

Using a Swap Instruction to Coalesce Loads and Stores

Apan Qasem, David Whalley, Xin Yuan, and Robert van Engelen

Department of Computer Science, Florida State University
Tallahassee, FL 32306-4530, U.S.A.

e-mail: {qasem,whalley,xyuan,engelen}@cs.fsu.edu, phone: (850) 644-3506

Abstract. A *swap* instruction, which exchanges a value in memory with a value of a register, is available on many architectures. The primary application of a swap instruction has been for process synchronization. In this paper we show that a swap instruction can often be used to coalesce loads and stores in a variety of applications. We describe the analysis necessary to detect opportunities to exploit a swap and the transformation required to coalesce a load and a store into a swap instruction. The results show that both the number of accesses to the memory system (data cache) and the number of executed instructions are reduced. In addition, the transformation reduces the register pressure by one register at the point the swap instruction is used, which sometimes enables other code-improving transformations to be performed.

1 INTRODUCTION

An instruction that exchanges a value in memory with a value in a register has been used on a variety of machines. The primary purpose for these *swap* instructions is to provide an atomic operation for reading from and writing to memory, which has been used to construct mutual-exclusion mechanisms in software for process synchronization. In fact, there are other forms of hardware instructions that have been used to support mutual exclusion, which include the classic *test-and-set* instruction. In this paper we show that a swap instruction can also be used by a low-level code-improving transformation to coalesce loads and stores into a single instruction.

A swap instruction exchanges a value in memory with a value in a register. This is illustrated in Figure 1, which depicts a load instruction, a store instruction, and a swap instruction using an RTL (register transfer list) notation. Each assignment in an RTL represents an effect on the machine. The list of effects within a single RTL are accomplished in parallel. Thus, the swap instruction is essentially a load and store accomplished in parallel.

A swap instruction can be efficiently integrated into a conventional RISC architecture. First, it can be encoded using the same format as a load or a store since all three instructions reference a register and a memory address. Only additional opcodes are required to support the encoding of a swap instruction.

$r[2] = M[x];$	$M[x] = r[2];$	$r[2] = M[x]; M[x] = r[2];$
(a) Load Instruction	(b) Store Instruction	(c) Swap Instruction

Fig. 1. Contrasting the Effects of Load, Store, and Swap Instructions

Second, access to a data cache can be performed efficiently for a swap instruction on most RISC machines. A direct-mapped data cache can send the value to be loaded from memory to the processor for a load or a swap instruction in parallel with the tag check. This value will not be used by the processor if a tag mismatch is later discovered [1]. A data cache is not updated with the value to be stored by a store or a swap instruction until after the tag check [1]. Thus, a swap instruction could be performed as efficiently as a store instruction on a machine with a direct-mapped data cache. In fact, a swap instruction requires the same number of cycles in the pipeline as a store instruction on the MicroSPARC I [2]. One should note that a swap instruction will likely perform less efficiently when it is used for process synchronization on a multiprocessor machine since it requires a signal over the bus to prevent other accesses to memory.

Finally, it is possible that a main memory access could also be performed efficiently for a swap instruction. Reads to DRAM are destructive, meaning that the value read must be written back afterwards. A DRAM organization could be constructed where the value that is written back could differ from the value that was read and sent to the processor. Thus, a load and a store to a single word of main memory could occur in one main memory access cycle.

The remainder of this paper has the following organization. First, we introduce related work that allows multiple accesses to memory to occur simultaneously. Second, we describe a variety of different opportunities for exploiting a swap instruction that commonly appear in applications. Third, we present an algorithm to detect and coalesce a load and store pair into a swap instruction and discuss issues related to implementing this code-improving transformation. Fourth, we present the results of applying the code-improving transformation on a variety of applications. Finally, we present the conclusions of the paper.

2 RELATED WORK

There has been some related work that allows multiple accesses to the memory system to occur in a single cycle. Superscalar and VLIW machines have been developed where a wider datapath between the data cache and the processor has been used to allow multiple simultaneous accesses to the data cache. Likewise, a wider datapath has been implemented between the data cache and main memory to allow multiple simultaneous accesses to main memory through the use of memory banks. A significant amount of compiler research has been spent on trying to schedule instructions so that multiple independent memory accesses to different banks can be performed simultaneously [3].

Memory access coalescing is a code-improving transformation that groups multiple memory references to consecutive memory locations into a single larger memory reference. This transformation was accomplished by recognizing a contiguous access pattern for a memory reference across iterations of a loop, unrolling the loop, and rescheduling instructions so that multiple loads or stores could be coalesced [4].

Direct hardware support of multiple simultaneous memory accesses in the form of superscalar or VLIW architectures requires that these simultaneous memory accesses be independent in that they access different memory locations. Likewise, memory access coalescing requires that the coalesced loads or stores access contiguous (and different) memory locations. In contrast, the use of a swap instruction allows a store and a load to the same memory location to be coalesced together and performed simultaneously. In a manner similar to the memory access coalescing transformation, a load and a store are coalesced together and explicitly represented in a single instruction.

3 OPPORTUNITIES FOR EXPLOITING THE SWAP INSTRUCTION

A swap instruction can potentially be exploited when a load is followed by a store to the same memory address and the value stored is not computed using the value that was loaded. We investigated how often this situation occurs and we have found many direct opportunities in a number of applications. Consider the following code segment in Figure 2 from an application that uses polynomial approximation from Chebyshev coefficients [5]. There is first a load of the $d[k]$ array element followed by a store to the same element, where the store does not use the value that was loaded. We have found comparable code segments containing such a load followed by a store in other diverse applications, such as Gauss-Jordan elimination [5] and tree traversals [6].

```

...
sv = d[k];
d[k] = 2.0*d[k-1] - dd[k];
...

```

Fig. 2. Code Segment in Polynomial Approximation from Chebyshev Coefficients

A more common operation where a swap instruction can be exploited is when the values of two variables are exchanged. Consider Figure 3(a), which depicts the exchange of the values in x and y at the source code level. Figure 3(b) indicates that the load and store of x can be coalesced together. Likewise, Figure 3(c) indicates that the load and store of y can also be coalesced together.

However, we will discover in the next section that only a single pair of load and store instructions in an exchange of values between variables can be coalesced together.

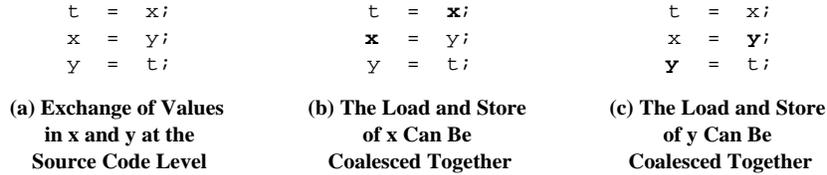


Fig. 3. Example of Exchanging the Values of Two Variables

There are numerous applications where the values of two variables are exchanged. Various sorts of an array or list of values are obvious applications in which a swap instruction could be exploited. Some other applications requiring an explicit exchange of values between two variables include transposing a matrix, the traveling salesperson problem, solving linear algebraic equations, fast fourier transforms, and the integration of differential equations. The above list is only a small subset of the applications that require this basic operation.

There are also opportunities for exploiting a swap instruction after other code-improving transformations have been performed. Consider the code segment in Figure 4(a) from an application that uses polynomial approximation from Chebyshev coefficients [5]. It would appear in this code segment that there is no opportunity for exploiting a swap instruction. However, consider the body of the loop executed across two iterations, which is shown in Figure 4(b) after unrolling the loop by a factor of two. For simplicity, we are assuming in this example that the original loop iterated an even number of times. Now the value loaded from `d[j-1]` in the first assignment statement in the loop is updated in the second assignment statement and the value computed in the first assignment is not used to compute the value stored in the second assignment. We have found opportunities for exploiting a swap instruction across loop iterations by loop unrolling in a number of applications, which includes linear prediction, interpolation and extrapolation, and solution of linear algebraic equations.

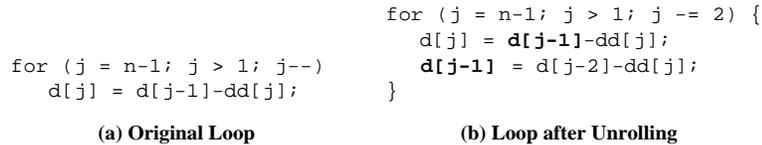


Fig. 4. Example of Unrolling a Loop to Provide an Opportunity to Exploit a Swap Instruction

Equation 1 indicates the number of memory references saved for each load and store pair that can be coalesced across loop iterations. In other words, one memory reference is saved each time the loop is unrolled twice. Of course, loop unrolling has additional benefits, such as reduced loop overhead and better opportunities for scheduling instructions.

$$\text{Memory References Saved} = \lfloor \frac{\text{loop unroll factor}}{2} \rfloor \quad (1)$$

Finally, we have also discovered opportunities for speculatively exploiting a swap instruction across basic blocks. Consider the code segment in Figure 5, which assigns values to an image according to a specified threshold [7]. `p[i][j]` is loaded in one block and a value is assigned to `p[i][j]` in both of its successor blocks. The load of `p[i][j]` and the store from the assignment to `p[i][j]` in the `then` or `else` portions of the `if` statement can be coalesced into a swap instruction since the value loaded is not used to compute the value stored. The store operation can be speculatively performed as part of a swap instruction in the block containing the load. We have found that stores can be performed speculatively in a number of other image processing applications, which include clipping and arithmetic operations.

```

for (i = 0; i < n; i++)
  for (i = 0; i < m; i++) {
    if p[i][j] <= t)
      p[i][j] = 1;
    else
      p[i][j] = 0;
  }

```

Fig. 5. Speculative Use of a Swap Instruction

4 A CODE-IMPROVING TRANSFORMATION TO EXPLOIT THE SWAP INSTRUCTION

Figure 6(a) illustrates the general form of a load followed by a store that can be coalesced. The memory reference is to the same variable or location and the register loaded (`r[a]`) and register stored (`r[b]`) differ. Figure 6(b) depicts the swap instruction that represents the coalesced load and store. Note that the register loaded has been renamed from `r[a]` to `r[b]`. This renaming is required since the swap instruction has to store from and load into the same register.

Figure 7(a), like Figure 3(a), shows an exchange of the values of two variables, `x` and `y`, at the source code level. Figure 7(b) shows similar code at the SPARC machine code level, which is represented in RTLs. The variable `t` has been allocated to register `r[1]`. Register `r[2]` is used to hold the temporary value loaded from `y` and stored in `x`. At this point a swap could be used to

<pre> r[a] = M[v]; ... M[v] = r[b]; </pre>	<pre> r[b] = M[v]; M[v] = r[b]; </pre>
(a) Load Followed by a Store	(b) Coalesced Load and Store

Fig. 6. Simple Example of Coalescing a Load and Store into a Swap Instruction

coalesce the load and store of x or the load and store of y . Figure 7(c) shows the RTLs after coalescing the load and store of x . One should note that $r[1]$ is no longer used since its live range has been renamed to $r[2]$. Due to the renaming of the register, the register pressure at this point in the program flow graph has been reduced by one. Reducing the register pressure can sometimes enable other code-improving transformations that require an available register to be applied. Note that the decision to coalesce the load and store of x prevents the coalescing of the load and store of y .

<pre> t = x; x = y; y = t; </pre>	<pre> r[1] = M[x]; r[2] = M[y]; M[x] = r[2]; M[y] = r[1]; </pre>
(a) Exchange of Values in x and y at the Source Code Level	(b) Exchange of Values in x and y at the Machine Code Level

```

r[2] = M[y];
M[x] = r[2]; r[2] = M[x];
M[y] = r[2];

```

(c) After Coalescing the
Load and Store of x

Fig. 7. Example of Exchanging the Values of Two Variables

The code-improving transformation to coalesce a load and a store into a swap instruction was accomplished using the algorithm in Figure 8. The algorithm finds a load followed by a store to the same address and coalesces the two memory references together into a single swap instruction if a variety of conditions are met. These conditions include: (1) the load and store must be within the same block or consecutively executed blocks, (2) the addresses of the memory references in the load and store instructions have to be the same, (3) the value in $r[b]$ that will be stored cannot depend on the value loaded into $r[a]$, (4) the value in $r[b]$ cannot be used after the store instruction, and (5) $r[a]$ has to be able to be renamed to $r[b]$. The following subsections describe issues relating to this code-improving transformation.

```

FOR B = each block in function DO
  FOR LD = each instruction in B DO
    IF LD is a load AND Find_Matching_Store(LD, B, LD->next, ST, P)
      AND Meet_Swap_Conds(LD, ST) THEN
      SW = Create("%s=M[%s];M[%s]=%s;",ST->r[b], LD->load_addr,
        LD->load_addr, ST->r[b]);
      Insert SW before P;
      Replace uses of L->r[a] with S->r[b] until L->r[a] dies;
      Delete LD and ST;

BOOL Find_Matching_Store(LD, B, FIRST, ST, P) {
FOR ST = FIRST to B->last DO
  IF ST is a store THEN
    IF ST->store_addr == LD->load_addr THEN
      If FIRST == B->first THEN
        RETURN TRUE;
      ELSE
        RETURN Find_Place_To_Insert_Swap(LD, ST, P);
    IF ST->store_addr != LD->load_addr THEN
      CONTINUE;
    IF cannot determine if the two addresses are same or different THEN
      RETURN FALSE;
  IF FIRST == B->first THEN
    RETURN FALSE;
FOR S = each successor of B DO
  IF !Find_Matching_Store(LD, S, S->first, ST, P) THEN
    RETURN FALSE;
FOR S = each successor of B DO
  IF Find_Place_To_Insert_Swap(LD, ST, P) THEN
    RETURN TRUE;
RETURN FALSE;
}

BOOL Meet_Swap_Conds(LD, ST){
RETURN (value in ST->r[b] is guaranteed to not depend on the value in LD->r[a]
  AND (ST->r[b] dies at the store)
  AND ((ST->r[b] is not reset before LD->r[a] dies)
  OR (other live range of S->r[b] can be renamed to use another register));
}

BOOL Find_Place_To_Insert_Swap(LD, ST, P){
IF LD->r[a] is not used between LD and ST THEN
  P = ST;
  RETURN TRUE;
IF ST->r[b] is not referenced between LD and ST THEN
  P = LD->next;
  RETURN TRUE;
IF first use of LD->r[a] after LD comes after the last reference to ST->r[b]
  before the store THEN
  P = instruction containing first use of LD->r[a] after LD;
  RETURN TRUE;
IF first use of LD->r[a] after LD can be moved after the last reference
  to ST->r[b] before the store THEN
  Move instructions as needed;
  P = instruction containing first use of LD->r[a] after LD;
  RETURN TRUE;
ELSE
  RETURN FALSE;
}

```

Fig. 8. Algorithm for Coalescing a Load and a Store into a Swap Instruction

4.1 Performing the Code-Improving Transformation Late in the Compilation Process

Sometimes apparent opportunities at the source code level for exploiting a swap instruction are not available after other code-improving transformations have been applied. Many code-improving transformations either eliminate memory references (e.g. register allocation) or move memory references (e.g. loop-invariant code motion). Coalescing loads and stores into swap instructions should only be performed after all other code-improving transformations that can affect the memory references have been applied. Figure 9(a) shows an exchange of values after the two values are compared in an `if` statement. Figure 9(b) shows a possible translation of this code segment to machine instructions. Due to common subexpression elimination, the loads of `x` and `y` in the block following the branch have been deleted in Figure 9(c). Thus, the swap instruction cannot be exploited within that block. This example illustrates why the swap instruction should be performed late in the compilation process when the actual loads and stores that will remain in the generated code are known.

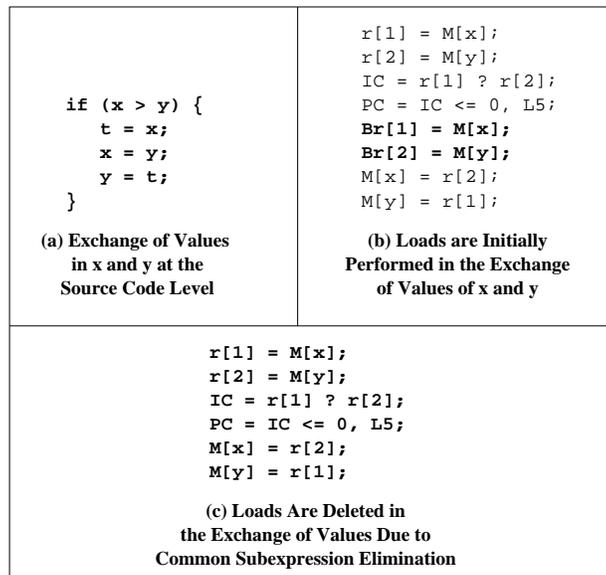


Fig. 9. Example Depicting Why the Swap Instruction Should Be Exploited as a Low-Level Code-Improving Transformation

4.2 Ensuring Memory Addresses Are Equivalent Or Are Different

One of the requirements for a load and store to be coalesced is that the load and store must refer to the same address. Figure 10(a) shows a load using the ad-

dress in register $r[2]$ and a store using the address in $r[4]$. The compiler must ensure that the value in $r[2]$ is the same as that in $r[4]$. This process of checking that two addresses are equivalent is complicated due to the code-improving transformation being performed late in the compilation process. Common sub-expression elimination and loop-invariant code motion may move the assignments of addresses to registers far from where they are actually dereferenced.

<pre> r[2] = r[3] << 2; r[2] = r[2] + _a; r[5] = M[r[2]]; ... r[4] = r[3] << 2; r[4] = r[4] + _a; M[r[4]] = r[6]; </pre>	<pre> r[a] = M[v]; ... M[r[c]] = r[d]; ... M[v] = r[b]; </pre>
<p>(a) Same Addresses after Expanding the Expressions</p>	<p>(b) Load and Store to the Same Variable with an Intervening Store</p>

Fig. 10. Examples of Detecting If Memory Addresses Are the Same or Differ

We implemented some techniques to determine if the addresses of two memory references were the same or if they differ. Addresses to memory were expanded by searching backwards for assignments to registers in the address until all registers are replaced or the beginning of a block with multiple predecessors is encountered. For instance, the address in the memory reference being loaded in Figure 10(a) is expanded as follows:

$$r[2] \Rightarrow r[2] + _a \Rightarrow (r[3] \ll 2) + _a$$

The address in the memory reference being stored would be expanded in a similar manner. Once the addresses of two memory references have been expanded, then they are compared to determine if they differ. If the expanded addresses are syntactically equivalent, then the compiler has ensured that they refer to the same address in memory.

We also associated the expanded addresses with memory references before code-improving transformations involving code motion were applied. The compiler tracked these expanded addresses with the memory references through a variety of code-improving transformations that would move the location of the memory references. Determining the expanded addresses early simplified the process of calculating addresses associated with memory references.

Another requirement for a load and a store to be coalesced is that there are no other possible intervening stores to the same address. Figure 10(b) shows a load of a variable v followed by a store to the same variable with an intervening store. The compiler must ensure that the value in $r[c]$ is not the address of the variable v . However, simply checking that two expanded addresses are not identical does not suffice to determine if they refer to differ locations in memory.

Various rules were used to determine if two addresses differed. Table 1 depicts some of these rules that were used.

Num	Rule	Example	
		First Address	Second Address
1	The addresses are to different classes (local variables, arguments, static variables, and global variables).	$M[_a]$	$M[r[30]+x]$
2	Both addresses are to the same class and their name differs.	$M[_a]$	$M[_b]$
3	One address is to a variable that has never had its address taken and the second address is not to the same variable.	$M[r[14]+v]$	$M[r[7]]$
4	The addresses are the same, except for different constant offsets.	$M[(r[3] \ll 2) + _a]$	$M[(r[3] \ll 2) + _a+4]$

Table 1. A Subset of the Rules Used for Memory Disambiguation

4.3 Finding a Location to Place the Swap Instruction

Another condition that has to be met for a load and a store to be coalesced into a swap instruction is that the instruction containing the first use of register assigned by the load has to occur after the last reference to the register to be stored. For example, consider the example in Figure 11(a). A use of $r[a]$ appears after the last reference to $r[b]$ before the store instruction, which prevents the load and store from being coalesced. Figure 11(b) shows that the compiler is sometimes able to reschedule the instructions between the load and the store to meet this condition. Now the load and the store can be moved where the load appears immediately before the store, as shown in Figure 11(c). Once the load and store are contiguous, the two instructions can be coalesced. Figure 11(d) shows the code sequence after the load and store has been deleted, the swap instruction inserted, and $r[a]$ has been renamed to $r[b]$.

4.4 Renaming Registers to Allow the Swap Instruction to Be Exploited

We encountered another complication due to coalescing loads and stores into swap instructions late in the compilation process. Pseudo registers, which contain temporary values, have already been assigned to hardware registers when the coalescing transformation is attempted. The compiler reuses hardware registers when assigning pseudo registers to hardware registers in an attempt to minimize the number of hardware registers used. Our implementation of the code-improving transformation sometimes renamed live ranges of registers to

<pre> r[a] = M[v]; = ... r[a] ...; ... = ... r[b] ...; ... M[v] = r[b]; </pre> <p>(a) Use of r[a] Appears before a Reference to r[b]</p>	<pre> r[a] = M[v]; = ... r[b] ...; ... = ... r[a] ...; ... M[v] = r[b]; </pre> <p>(b) First Use of r[a] Appears after the Last Reference to r[b]</p>
<pre> = ... r[b] ...; r[a] = M[v]; M[v] = r[b]; ... = ... r[a] ...; ... </pre> <p>(c) Load and Store Can Now Be Made Contiguous</p>	<pre> = ... r[b] ...; r[b] = M[v]; M[v] = r[b]; ... = ... r[b] ...; ... </pre> <p>(d) After Coalescing the Load and Store and Renaming r[a] to Be r[b]</p>

Fig. 11. Examples of Finding a Location to Place the Swap Instruction

permit the use of a swap instruction. Consider the example in Figure 12(a), which contains a set of $r[b]$ after the store and before the last use of the value assigned to $r[a]$. In this situation, we attempt to rename the second live range of $r[b]$ to a different available register. Figure 12(b) shows this live range being renamed to $r[c]$. Figure 12(c) depicts that the load and store can now be coalesced since $r[a]$ can be renamed to $r[b]$.

Sometimes we had to move sequences of instructions past other instructions in order for the load and store to be coalesced. Consider the unrolled loop in Figure 4. Figure 13(a) shows the same loop, but in a load/store fashion, where the temporaries are registers. The load and store cannot be made contiguous due to reuse of the same registers. Figure 13(b) shows the same code after renaming the registers on which the value to be stored depends. Now the instructions can be scheduled so that the load and store can be made contiguous as shown in Figure 13(c). Figure 13(d) shows the load and store coalesced and the loaded register renamed.

5 RESULTS

Table 2 describes the numerous benchmarks and applications that we used to evaluate the impact of applying the code-improving transformation to coalesce loads and stores into a swap instruction. The programs depicted in boldface were directly obtained from the Numerical Recipes in C text [5]. The code in many of these benchmarks are used as utilities in a variety of programs. Thus, coalescing loads and stores into swaps can be performed on a diverse set of applications.

Measurements were collected using the *ease* system that is available with the *vpo* compiler. In some cases, we emulated a swap instruction when it did not exist. For instance, the SPARC does not have swap instructions that swaps

<pre> r[a] = M[v]; ... M[v] = r[b]; ... r[b] = ...; = ... r[a] ...; = ... r[b] ...; </pre> <p style="text-align: center;">(a) r[b] Is Set in the Live Range of r[a]</p>	<pre> r[a] = M[v]; ... M[v] = r[b]; ... r[c] = ...; = ... r[a] ...; = ... r[c] ...; </pre> <p style="text-align: center;">(b) Live Range of r[b] after Store Has Been Renamed to r[c]</p>
<pre> ... r[b] = M[v]; M[v] = r[b]; ... r[c] = ...; = ... r[b] ...; = ... r[c] ...; </pre> <p style="text-align: center;">(c) After Coalescing the Load and Store and Renaming the Live Range of r[a] to r[b]</p>	

Fig. 12. Example of Applying Register Renaming to Permit the Use of a Swap Instruction

<pre> for (j = n-1; j > 1; j -= 2) { r[1] = d[j-1]; r[2] = dd[j]; r[1] = r[1]-r[2]; d[j] = r[1]; r[1] = d[j-2]; r[2] = dd[j-1]; r[1] = r[1]-r[2]; d[j-1] = r[1]; } </pre> <p style="text-align: center;">(a) After Loop Unrolling</p>	<pre> for (j = n-1; j > 1; j -= 2) { r[1] = d[j-1]; r[2] = dd[j]; r[1] = r[1]-r[2]; d[j] = r[1]; r[3] = dd[j-1]; r[4] = d[j-2]; r[3] = r[3]-r[4]; d[j-1] = r[3]; } </pre> <p style="text-align: center;">(b) After Register Renaming</p>
<pre> for (j = n-1; j > 1; j -= 2) { r[3] = d[j-2]; r[4] = dd[j-1]; r[3] = r[3]-r[4]; r[1] = d[j-1]; d[j-1] = r[3]; r[2] = dd[j]; r[1] = r[1]-r[2]; d[j] = r[1]; } </pre> <p style="text-align: center;">(c) After Scheduling the Instructions</p>	<pre> for (j = n-1; j > 1; j -= 2) { r[3] = d[j-2]; r[4] = dd[j-1]; r[3] = r[3]-r[4]; r[3] = d[j-1]; d[j-1] = r[3]; r[2] = dd[j]; r[1] = r[3]-r[2]; d[j] = r[1]; } </pre> <p style="text-align: center;">(d) After Coalescing the Load and Store</p>

Fig. 13. Another Example of Applying Register Renaming to Permit the Use of a Swap Instruction

Program	Description
bandec	constructs an LU decomposition of a sparse representation of a band diagonal matrix
bubblesort	sorts an integer array in ascending order using a bubble sort
chebpc	polynomial approximation from Chebyshev coefficients
elmhes	reduces an $N \times N$ matrix to Hessenberg form
fft	fast fourier transform
gaussj	solves linear equations using Gauss-Jordan elimination
indexx	cal. indices for the array such that the indices are in ascending order
ludcmp	performs LU decomposition of an $N \times N$ matrix
mmid	modified midpoint method
predic	performs linear prediction of a set of data points
rtflsp	finds the root of a function using the false position method
select	returns the k smallest value in an array
thresh	adjusts an image according to a threshold value
transpose	transposes a matrix
traverse	binary tree traversal without a stack
tsp	traveling salesman problem

Table 2. Test Programs

bytes, halfwords, floats, or doublewords. The *ease* system provides the ability to gather measurements on proposed architectural features that do not exist on a host machine [8,9]. Note that it is sometimes possible to use the SPARC swap instruction, which exchanges a word in an integer register with a word in memory, for exchanging a floating-point value with a value in memory. When the floating-point values that are loaded and stored are not used in any operations, then these values could be loaded and stored using integer registers instead of floating-point registers and the swap instruction could be exploited.

Table 3 depicts the results that were obtained on the test programs for coalescing loads and stores into swap instructions. We unrolled several loops in these programs by an unroll factor of two to provide opportunities for coalescing a load and a store across the original iterations of the loop. In these cases, the *Not Coalesced* column includes the unrolling of these loops to provide a fair comparison. The results show decreases in the number of instructions executed and memory references performed for a wide variety of applications. The amount of the decrease varied depending on the execution frequency of the load and store instructions that were coalesced.

6 CONCLUSIONS

In this paper we have shown how to exploit a swap instruction, which exchanges the values between a register and a location in memory. We have discussed how a swap instruction could be efficiently integrated into a conventional load/store architecture. A number of different types of opportunities for exploiting the swap instruction were shown to be available. An algorithm for coalescing a load and a store into a swap instruction was given and a number of issues related to implementing the coalescing transformations were described. The results show that this code-improving transformation could be applied on a variety of applications and benchmarks and reductions in the number of instructions executed and memory references performed were observed.

Program	Instructions Executed			Memory References Performed		
	Not Coalesced	Coalesced	Decrease	Not Coalesced	Coalesced	Decrease
bandec	69,189	68,459	1.06%	18,054	17,324	4.04%
bubblesort	2,439,005	2,376,705	2.55%	498,734	436,434	12.49%
chebpc	7,531,984	7,029,990	6.66%	3,008,052	2,507,056	16.66%
elmhes	18,527	18,044	2.61%	3,010	2,891	3.95%
fft	4,176,112	4,148,112	0.67%	672,132	660,932	1.67%
gaussj	27,143	26,756	1.43%	7,884	7,587	3.77%
indexx	70,322	68,676	2.34%	17,132	15,981	6.72%
ludcmp	10,521,952	10,439,152	0.79%	854,915	845,715	1.08%
mmid	267,563	258,554	3.37%	88,622	79,613	10.17%
predic	40,827	38,927	4.65%	13,894	11,994	13.67%
rtflsp	81,117	80,116	1.23%	66,184	65,183	1.51%
select	19,939	19,434	2.53%	3,618	3,121	13.74%
thresh	7,958,909	7,661,796	3.73%	1,523,554	1,226,594	19.49%
transpose	42,883	37,933	11.54%	19,832	14,882	24.96%
traverse	94,159	91,090	3.26%	98,311	96,265	2.08%
tsp	64,294,814	63,950,122	0.54%	52,144,375	51,969,529	0.34%
average	6,103,402	6,019,616	3.06%	3,689,893	3,622,568	8.52%

Table 3. Results

References

1. M. D. Hill, “A Case for Direct-Mapped Caches,” *IEEE Computer*, 21(11), pages 25–40, December 1988.
2. Texas Instruments, Inc., *Product Preview of the TMS390S10 Integrated SPARC Processor*, 1993.
3. J. Hennessy and D. Patterson, *Computer Architecture: A Quantitative Approach, Second Edition*, Morgan Kaufmann, San Francisco, CA, 1996.
4. J.W. Davidson and S. Jinturkar, “Memory Access Coalescing: A Technique for Eliminating Redundant Memory Accesses,” *Proceedings of the SIGPLAN’94 Symposium on Programming Language Design and Implementation*, pages 186–195, June 1994.
5. W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing, Second Edition*, Cambridge University Press, New York, NY, 1996.
6. B. Dwyer, “Simple Algorithms for Traversing a Tree without a Stack,” *Information Processing Letters*, 2(5), pages 143–145, 1973.
7. I. Pitas, *Digital Image Processing Algorithms and Applications*, John Wiley & Sons, Inc., New York, NY, 2000.
8. J.W. Davidson and D.B. Whalley, “Ease: An Environment for Architecture Study and Experimentation,” *Proceedings SIGMETRICS’90 Conference on Measurement and Modeling of Computer Systems*, Pages 259–260, May 1990.
9. J.W. Davidson and D.B. Whalley, “A Design Environment for Addressing Architecture and Compiler Interactions,” *Microprocessors and Microsystems*, 15(9), pages 459–472, November 1991.