encore implementation.

The simple solution adopted here is for $T$ to "bite its tail." $T$'s stack is split above the most recent lazy continuation (at the tail of the lazy task queue), and only the top piece is blocked along with $T$. As with a steal operation, a placeholder is created to communicate a value between the two pieces of the split stack. The executing processor $P$ can continue using the bottom piece of the stack, which contains all of the continuations on the lazy task queue. No queued continuations are inaccessible to potential consumers. $P$ dequeues the tail lazy continuation and returns to it, passing the placeholder as an argument.

In essence, $P$ has stolen a task from the *tail* of $T$'s lazy task queue. One problem with this solution is that it goes against our preference for oldest-first scheduling, since we have effectively created a task at the newest potential fork point. Performance can suffer because this task is more likely to have small granularity. And further blocking may result, possibly leading to the dismantling of the entire lazy task queue. An improved solution which avoids these problems has been implemented for ALEWIFE, and is discussed in the next section.

## 4.3 ALEWIFE implementation

The Encore implementation of lazy task creation performs reasonably well by lowering the overhead of using the **future** construct, but it still has several other sources of overhead. Compiler support for future and stack checking is costly (see Section 5.1), and locking operations can be costly because a global resource (the bus) is used.

The ALEWIFE machine [1]—a cache-coherent machine being developed at MIT with distributed, globally shared memory—is designed to address these problems. Its processing elements are modified SPARC[8] chips [2]; the modifications of interest here are fast traps for strict operations on futures and support for full/empty bits in each memory word. If a strict arithmetic operation or memory reference operates on a future a trap occurs; thus, explicit checks are not needed. The full/empty bits allow fine-grained locking: ALEWIFE includes memory-referencing instructions that trap when the full/empty state of the referenced location is not as expected. It should be noted that this modified SPARC is not "special purpose" hardware for Mul-T programs. The modifications do not affect the cycle time of the processor and would be useful for the implementation of lazy task creation in the context of any language.

For the ALEWIFE implementation of lazy task creation, a stack is represented as a doubly linked list of stack frames (inspired by [22]) in order to minimize copying in the stealing operation. In this scheme, each frame has a link to the previously allocated frame and another link to the next frame to be allocated. Thus push-frame and pop-frame operations are simply load instructions. An important feature of this scheme is that stack frames are not deallocated when popped. A subsequent push will re-use the frame, meaning that in the average case the cost of stack operations associated with procedure call and return is very close to the cost of such operations with conventional stacks. The "next frame" link is set to empty when no next frame has been allocated. This strategy avoids the need to check explicitly for stack overflow when doing a push-frame operation: in the (uncommon) case where no next frame is available the push-frame operation will trap and the trap handler will allocate a new frame.

[8] SPARC is a trademark of Sun Microsystems, Inc.

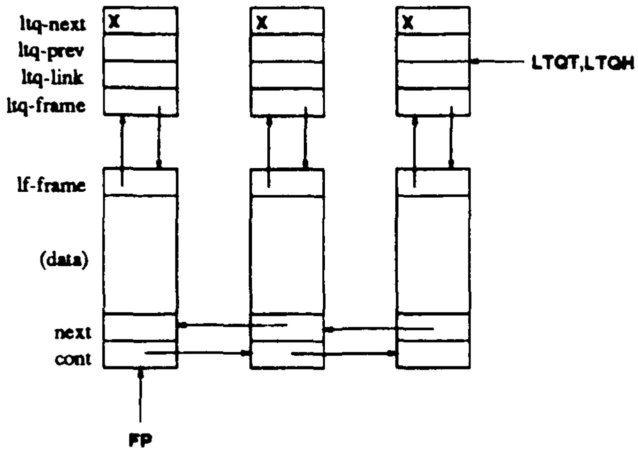



Figure 7: Just before lazy future call.

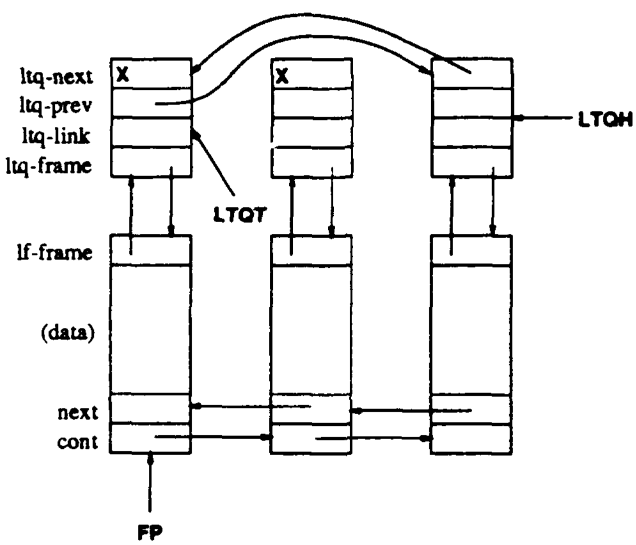



Figure 8: Just after lazy future call.

An earlier version of this paper [20] described an initial ALEWIFE implementation. In that version, stealing a lazy task involved copying the topmost stack frame. The version described here avoids this copying and also fixes a subtle bug in the original version.

Each frame is divided into two separate data structures, referred to as the *stack frame* and the *frame stub*. The stack frames form a doubly linked list as described at the beginning of this section. Each stack frame contains local and temporary variables as in an ordinary stack frame; in addition, each stack frame contains a pointer to its associated frame stub. Each frame stub also has a pointer back to its associated stack frame. Separating these two structures is important in allowing a non-copying steal operation.

In this implementation the lazy task queue is threaded through the frame stubs. Figures 7–10 show the lazy future call and stealing operations graphically.

In these figures we use the following register names:

`FP` Frame pointer register. Points to the current stack frame (not frame stub).

`LTQT` Lazy task queue tail register. Modified only by the producer. Points to the current frame stub.

`LTQH` Head of the lazy task queue. This must be in memory so that consumers on other processors can steal frames from the head of the queue. Its full/empty bit serves as the lock limiting access to one potential consumer at a time.

Each stack frame has the following slots:

`next` This slot points to the "next" frame, which will become current if a stack-frame push operation is performed. The push-frame operation is thus performed simply by loading `next[FP]` into `FP`. If the next frame has not yet been allocated, `next[FP]` is marked as empty.

`cont` This slot points to the "continuation" frame, which will become current if a stack-frame pop operation is performed. The pop-frame operation is thus a load of `cont[FP]` into `FP`.

`data` Some number of slots for local variable bindings and temporary results.

`lf-frame` This slot points to the associated frame stub.

Each frame stub has the following slots:

`ltq-next` This slot points to the next frame stub on the lazy task queue (toward the tail of the queue). This location's full/empty bit is the lock arbitrating between a consumer stealing a continuation and the producer trying to invoke that continuation.

`ltq-prev` This slot points to the previous frame stub on the lazy task queue (toward the head of the queue).

`ltq-link` The lazy future call code stores in this slot the return address that the consumer should use if it steals this frame's continuation. If the continuation is stolen, the consumer reads out this return address and replaces it with the placeholder object it creates.

`frame` This slot points to the associated stack frame.

In this implementation, every call—whether a lazy future call or an ordinary procedure call—is preceded by a push-frame operation and followed by a pop-frame operation; this contrasts with the more common approach of pushing a frame upon procedure entry and deallocating it at procedure exit. (The details motivating this choice and a discussion of its cost may be found in [21].)

Figure 7 shows how the stack frames and relevant registers might look just before a lazy future call (but after that call's push-frame operation has already occurred). Note that each stack frame's **next** pointer points to the next frame toward the top of the stack and each **cont** pointer points to the next stack frame toward the bottom of the stack. If a memory location's contents are left blank in the figure, its contents are either unimportant (they will never be used) or indeterminate: for example, the **next** slot of the leftmost frame in Figure 7 could either be empty or point to another, currently unused frame. An "X" in the left-hand part of a frame slot (see, for example, the **ltq-next** slots in Figure 7) indicates that the full/empty bit of the corresponding memory word is set to "empty."

The lazy task queue in Figure 7 has no frames in it. A consumer would discover this by seeing that the **ltq-next** slot of the frame stub pointed to by **LTQH** is empty—if this task had stealable frames, this slot would point to the first such frame.

Figure 8 shows the situation just after the lazy future call. The frame stub associated with the current stack frame (pointed to by **FP**) has joined the lazy task queue. Accordingly, **LTQT** has changed to point to that frame stub, and the **ltq-next** and **ltq-prev** links have been updated as needed to maintain the doubly linked lazy task queue. Note that the rightmost frame stub in Figure 8 is not logically part of the lazy task queue—it is serving as a convenient header object for the doubly linked queue. The middle frame is also not part of the lazy task queue; it is simply part of the stack. The current frame stub's **ltq-link** field contains the address for the lazy future call's continuation, as required.

If no consumer steals this continuation, then this lazy future call will eventually return. The code for the return will restore the state of affairs depicted in Figure 7, after which the pop-frame operation associated with the lazy future call can be performed.

Figure 9 shows the state of the producer and consumer tasks if instead a consumer steals the continuation from the task shown in Figure 8. The consumer task's state variables are shown with a c appended, as in **LTQHc**. The shaded areas and shaded arrows show structures that have been created by the consumer. An alternate view of this situation is shown in Figure 10. Note that the consumer's stack (the part that is not blacked out in Figure 10) now looks just like the producer's stack did in Figure 7 just before the original lazy future call (and just like the producer's stack would have looked in the case of a normal return from the lazy future call). Effectively, the consumer has "taken over" the continuation, created a placeholder to stand for the value of the called computation (which is still being performed by the producer), and forced an early return from the lazy future call, supplying the placeholder as the call's returned value. (No arrow is shown from any of the consumer's data structures to the placeholder because that value is returned in one of the consumer's registers.)

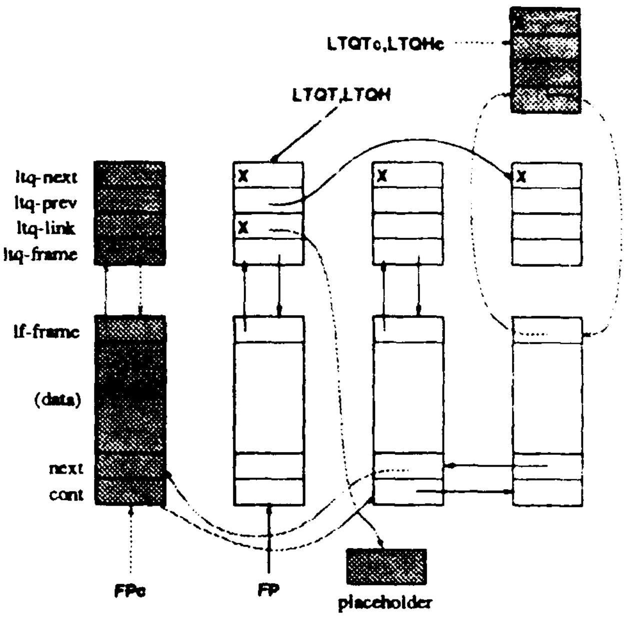



Figure 9: After steal; new structures shaded.

The consumer has also made the producer's **ltq-link** field point to the newly created placeholder. When the producer completes its computation and finds that its continuation has been stolen, it looks here to find the placeholder that should resolve to this computation's value. The synchronization here is unusual in that **ltq-link** is marked "empty" even though it contains useful data. This technique handles close races between a returning producer and a stealing consumer. By inspecting **ltq-next** and **ltq-prev** pointers, a returning producer can discover that its continuation has been stolen before the consumer has actually stored the placeholder in the **ltq-link** field. Correct operation is ensured by having the consumer set the **ltq-link** field's "empty" flag when the placeholder is installed, and having the producer wait for this "empty" flag before attempting to read out the placeholder.

A producer returning from a lazy future call distinguishes between the situations shown in Figures 8 and 9 by locating the frame stub $F$ pointed to by the **ltq-prev** field of the frame stub pointed to by **LTQT**

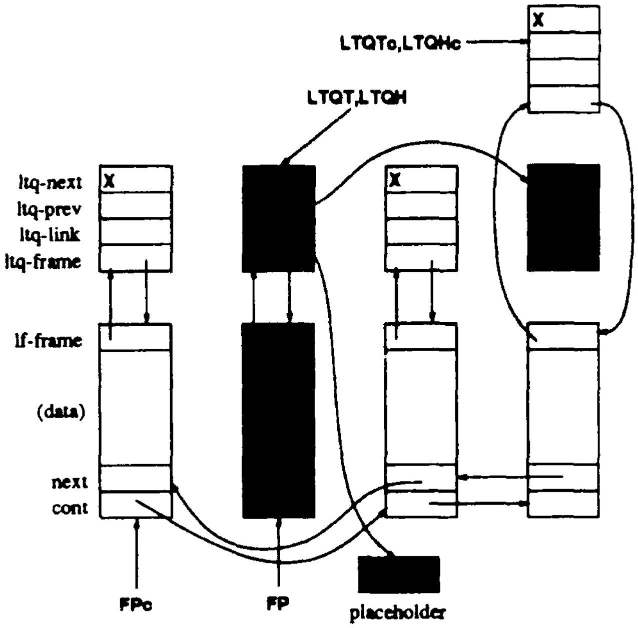



Figure 10: After steal; frames belonging to producer shown in black.

and looking at the `ltq-next` field of $F$. In Figure 9, where the continuation has been stolen, this field in $F$ is empty; in Figure 8, where the continuation has not been stolen, it is not.

The algorithm for lazy future calls is spelled out in more detail in the pseudo-code shown in Figure 11. The in-line code for a lazy future call starts at the label `lf-call`; the code at `stolen` is out-of-line code shared by all lazy future calls. The algorithm for a consumer to find and steal a continuation is given in Figure 12.

Since the producer is not explicitly notified when a steal operation is performed on its stack, any resources the producer may continue to use after a continuation is stolen may not be used by the consumer. Some complexity in the algorithm for stealing results from this fact. In particular, the consumer must copy the rightmost frame stub in Figure 9 so it can use the `ltq-next` slot in that frame stub when it performs lazy future calls. If this frame stub were shared with the producer, such calls by the consumer could confuse the producer.

This approach has the drawback that a push-frame operation occurs at every procedure call (lazy or not) instead of at the entry point of a procedure, but there are several mitigating factors:

1. Push-frame and pop-frame operations are inexpensive (one instruction).
2. Compiler optimizations can eliminate some of them (*e.g.*, pop-push sequences that cancel out can be detected and eliminated).

1. Select a processor for inspection and load its `LTQH` pointer into a register $H$, leaving `LTQH` empty. If `LTQH` is found already empty, move on to another processor. (This enforces mutual exclusion among consumers.)
2. Load `ltq-next[`$H$`]` into a register $F$, leaving `ltq-next[`$H$`]` empty. If `ltq-next[`$H$`]` is found already empty, then this processor's lazy task queue is empty—write $H$ back into `LTQH` and move on to another processor.
3. Store $F$ into `LTQH`. This step commits the steal operation and ends the exclusion of other consumers. Other consumers can now steal other continuations from this processor, even as this consumer continues its steal operation.
4. Load `ltq-link[`$F$`]` into a register *retpc*. This is the consumer's return address from the lazy future call.
5. Create a placeholder object and save its address in a register *retval*.
6. Store *retval* into `ltq-link[`$F$`]`, leaving `ltq-link[`$F$`]` empty. If the producer is already trying to return to the continuation being stolen, this step frees the producer to proceed and deposit its result into the placeholder.
7. Create the stack frame and the two frame stubs shown shaded in Figure 9 and link them together as shown by the shaded arrows in Figure 9. These data structures are accessed only by this consumer and hence neither the producer nor other consumers need to synchronize with these operations.
8. Set the `FPc`, `LTQTc`, and `LTQHc` pointers properly and execute a return to the address *retpc*, giving *retval* as the returned value.

Figure 12: Algorithm for steal operation in ALEWIFE implementation.

3. Some push-frame operations at procedure entry turn out to be unnecessary due to conditional branches; this approach delays them until they are sure to be necessary.

We do not know the net effect of using this approach but we believe that the difference is not significant.

Finally, we return to the issue of what to do with the

```
lf-call:
  load      next[FP],FP                   # Push stack frame.
  load      lf-frame[FP],temp             # Address of new frame stub.
  store     $continue,ltq-link[temp]      # PC for consumer's return.
  store     LTQT,ltq-prev[temp]           # Make lazy task queue backward ...
  store     temp,ltq-next[LTQT]           # ... and forward links.
  move      temp,LTQT                     # Advance lazy task queue tail pointer.
  Call the procedure.
  load      ltq-prev[LTQT],temp           # Dequeue from lazy task queue tail,
  empty     ltq-next[temp]                # trap to stolen if continuation stolen.
  move      temp,LTQT                     # Reset lazy task queue tail pointer.
continue:
  load      cont[FP],FP                   # Pop stack frame.

stolen:
  Wait for ltq-link[LTQT] to be empty.
  load-e    ltq-link[LTQT],temp           # Get placeholder to resolve.
  Resolve the placeholder in temp to the value returned by the procedure.
  Terminate the current task and find new work to do.
```

Figure 11: Assembler pseudo-code showing lazy future call and return in the ALEWIFE implementation.

lazy task queue when a task blocks on an unresolved future. To preserve both oldest-first scheduling and laziness in task creation we would like to make the lazy task queue accessible for normal stealing by consumers. This is accomplished by placing the entire blocked task, lazy task queue and all, on the task queue of an appropriate processor.[9] Consumers may steal either from a task that is actually running or from a queued blocked task; a processor may steal from the lazy task queue of one of its own blocked tasks if it runs out of other useful work. This solution addresses the problems raised in Section 4.2.3.

## 4.4 Discussion

What are the advantages and disadvantages of these implementations? The main disadvantage of the conventional stack implementation is the copying it performs. It would appear that the amount of copying required for a stealing operation is potentially unlimited, so that the cost of stealing a lazy task is also unlimited. While this is technically true it is somewhat misleading; the overhead of copying when stealing a continuation should be viewed against the cost of creating the continuation in the first place. A program with fine source granularity does little work between lazy future calls, and so is not able to push enough items onto the stack to require significant copying. A program which creates large continuations (requiring stealers to do lots of copying) must do a fair amount of work to push all that information on the stack, and the cost of copying is unlikely to be significant in comparison.

[9] Of course, the task is marked as blocked, so the processor will not attempt to run it.

One exception to this argument is a program that builds up a lot of stack and then enters a loop that generates futures:

```
(define (example)
  (build-up-stack-and-then-call loop))
(define (loop)
  (future ... )
  (loop))
```

Stealing the first lazy task's continuation requires copying the built-up stack. As argued, that cost is unlikely to be significant compared with the cost of building up the stack in the first place. But in this example the stolen continuation immediately creates another lazy task, so the next steal must copy the same information again. In fact, spreading work to $n$ processors in this example via lazy tasks requires the built-up stack information to be copied $n$ times.

There are two easy solutions to this problem. First, `loop` can be rewritten so that `future` appears around the recursive call to `loop`, resulting in a program where the built-up stack is never copied. Or, `future` could be inserted around the original call to `loop`, resulting in a program where the built-up stack is copied only once.

It appears then that the effects of copying in a conventional stack implementation can be minimized. But it is still attractive to eliminate copying altogether using the linked-frame implementation described for

ALEWIFE. Such an implementation is certainly more efficient on lazy task operations. It is somewhat more difficult to gauge exactly the overhead introduced in sequential sections of code. One ramification of re-using stack frames is that all frames have a fixed size; choosing the correct frame size involves a trade-off. If a small frame size is chosen, frames needing more space will need to create an overflow vector, increasing costs for accessing frame elements and for memory allocation. If a large frame size is chosen, most frames will contain a lot of unused slots. This could lead to more frequent garbage collection and might use up valuable space in cache and/or virtual memory, although these latter factors could well be minimal in today's memory-rich systems. The current ALEWIFE implementation uses a frame size of 17 slots. We must accumulate more experience with this promising implementation technique before making a final evaluation.

# 5 Performance

In this section we present performance figures for both Mul-T implementations. Measurements of Encore Mul-T used Yale's Encore Multimax system, configured with 18 NS-32332 processors and 64 megabytes of memory.

Figures for ALEWIFE Mul-T were obtained using a detailed simulator of the ALEWIFE machine. Both the Mul-T run-time system and code for the benchmarks are compiled to SPARC instructions that are interpreted by the simulator. Overheads due to future creation, blocking, scheduling, *etc.*, are accurately reflected in the statistics. Memory-referencing delays were not simulated in these experiments.[10]

When assessing the performance of a multiprocessor system it is important to make comparisons with the "best" sequential implementation. To compare a parallel Mul-T program with, say, a sequential C program, four categories of overhead must be considered:

1. The cost of using Lisp instead of a language like C, *e.g.* automatic storage reclamation, manipulation of run-time tags, dynamic linking.
2. The cost of using sequential Mul-T instead of T, *e.g.* run-time checks for futures and stack overflow.
3. The cost of using a parallel algorithm instead of a sequential algorithm, *e.g.* using recursive divide and conquer instead of an iterative loop.
4. The run-time costs of multiprocessing, *e.g.* task creation, idle processors, contention for shared resources.

To ensure that measurements of our task creation strategies are meaningful we must distinguish overhead due to task creation from overhead due to these other sources—it is important to be sure that the overhead of task creation really *is* low, rather than just *looking* low because it is masked by overhead in the rest of the system.

The first two categories of overhead are addressed in Section 5.1 while the last two are considered in the context of specific benchmarks in Section 5.3. Section 5.2 deals specifically with the overhead of lazy task creation.

## 5.1 Overhead in Sequential Code

Despite its reputation for inefficiency, overhead due to Lisp is not a major factor in the benchmarks to be presented. First, we note that code produced by T's Orbit compiler is comparable in quality to code produced by other compilers for the same hardware [19]. Second, we have minimized run-time overhead in our benchmarks by using type-specific arithmetic, avoiding run-time storage allocation, and excluding garbage-collection time from performance statistics. The programs were carefully written for maximum efficiency. As a direct comparison, the "best" version of `tridiag` (see Section 5.3) was coded in C (3.33 sec) as well as T (3.92 sec).

The second category of overhead is significant for Encore Mul-T but insignificant for ALEWIFE Mul-T. Overhead is introduced in sequential Mul-T code by the Encore implementation because compiler support is provided for futures and multiple stacks. The compiler inserts `future?` checks on arguments to strict operations, and inserts tests for stack overflow. Although the Encore implementation is engineered to minimize these sources of overhead [17], the cost can be non-trivial. (We note however that compiler support for future checking is orthogonal to support for lazy task creation—lazy task creation also performs well in the Encore implementation when the overhead of future checking is eliminated by using explicit `touch` operations instead of implicit compiler checks.)

Table 1 compares running times of several sequential programs[11] in T3.1 with the same programs run in Mul-T on one processor. Because of future and stack checking overhead, the Mul-T programs run between 1.4 and 2.2 times as long as their T3.1 counterparts.

[10] Minimizing memory-referencing delays is crucial to good performance in a distributed-memory machine. ALEWIFE's distributed caching scheme [5] reduces the need for remote references; preliminary results of current research at MIT show that ALEWIFE Mul-T also performs well when network delays are simulated.

[11] Some of these programs are described in Section 5.3; the rest are described in [17].

| Program | Time (seconds) Mul-T | T | ratio |
|---|---|---|---|
| **abisort** | 12.67 | 6.98 | 1.62 |
| **compiler** | 159 | 98 | 1.62 |
| **fib** | 0.24 | 0.12 | 2.00 |
| **mergesort** | 1.82 | 0.99 | 1.84 |
| **permute** | 11,600 | 8,500 | 1.36 |
| **queens** | 3.95 | 2.44 | 1.62 |
| **speech** | 95.9 | 43.4 | 2.21 |
| **tridiag** (best) | 6.01 | 3.92 | 1.53 |

Table 1: Comparison of Running Times for Encore Mul-T and T3.1.

The ALEWIFE implementation of Mul-T does not incur these overheads, as hardware traps eliminate the need for explicit checks.[12] The analog of Table 1 for ALEWIFE would show identical "parallel" and "sequential" times for these programs.

All measurements of Encore Mul-T in Section 5.3 include the overhead of future and stack checking; this means that the relative granularity of tasks is somewhat larger for Encore than for ALEWIFE.

### 5.2 Cost of Lazy Task Queue Operations

As mentioned earlier, it is crucial to minimize the overhead of lazy future calls. Below are statistics for both implementations on the additional cost of a lazy future call over that of a conventional call, namely pushing a continuation onto the lazy task queue and popping it off.

| | |
|---|---|
| Encore | 12 instructions, 12.6 $\mu$sec |
| ALEWIFE | 9 instructions, 15 SPARC cycles |

For the Encore, 4 instructions could be eliminated by using the compiler optimization mentioned in section 5, saving roughly 3 $\mu$sec. Still, the ALEWIFE sequence is probably the cheaper of the two, since the RISC instructions of the SPARC processor are simpler than NS-32332 instructions.

The cost of stealing a continuation from another processor's task queue is not as critical, since steals are relatively rare. Still, as seen below, stealing a task in the Encore implementation has comparable cost to creating an eager future. Stealing a task in the ALEWIFE implementation is noticeably cheaper; the linked-frame stack implementation allows a much cleaner steal.

[12] It is interesting to note that the presence of hardware tag checking may be more significant in machines supporting parallel Lisp than in machines supporting sequential Lisp.

| Machine / Operation | Number of Instructions |
|---|---|
| Encore / Eager Future | 118 |
| Encore / Steal | 150 + 4 per word copied |
| ALEWIFE / Steal | 100 |

These instruction counts include all aspects of creating and executing a task; *e.g.*, allocating and initializing placeholder, task, and stack objects, queueing and dequeuing the task, and resolving the placeholder.

### 5.3 Benchmarks

We begin our discussion of actual Mul-T programs with the synthetic benchmark **grain**, designed to measure the effectiveness of the various task-creation strategies over a range of task granularities. **grain** adds up a perfect binary tree of 1's using a parallel divide and conquer structure very similar to **psum-tree**, but before returning 1 at any leaf it executes a delay loop of a specified length, allowing granularity control. By timing trials using a range of granularities we can get an "efficiency profile" for each task-creation strategy. The efficiency $E$ for a given trial is calculated using the formula

$$E = \frac{t_{seq}}{n t_{par}}$$

where in this case the sequential time $t_{seq}$ is for a Mul-T program without futures and the parallel time $t_{par}$ was measured using $n = 16$ processors. Efficiency of 1.0 means perfect speedup. The tree depth of 16 (65,536 1's) used in these trials ensures that processor idle time at start-up and tail-off is minimal, so close-to-perfect speedup should be achievable.

The granularity figures across the top of Table 2 tell how many NS-32332 instructions were used at the leaves to execute the delay loop and return 1; they do not include the instructions which implement the basic divide and conquer loop. The average source granularity is actually half of the given figure because internal nodes of the tree (where no delay loop is executed) account for half of the futures in this program. The instruction counts would be different for ALEWIFE due to its RISC instruction set, but because the source code is the same the efficiency figures are roughly comparable.

As expected, the high cost of eager task creation leads to poor efficiency at fine granularities. With load-based inlining 90–95% of the $2^{16}$ tasks are eliminated, improving efficiency substantially. Lazy task creation makes an additional improvement by eliminating more than 99% of the tasks. Still, the overhead of lazy future calls is significant, hurting efficiency at the finest

| Machine / Strategy | Leaf task granularity (number of NS-32332 instructions) 6 | 12 | 24 | 48 | 96 | 192 | 384 | 768 | 1536 | 3072 |
|---|---|---|---|---|---|---|---|---|---|---|
| Encore / Eager | .06 | .07 | .09 | .12 | .17 | .27 | .44 | .62 | .77 | .87 |
| Encore / LBI ($T = 2$) | .51 | .52 | .54 | .63 | .69 | .80 | .88 | 93 | .96 | .98 |
| Encore / Lazy | .56 | .59 | .62 | .71 | .81 | .84 | .92 | .95 | .97 | .99 |
| ALEWIFE / Lazy | .74 | .78 | .82 | .86 | .91 | .95 | .97 | .98 | .99 | 1.00 |

Table 2: Efficiency of the **grain** benchmark on 16 processors.

granularities. The lower overhead of lazy future calls in ALEWIFE leads to yet higher efficiency.

Table 3 shows performance statistics for several Mul-T programs. For each task creation strategy, the column marked $t$ shows the elapsed time (in seconds for Encore and in millions of simulated SPARC cycles for ALEWIFE) as well as the relative speedup in parentheses. The column marked $f$ shows the number of tasks (futures) created. Statistics are given for 1, 2, 4, 8, and 16 processors; in addition, the row marked "seq" gives the Mul-T time on one processor when **future** is ignored, and the row marked "best" gives the Mul-T time for running the best sequential version of the benchmark.

In our experience, Encore timings vary somewhat between trials even when each process acting as a virtual Mul-T processor is given exclusive control of an actual Multimax processor. It appears that changes in program and data locations from trial to trial substantially affect the miss ratio in the Multimax's direct-mapped, physically-addressed cache. Each figure shown here is the average of several trials; code was reloaded between each trial.

Knowing the source granularity of a benchmark (see Section 1) is important in interpreting the performance results. To get a measure of source granularity we can divide the sequential execution time of a benchmark by its total number of calls to **future**:

$$g = \frac{t_{seq}}{f_{ETC}}$$

$g$ estimates average task execution time, excluding task creation overhead. For these benchmarks our Encore runs at about 1 Mips, so $g$ is roughly comparable to the average number of NS32332 instructions per task as well.

**queens** ($g = 113$) finds all solutions to the $n$ queens problem, with $n = 10$ in this case.[13] A queen is placed on one row of the board at a time; each time a queen is legally placed, **future** appears around a recursive call to find all solutions stemming from the current configuration.

**abisort** ($g = 119$) performs an adaptive bitonic sort [4] of $n = 16,384$ numbers. The "adaptive" algorithm has complexity $O(n \log n)$ rather than the $O(n \log^2 n)$ of the standard bitonic sort algorithm. For comparison, the "best sequential time" shown in Table 3 is for an optimized merge sort. Adaptive bitonic sort performs about twice as many comparisons as merge sort, and has somewhat greater bookkeeping costs. However, its merge operation has substantial parallelism which allows close to linear speedup; such speedup is not possible with straightforward implementations (on hardware like ours) of other divide-and-conquer sorts such as merge sort or quicksort.

**tridiag** ($g = 314$) solves a tridiagonal system of $n = 65,535$ equations using cyclic reduction [15] and backsubstitution. "best" measures the standard Gaussian elimination algorithm, which performs fewer operations per equation than cyclic reduction (8 as opposed to 17) but is inherently sequential. The "seq" time reflects this difference, as well as some overhead due to the use of recursion in cyclic reduction. The large value of $n$ simply shows our preference for non-trivial problems; good performance was also achieved for smaller values of $n$.

The performance figures show fairly consistent results for these first three finer-grained benchmarks. Comparing the "seq" and 1-processor rows for these programs gives an indication of the overhead of task creation for each strategy; in **queens** for example, creating tasks eagerly nearly triples the running time. Load-based inlining greatly reduces this impact (to only 3%) because there is very little overhead when no task is created. Lazy task creation has somewhat higher overhead, though not overwhelmingly so (11%).

Load-based inlining improves running times substantially over the eager task creation times, but it consistently suffers significant task-creation overhead due to the mechanism discussed in Section 3.4. For these programs, LBI eliminates only 80-87% of the possible tasks when 16 processors are used. Lazy task creation performs much better, eliminating 98-99% of the possible tasks. Despite its higher overhead, lazy task creation

[13] This version of **queens** is different from the ones measured in [20] and [21].

**queens**

| n | Encore Eager t | Encore Eager f | Encore LBI ($T = 2$) t | Encore LBI ($T = 2$) f | Encore Lazy t | Encore Lazy f | ALEWIFE Lazy t | ALEWIFE Lazy f |
|---|---|---|---|---|---|---|---|---|
| seq | 3.95 | | 3.95 | | 4.28 | | 6.47 | |
| 1 | 11.69 (1.00) | 34814 | 4.06 (1.00) | 30 | 4.75 (1.00) | 0 | 7.20 (1.00) | 1 |
| 2 | 5.91 (1.98) | 34814 | 2.09 (1.94) | 528 | 2.38 (2.00) | 11 | 3.60 (2.00) | 11 |
| 4 | 3.05 (3.83) | 34814 | 1.15 (3.53) | 2135 | 1.20 (3.96) | 45 | 1.84 (3.91) | 104 |
| 8 | 1.63 (7.17) | 34814 | .67 (6.06) | 4771 | .61 (7.79) | 98 | .94 (7.66) | 229 |
| 16 | 1.08 (10.82) | 34814 | .43 (9.44) | 6826 | .34 (13.97) | 377 | .50 (14.40) | 554 |

**abisort**

| n | Encore Eager t | Encore Eager f | Encore LBI ($T = 2$) t | Encore LBI ($T = 2$) f | Encore Lazy t | Encore Lazy f | ALEWIFE Lazy t | ALEWIFE Lazy f |
|---|---|---|---|---|---|---|---|---|
| best | 3.67 | | 3.67 | | 3.56 | | 6.03 | |
| seq | 12.67 | | 12.67 | | 12.63 | | 20.27 | |
| 1 | 37.99 (1.00) | 106K | 15.29 (1.00) | 910 | 14.15 (1.00) | 0 | 23.64 (1.00) | 1 |
| 2 | 19.29 (1.97) | 106K | 7.84 (1.95) | 2449 | 7.13 (1.98) | 8 | 11.82 (2.00) | 13 |
| 4 | 9.85 (3.86) | 106K | 4.19 (3.65) | 6309 | 3.61 (3.92) | 55 | 5.93 (3.99) | 55 |
| 8 | 5.17 (7.35) | 106K | 2.31 (6.62) | 12627 | 1.84 (7.69) | 263 | 3.02 (7.83) | 388 |
| 16 | 3.36 (11.31) | 106K | 1.37 (11.16) | 19248 | 1.00 (14.15) | 894 | 1.57 (15.06) | 1018 |

**tridiag**

| n | Encore Eager t | Encore Eager f | Encore LBI ($T = 2$) t | Encore LBI ($T = 2$) f | Encore Lazy t | Encore Lazy f | ALEWIFE Lazy t | ALEWIFE Lazy f |
|---|---|---|---|---|---|---|---|---|
| best | 6.01 | | 6.01 | | 6.02 | | 4.78 | |
| seq | 15.43 | | 15.43 | | 15.40 | | 14.20 | |
| 1 | 26.58 (1.00) | 49150 | 16.04 (1.00) | 225 | 17.18 (1.00) | 0 | 15.35 (1.00) | 1 |
| 2 | 13.56 (1.96) | 49150 | 8.10 (1.98) | 792 | 8.61 (2.00) | 5 | 7.68 (2.00) | 4 |
| 4 | 6.87 (3.87) | 49150 | 4.18 (3.84) | 2524 | 4.29 (4.00) | 13 | 3.84 (4.00) | 9 |
| 8 | 3.54 (7.51) | 49150 | 2.17 (7.39) | 4157 | 2.18 (7.88) | 133 | 1.97 (7.79) | 190 |
| 16 | 2.01 (13.22) | 49150 | 1.16 (13.83) | 6637 | 1.14 (15.07) | 550 | 1.02 (15.05) | 461 |

**speech**

| n | Encore Eager t | Encore Eager f | Encore LBI ($T = 2$) t | Encore LBI ($T = 2$) f | Encore Lazy t | Encore Lazy f | ALEWIFE Lazy t | ALEWIFE Lazy f |
|---|---|---|---|---|---|---|---|---|
| seq | 96.0 | | 96.0 | | 97.2 | | 85.3 | |
| 1 | 106.1 (1.00) | 39856 | 97.7 (1.00) | 6254 | 96.8 (1.00) | 0 | 85.6 (1.00) | 0 |
| 2 | 53.9 (1.97) | 39856 | 50.8 (1.92) | 13562 | 49.8 (1.94) | 632 | 44.0 (1.95) | 613 |
| 4 | 27.8 (3.82) | 39856 | 26.6 (3.67) | 19481 | 26.1 (3.71) | 1917 | 23.3 (3.67) | 1946 |
| 8 | 15.1 (7.03) | 39856 | 14.6 (6.70) | 24150 | 14.2 (6.81) | 4414 | 13.0 (6.58) | 4807 |
| 16 | 8.9 (11.95) | 39856 | 8.7 (11.19) | 29324 | 8.3 (11.65) | 7911 | 7.8 (10.97) | 9930 |

Table 3: Performance of Mul-T benchmarks (absolute times are in seconds for Encore and in millions of simulated SPARC cycles for ALEWIFE).

consistently has the best time on 16 processors. In addition, lazy task creation shows better relative speedup than LBI, suggesting that it will scale better to larger systems.

**speech** ($g = 2410$) is part of a multi-stage speech understanding system under development at MIT. This stage is essentially a graph-matching problem, finding the closest dictionary entry to a spoken utterance. The program contains about 150 steps separated by barrier synchronizations; each step contains 200-300 parallel tasks of rather coarse average granularity. The coarse granularity means that eager task creation doesn't perform too badly, so the improvement with lazy task creation is modest. The barrier synchronizations cause significant idleness, hurting speedup for all strategies.

The statistics we have gathered do not allow precise conclusions about the extent of multiprocessing overhead from sources such as cache turbulence and contention for shared resources. However, because speedup for the finer-grained benchmarks is close to linear with lazy task creation, we can conclude that the effect of these other sources of overhead is fairly small for that strategy.

# 6 Related Work

Load-based inlining has been studied previously in the Mul-T parallel Lisp system [17], and is also available in Qlisp by using **(deque-size)** or **(qemptyp)** to sense the current load [10, 29]. An analytical model of load-based inlining for programs like **psum-tree** has been developed by Weening [29, 30]. His analytical results generally agree with empirical observations of load-based inlining in both Mul-T and Qlisp; however, neither the prior Mul-T work nor the prior Qlisp work have explored the alternative of lazy task creation.

Pehoushek and Weening [29] also present a strategy which reduces task creation overhead when a queued task is executed by the processor that created it. This strategy takes advantage of the same phenomenon that lazy task creation leverages: that when parallelism is abundant most tasks are executed locally. Executing such tasks with lazy task creation appears to be cheaper than with their scheme; furthermore, their scheme only works in programs with a fork/join style of parallelism. Lazy task creation has no such restriction, interacting well with the unlimited lifetime of futures in Mul-T.

WorkCrews [27] is a package that does perform lazy task creation, intended for use with a fork-join or **cobegin** style of programming. It is implemented on top of Modula-2+ (an extension of Modula-2). For every task that is to be created lazily, a WorkCrews program calls **RequestHelp**(*proc*, *data*) and then proceeds with other work. A free processor looks for unanswered help requests, "steals" one, and applies its *proc* to its *data*. When the requester finishes its other work, it calls **GotHelp** to see whether the **RequestHelp** task was stolen. If not, it proceeds to do the work itself; if so, it looks for other work to do. The performance of WorkCrews was evaluated on several parallel Quicksort programs and on **MultiGrep**, a program that searches for occurrences of a given string in a group of files [27].

The principal difference between WorkCrews-style lazy task creation and Mul-T's lazy futures is that invoking lazy task creation in WorkCrews requires a significantly larger amount of source code to be written—the work performed by *proc* must be broken out into a separate procedure, the argument block to be passed as *data* must be explicitly allocated and filled in, and finally the **RequestHelp** and **GotHelp** procedures must be called. Moreover, synchronization with and value retrieval from the lazily created task are explicit responsibilities of the programmer. By contrast, in Mul-T it is only necessary to insert the keyword **future** to begin enjoying the benefits of lazy task creation.

These stylistic differences lead to some implementation differences: our lazy future implementations directly manipulate implementation objects such as stack frames and are thus more "built in" to the implementation than in the case of WorkCrews. We think some efficiency improvements result from our approach, but the systems are different enough that it is hard to make a conclusive comparison. In any case, although the mechanics of the two systems are rather different, there is a very close relationship between their underlying philosophies.

Motivated by the idea of lazy futures presented in [17], Feeley has independently implemented lazy task creation in a parallel version of Scheme which runs on the BBN Butterfly [12]. His implementation is roughly similar to our Encore implementation, and contains some innovative features.

Our philosophy of encouraging programmers to expose parallelism while relying on the implementation to curb excess parallelism resembles that of data-flow researchers who have been concerned with *throttling* [6, 23]. However, the main purpose of throttling is to reduce the memory requirements of parallel computations, not to increase granularity (which is generally fixed at a very fine level by data-flow architectures [3, 11]). Throttling thus serves the same purpose as our preference for depth-first scheduling and is not directly related to lazy task creation.

# 7 Conclusions and Future Work

We are encouraged that our performance statistics support the theoretical benefits of lazy task creation. For programs with bushy call trees the programmer can use `future` to identify parallelism, effectively ignoring granularity considerations.

A remaining challenge are fine-grained programs without bushy call trees, such as those with data-level parallelism expressed iteratively (see Section 3.5). For example, consider a program fragment which performs a fine-grained operation on all elements of an array using an iterative loop, creating one task per element. This program will not execute efficiently in parallel unless its granularity is increased so that tasks handle several array elements instead of just one, but dynamic methods alone are unlikely to partition this program effectively because they are unable to change program structure. If the iterative structure of the program is obeyed, parallelism is inherently limited.

If instead of using iteration this program were restructured to perform a divide-and-conquer division of the array's index set, we know that lazy task creation would achieve the desired partition. But such a restructuring has two problems: it raises program complexity and it lowers program efficiency. To address the complexity problem we envision expressing such parallel operations on data aggregates at a higher level, converting the high-level expressions to appropriate divide-and-conquer divisions at compile time. Ideas for how to express such high-level operations appear in the work of Waters [28], Steele and Hillis [26], and Sabot [24].

The efficiency problem arises because the execution overhead of a divide-and-conquer division is large compared to the low overhead of an iterative loop. This overhead can be reduced substantially by smart compilation, but it will still be significant if the inner loop code is fine-grained. We observe though that a fine-grained inner loop is very likely to contain straight-line code rather than additional loops or calls to unknown procedures, so estimating its cost should be straightforward. Knowing the inner loop cost allows the compiler to unroll enough iterations to balance out the overhead of a divide-and-conquer division.

There is also the important issue of scalability. In both the Encore machine and the ALEWIFE simulator (with memory delays turned off), all memory references are of approximately equal cost, an unreasonable assumption for a large-scale multiprocessor. We are investigating how our lazy task creation strategy can be augmented to take advantage of locality in shared address-space systems where the physical memory is distributed, such as the ALEWIFE machine.

Because of their extra record-keeping burden, lazy future calls are unlikely ever to be as cheap as the cheapest implementation of normal calls, but the incremental cost of a lazy future call can be strongly influenced by a multiprocessor's hardware architecture. For example, the linked-frame implementation shown in Section 4.3 benefits greatly from the ALEWIFE architecture's support for full/empty bits in memory that can be accessed efficiently as a side effect of a load or store instruction.

Nevertheless, the linked-frame implementation still requires some memory operations for every call, and even a few more memory operations for every lazy future call. For architectures whose processors have register windows we have contemplated another approach with the potential of eliminating most memory operations: each register window could have an associated bit in a processor register indicating whether it is logically part of the lazy task queue, but only when a register window was unloaded due to a window overflow trap would the frame actually be linked into the in-memory data structure representing the queue. This would further reduce the cost of lazy future calls, since one might expect a large fraction of lazy future calls to return without their associated register window ever having been unloaded. However, some mechanism would have to be provided for querying a processor to see if it contains any stealable continuations (in the event that none are found in memory) and for interrupting a processor to request it to unload stealable continuations needed by other processors. The costs and benefits of this idea are not currently known.

The larger quest in which we have been engaged is to provide the expressive power and elegance of `future` at the lowest possible cost. Complete success in this endeavor would make it unnecessary for programmers ever to shun `future` in favor of lower-level, but more efficient, constructs. Success would also encourage programmers to express the parallelism in programs at all levels of granularity, rather than forcing them to hand-tune the granularity (at the source-code level) for the best performance. Lazy task creation moves us closer to this ideal, producing very acceptable performance and greatly reducing the number of tasks created for all of the benchmark programs in Section 5. And while the ideal may never be achieved completely, every step in the direction of making `future` cheaper increases the number of situations in which the cost of `future` is no bar to its use.

# 8 Acknowledgments

Special thanks to Dan Nussbaum for his contributions to the ALEWIFE implementation. Thanks to Randy Osborne, Richard Kelsey, and the anonymous referees for helpful comments on drafts of the paper, to Kirk Johnson for the speech application, to Marc Feeley for the elegant bit-vector version of $n$ queens, and to the Sloan foundation, IBM, Digital Equipment Corporation, the Department of Energy (FG02-86ER25012) and DARPA (N00014-87-K-0825) for their support.

# References


[1] A. Agarwal, et. al., "The MIT Alewife Machine: A Large-Scale Distributed-Memory Multiprocessor," in M. DuBois and S. Thakkar, eds., *Scalable Shared Memory Multiprocessors*, Kluwer Academic Publishers, 1991.

[2] A. Agarwal, B. H. Lim, D. Kranz, and J. Kubiatowicz, "APRIL: A Processor Architecture for Multiprocessing," *17th Annual Int'l. Symp. on Computer Architecture*, Seattle, Wa., May 1990, pp. 104–114.

[3] Arvind and D. Culler, "Dataflow Architectures," *Annual Reviews in Computer Science*, Annual Reviews, Inc., Palo Alto, Ca., 1986, pp. 225–253.

[4] G. Bilardi and A. Nicolau, "Adaptive Bitonic Sorting: An Optimal Parallel Algorithm for Shared-Memory Machines," *SIAM J. Comput.*, 18:2, April 1989, pp. 216–228.

[5] D. Chaiken, J. Kubiatowicz, and A. Agarwal, "LimitLESS Directories: A Scalable Cache Coherence Scheme," *4th Int'l. Conf. on Architectural Support for Programming Languages and Operating Systems*, Santa Clara, Ca., April 1991.

[6] D. E. Culler, "Managing Parallelism and Resources in Scientific Dataflow Programs," Ph.D. thesis, M.I.T. Dept. of Electrical Engineering and Computer Science, Cambridge, Mass., June 1989.

[7] R. P. Gabriel and J. McCarthy, "Queue-based Multi-processing Lisp," *1984 ACM Symp. on Lisp and Functional Programming*, Austin, Tex., Aug. 1984, pp. 25–44.

[8] B. Goldberg, "Multiprocessor Execution of Functional Programs," *Int'l. J. of Parallel Programming* 17:5, Oct. 1988, pp. 425–473.

[9] R. Goldman and R. P. Gabriel, "Preliminary Results with the Initial Implementation of Qlisp," *1988 ACM Symp. on Lisp and Functional Programming*, Snowbird, Utah, July 1988, pp. 143–152.

[10] R. Goldman, R. Gabriel, and C. Sexton, "Qlisp: An Interim Report," in T. Ito and R. Halstead, eds., *Proceedings of U.S./Japan Workshop on Parallel Lisp* (Springer-Verlag Lecture Notes in Computer Science 441), Sendai, Japan, June 1989, pp. 161–181.

[11] J. Gurd, C. Kirkham, and I. Watson, "The Manchester Prototype Dataflow Computer," *Comm. ACM* 28:1, January 1985, pp. 34–52.

[12] M. Feeley, "Fine grain parallelism in Multilisp on a shared memory MIMD computer," Ph.D. thesis, Brandeis University (in preparation).

[13] R. Halstead, "Multilisp: A Language for Concurrent Symbolic Computation," *ACM Trans. on Prog. Languages and Systems* 7:4, October 1985, pp. 501–538.

[14] R. Halstead, "An Assessment of Multilisp: Lessons from Experience," *Int'l. J. of Parallel Programming* 15:6, Dec. 1986, pp. 459–501.

[15] R. W. Hockney and C. R. Jesshope, *Parallel Computers 2*, Adam Hilger, Bristol and Philadelphia, 1988, pp. 475-83.

[16] P. Hudak and B. Goldberg, "Serial Combinators: 'Optimal' Grains of Parallelism," *Functional Programming Languages and Computer Architecture*, Springer-Verlag LNCS 201, September 1985, pp. 382–388.

[17] D. Kranz, R. Halstead, and E. Mohr, "Mul-T, A High-Performance Parallel Lisp", *ACM SIGPLAN '89 Conference on Programming Language Design and Implementation*, Portland, OR, June 1989, pp. 81–90.

[18] D. Kranz, R. Kelsey, J. Rees, P. Hudak, J. Philbin, and N. Adams, "Orbit: An Optimizing Compiler for Scheme," *Proc. SIGPLAN '86 Symp. on Compiler Construction*, June 1986, pp. 219–233.

[19] D. Kranz, "ORBIT: An Optimizing Compiler for Scheme," Ph.D. Thesis, Yale University Technical Report YALEU/DCS/RR-632, February 1988.

[20] E. Mohr, D. Kranz, and R. Halstead, "Lazy Task Creation: A Technique for Increasing the Granularity of Parallel Programs," *Proceedings of ACM Symposium on Lisp and Functional Programming*, June 1990, pp. 185–197.

[21] E. Mohr, D. Kranz, and R. Halstead, "Lazy Task Creation: A Technique for Increasing the Granularity of Parallel Programs," DEC Cambridge Research Lab Technical Report CRL 90/7, November 1990.

[22] J. E. B. Moss, "Managing Stack Frames in Smalltalk," *Proc. SIGPLAN '87 Symp. on Interpreters and Interpretive Techniques*, June 1987, pp. 229–240.

[23] C. A. Ruggiero and J. Sargeant, "Control of Parallelism in the Manchester Dataflow Machine," Springer-Verlag LNCS 274, *Functional Programming Languages and Computer Architecture*, Portland, Oregon, September 1987, pp. 1–15.

[24] G. Sabot, *The Paralation Model*, M.I.T. Press, 1988.

[25] V. Sarkar and J. Hennessy, "Compile-time Partitioning and Scheduling of Parallel Programs," *SIGPLAN '86 Symposium on Compiler Construction*, July 1986, pp. 17-26.

[26] G. L. Steele, Jr. and W. D. Hillis, "Connection Machine Lisp: Fine-Grained Parallel Symbolic Processing," *1986 ACM Symp. on Lisp and Functional Programming*, Cambridge, MA, August 1986, pp. 279-297.

[27] M. Vandevoorde and E. Roberts, "WorkCrews: An Abstraction for Controlling Parallelism," *Int'l. J. of Parallel Programming* 17:4, August 1988, pp. 347–366.

[28] R. C. Waters, "Series", in G. Steele, Jr., *Common Lisp: the Language*, Second Edition, Digital Press, Maynard MA, pp. 923-955.

[29] J. Weening, "An Analysis of Dynamic Partitioning," in T. Ito and R. Halstead, eds., *Proceedings of U.S./Japan Workshop on Parallel Lisp* (Springer-Verlag Lecture Notes in Computer Science 441), Sendai, Japan, June 1989, pp. 182–199.

[30] J. Weening, "Parallel Execution of Lisp Programs," *Stanford Computer Science Report* STAN-CS-89-1265, June 1989.

# OFFICIAL DISTRIBUTION LIST

| | |
|---|---|
| DIRECTOR<br>Information Processing Techniques Office<br>Defense Advanced Research Projects Agency (DARPA)<br>1400 Wilson Boulevard<br>Arlington, VA 22209 | 2 copies |
| OFFICE OF NAVAL RESEARCH<br>800 North Quincy Street<br>Arlington, VA 22217<br>Attn: Dr. Gary Koop, Code 433 | 2 copies |
| DIRECTOR, CODE 2627<br>Naval Research Laboratory<br>Washington, DC 20375 | 6 copies |
| DEFENSE TECHNICAL INFORMATION CENTER<br>Cameron Station<br>Alexandria, VA 22314 | 12 copies |
| NATIONAL SCIENCE FOUNDATION<br>Office of Computing Activities<br>1800 G. Street, N.W.<br>Washington, DC 20550<br>Attn: Program Director | 2 copies |
| HEAD, CODE 38<br>Research Department<br>Naval Weapons Center<br>China Lake, CA 93555 | 1 copy |