Effectively Exploiting Indirect Jumps

GANG-RYUNG UH
Lucent Technologies, Allentown, PA 18103
tel: 610-712-2447, email:*uh*@lucent.com
and
DAVID B. WHALLEY
Department of Computer Science,
Florida State University,
tel: 850-644-3506, email:*whalley*@cs.fsu.edu

This paper describes a general code-improving transformation that can coalesce conditional branches into an indirect jump from a table. Applying this transformation allows an optimizer to exploit indirect jumps for many other coalescing opportunities besides the translation of multiway branch statements. First, dataflow analysis is performed to detect a set of coalescent conditional branches, which are often separated by blocks of intervening instructions. Second, several techniques are applied to reduce the cost of performing an indirect jump operation, often requiring the execution of only two instructions on a SPARC. Finally, the control flow is restructured using code duplication to replace the set of branches with an indirect jump. Thus, the transformation essentially provides early resolution of conditional branches that may originally have been some distance from the point where the indirect jump is inserted. The transformation can be frequently applied with often significant reductions in the number of instructions executed, total cache work, and execution time. In addition, we show that with branch target buffer support, indirect jumps improve branch prediction since they cause fewer mispredictions than the set of branches they replaced.

Categories and Subject Descriptors: Programming Languages [**Compilers**]: Optimization—*Languages*

General Terms: Compiler, Code improving transformation, Measurements

Key Words: Indirect jump, branch prediction, branch target buffer

1. INTRODUCTION

Most high-level languages provide *multiway branch* statements to allow programmers to write more readable code. The characteristic feature of a multiway statement is the ability to select an action based on the value of a control expression. Without performing any optimization, a compiler would translate each case label of the multiway statement into a conditional branch. Because of the widespread usage of multiway statements, instruction sets commonly support an *indirect jump* from a table in order to reduce the cost of such sequences of conditional branches. As a result, compiler front-ends typically generate an indirect jump from a table as one translation alternative[1] for multiway statements [Sale 1981; Spuler 1994].

This traditional approach for using indirect jumps poses two problems. First, it is difficult to determine when the indirect jump can be effectively used in a machine-independent fashion since an accurate cost-benefit estimate can only be made after generating machine code. Second, many code-improving opportunities

[1]The other popular alternatives include *linear search*, *binary search*, and *hashing.*

suitable for the indirect jump may be missed when only considering this operation for the translation of a multiway statement.

This paper describes a general code-improving transformation that exploits indirect jumps after code generation. As the instruction issue rate and pipeline depth of processors increase, efficient handling of branches becomes more vital. Our improving transformation reduces the number of branches and mispredictions by coalescing several conditional branches into an indirect jump. First, dataflow analysis is performed to detect a set of possibly noncontiguous conditional branches that can be potentially coalesced into a single indirect jump. Second, control-flow analysis is used to determine how the control flow should be restructured to perform the coalescing. Third, analysis is accomplished to determine how to most efficiently generate the indirect jump operation. The cost of the original branches is also estimated and the indirect jump transformation is applied when deemed worthwhile. Finally, the original control flow is modified by duplicating basic blocks when necessary.

1.1 Motivation

Exploiting indirect jumps after code generation can be quite beneficial since additional branches from other control statements besides multiway statements can be coalesced into a single indirect jump. The examples in this section are given in C to more concisely depict branches that can be coalesced into indirect jumps. The control flow of the restructured C code segments would be comparable to a restructured flow graph of basic blocks with an indirect jump from a table.

I. Indirect Jumps with Branches (Figure 1)

Consider the *Original* code segment from *ctags* (C tag generator). A typical C compiler would translate the `switch` statement into an indirect jump from a table and would generate a conditional branch for the `for` statement. Yet, the conditional branch comparing `*sp` with zero would immediately precede the indirect jump. An optimizer could recognize this sequence of branches and be able to coalesce the extra conditional branch that compares the variable with zero into the indirect jump. Note that one can view this branch as another `case` for the `switch` statement as shown in the *Restructured* code segment.

Original

```
for (sp = line; *sp; sp++) {
   switch (*sp) {
   case ’p’:
            ...
   case ’k’:
            ...
   ...
  }
}
```

Restructured

```
for (sp = line;  ; sp++) {
   switch (*sp) {
   case ’\0’:
             goto out;
   case ’p’:
             ...
   case ’k’:
            ...
}
out:
```

Fig. 1. Code Fragment from UNIX utility *ctags*

II. Sequence of Contiguous Branches (Figure 2)

Other common instances may occur due to programming style. The *Original* code segment from *grep* (pattern search utility) shows a series of `if` statements comparing the same variable to different constants. A typical C compiler would translate these `if` statements as a sequence of conditional branches. However, the code could have been equivalently written as a single `switch` statement as shown in the *Restructured* code segment. An optimizer could detect the original sequence of conditional branches and could coalesce such contiguous branches into a single indirect jump. Use of multiple macros may also result in several consecutive comparisons being performed. Thus, branch coalescing is appealing since performance is less affected by program style (whether or not multiway branches are used).

Original

```
if ((c = *sp++) == 0)
    goto cerror;
if (c == ’<’) { ... }
if (c == ’>’) { ... }
if (c == ’(’) { ... }
if (c == ’)’) { ... }
if (c >= ’1’ && c <= ’9’) { ... }
...
```

Restructured

```
c = *sp++;
switch (c) {
case 0:          goto cerror;
case ’<’:          ...
case ’>’:          ...
case ’(’:          ...
case ’)’:          ...
case ’1’: case ’2’: case ’3’:
case ’4’: case ’5’: case ’6’:
case ’7’: case ’8’: case ’9’:
                   ...
default:           ...
}
```

Fig. 2. Code fragment from UNIX utility *grep*

III. Set of Contiguous and Noncontiguous Branches (Figure 3)

Often there are paths in which intervening instructions exist between branches that compare the same variable to constants and these intervening instructions do not update this variable. Consider the following *Original* code segment from *wc* (word count utility). A typical C compiler would translate each `if` statement into conditional branch(es). At first, it may appear that only the sequence of conditional branches shown in the shaded boxes can be coalesced into an indirect jump. However, the statement `charct++;` does not affect the branch variable `c`. An optimizer could determine the existence of path(s) between branches comparing the same variable to constants where the variable is unaffected. The optimizer could modify the original control flow by duplicating code to allow the branch for the `EOF` check to also be coalescent. As shown in the *Duplicated* code segment, all of the branches in the shaded boxes can be effectively considered as being contiguous and coalescent for a single indirect jump. The *Restructured* code segment shows equivalent code written with a `switch` statement.

Original

```
for (; ;) {
  c = getc(fp)
  if (c == EOF)
      break;
  charct++;
  if (' '<c&&c<0177) {
      if (!token) {
          wordct++;
          token++;
      }
      continue;
  }
  if (c=='\n')
      linect++;
  else if (c!=' '&&
           c!='\t')
      continue;
  token = 0;
}
```

Duplicated

```
for (; ;) {
  c = getc(fp)
  if (c == EOF)
      break;
  if (' '<c&&c<0177) {
      charct++;
      if (!token) {
          wordct++;
          token++;
      }
      continue;
  }
  if (c=='\n') {
      charct++;
      linect++;
  }
  else if (c!=' '&&
           c!='\t') {
      charct++;
      continue;
  }
  else
      charct++;
  token = 0;
}
```

Restructured

```
for (; ;) {
  c = getc(fp);
  switch (c) {
  case EOF:      goto out;
  case 041:
  ...
  case 0176:     charct++;
                 if (!token){
                     wordct++;
                     token++;
                 }
                 continue;
  case '\n':     charct++;
                 linect++;
                 goto end;
  default:       charct++;
                 continue;
  case ' ':
  case '\t':     charct++;
  }
  end:  token = 0;
}
out:
```

Fig. 3. Code fragment from UNIX utility *wc*

1.2 Organization of Paper

The paper is organized in the following manner. Section 2 gives a description of the compiler that has been used and modified to exploit indirect jumps after code generation. Section 3 briefly describes related compiler optimizations to reduce the cost of conditional branches. In order to detect and replace more branches into a single indirect jump than would be done in the traditional way, several detection and restructuring algorithms are introduced in Section 4 that can allow a compiler to detect a set of potentially noncontiguous coalescent conditional branches, which are often separated by blocks of intervening instructions, and to restructure the control flow by code duplication when necessary. Section 5 presents several techniques that reduce the cost of performing an indirect jump operation, often requiring the execution of only two instructions on a SPARC. The task of filling delay slots for indirect jumps is also dealt with in this section. Section 6 shows execution time results from performing dual loop tests [Clapp et al. 1986; Altman and Weiderman 1987] on SPARCstations to estimate the impact on pipeline stalls when the branch coalescing transformation was applied as another code improving transformation. Furthermore, the benefits of target buffer support for indirect jumps are discussed in this section. Various performance measurements are given in Section 7 that justify the validity of applying the code-improving transformation that is described in this paper. Section 8 suggests topics for future research. Finally, Section 9 concludes the paper.

2. MODIFICATIONS TO THE COMPILER

Figure 4 shows the overall structure of the *vpo* (Very Portable Optimizer [Benitez and Davidson 1988]) compiler system. The front-end of the compiler, *cfe* [Davidson and Whalley 1989], produces intermediate code from a given C preprocessed file. The *code expander* translates the intermediate code into unoptimized lists of machine-dependent effects, called *RTL*s (Register Transfer Lists). RTLs have the form of conventional expressions and assignments over the hardware's storage cells. For example, the RTLs

```
IC=r[8]?10;          ! SPARC assembly: cmp  %l0,10
PC=IC:0,L001;        ! SPARC assembly: be   L001
```

represent two machine instructions, where `IC` denotes condition code and `PC` denotes program counter, respectively.[2]

The first RTL depicts the effect of setting a condition code by comparing a register (`r[8]`) to constant `10`. The second RTL describes the effect of transfering the control to the address `L001` when `r[8]` is equal to `10`. While any particular RTL is machine specific, the general form of the RTL is machine-independent. This allows general machine-independent algorithms to be written that implement code improving transformations on machine-dependent code.

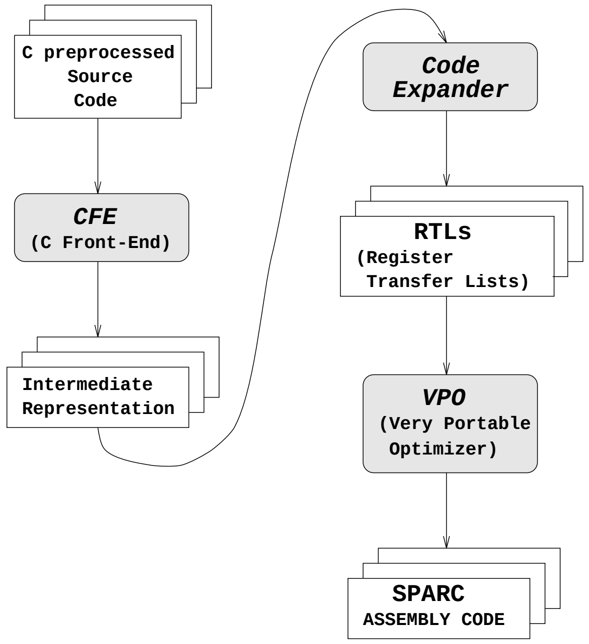

C preprocessed Source Code
CFE
(C Front-End)
Intermediate Representation
Code Expander
RTLs
(Register Transfer Lists)
VPO
(Very Portable Optimizer)
SPARC
ASSEMBLY CODE

Fig. 4. *VPCC* (Very Portable C Compiler)

[2]These instructions are generated by *cfe* when translating high level control statements, such as `if` or `if-then-else` statements.

All phases of the back-end of the compiler, *vpo* (Very Portable Optimizer), manipulate RTLs. The RTLs are stored in a data structure that also contains information about the order and control flow of the RTLs within a function. By manipulating RTLs as the sole intermediate representation, the following benefits can be achieved.

(1) Most optimizations can be invoked in any order and can be allowed to iterate until no further improvement can be found. Therefore, many phase ordering problems are eliminated.[3]
(2) The effect of a modification to the set of RTLs comprising a function is relatively simple to grasp.[4]

In order to exploit the indirect jump operation the following modifications were made to the compiler. The front-end of the compiler, *cfe*, was modified to always produce a linear sequence of conditional branches when translating a C `switch` statement instead of sometimes producing an indirect jump.[5]

An additional code-improving transformation phase to coalesce branches into an indirect jump from a table was added to the back-end of the compiler, *vpo*. After branches were coalesced into an indirect jump in a loop, loop invariant code motion was reinvoked to move the instructions that calculate the address of the jump table out of that loop before coalescing branches at an outer loop level. Most compiler optimizers perform these transformations starting with the innermost loops first to secure registers for the most frequently executed code segments. When an indirect jump occurs inside a loop, performing code motion on the loop-invariant instructions for calculating the jump table address requires a register. Thus, as depicted in Figure 5, the authors coalesced branches from the innermost loop outward after all other transformations for a given loop have been initially attempted. Afterwards, branch coalescing was also attempted on the outermost level of an entire function.

3. RELATED WORK

There has been some research on other techniques for avoiding conditional branches. Loop unrolling has been used to avoid executions of the conditional branch associated with a loop termination condition [Davidson and Jinturkar 1996]. Loop unswitching moves a conditional branch with a loop-invariant condition before the loop and duplicates the loop in each of the two destinations of the branch [Allen and Cocke 1971]. Conditional branches have also been avoided by code duplication

[3]In contrast, a more traditional compiler system will perform optimizations on various different representations. For instance, machine-independent transformations are often performed on intermediate code and machine-dependent transformations, such as peephole optimizations, are often performed on assembly code.
[4]In contrast, most traditional compiler systems generate code after optimizations. Thus, the optimizations are actually performed on intermediate code. Since there is typically not a one-to-one mapping between an intermediate code operation and a machine instruction, the effect of a modification on the final code that will be generated may not be obvious in these systems.
[5]*cfe* originally translates a C `switch` statement into one of the following three alternative forms:

(1) indirect jump using a jump table,
(2) binary search, or
(3) linear search.

```
Branch Chaining
Useless Jump Elimination
Dead Code Elimination
Eliminate Unconditional Jumps by Reordering Code
Instruction Selection
Evaluation Order Determination
Global Instruction Selection
Register Assignment
Jump Minimization
Instruction Selection

DO {
   Register Allocation
   Instruction Selection
   Common Subexpression Elimination
   Dead Variable Elimination
   Loop Optimizations:
      Code Motion
      Recurrences
      Loop Strength Reduction
      Induction Variable Elimination
      If (First Pass)
        Branch Coalescing
   Useless Jump Elimination
   Cheaper Instruction Replacement
   Instruction Selection
} While (change)

Branch Coalescing
Setup Entry and Exit
Instruction Scheduling
Fill Slots
Useless Jumps
```

Fig. 5. Modified *vpo*

[Mueller and Whalley 1995]. This method determines if there are paths where the result of a conditional branch will be known and duplicates code to avoid execution of the branch. The method of avoiding conditional branches using code duplication has been extended using interprocedural analysis [Bodik et al. 1997]. Finally, sequences of branches have been reordered using profile data to reduce the number of branches executed [Yang et al. 1998].

Our approach in this paper is similar to the above techniques in that it improves performance despite the penalty of increasing code size. However, there are often situations where several branches can be coalesced into a single indirect jump to avoid the execution of branches that these other techniques could not. Our approach essentially provides early resolution of branches that may originally have been some distance away in the control flow from the point where the indirect jump is inserted.

4. COALESCING A SET OF NONCONTIGUOUS CONDITIONAL BRANCHES

Some definitions are now presented before describing how a set of branches to be coalesced is found.

Definition 4.0.1. A *branch variable* is the register or variable associated with a conditional branch that is being compared to a constant.

Definition 4.0.2. A basic block will have an *effect* on a branch variable if the block has an instruction that updates the branch variable.

Definition 4.0.3. A conditional branch is considered *reachable* from a point in the given control flow if there exists a path from that point to the conditional branch with no effect on the branch variable.

Definition 4.0.4. A *related* set of branches for basic block `B` are those branches that are reachable from `B` and have the same branch variable as `B`.

The task for coalescing a set of conditional branches into an indirect jump is accomplished in the following manner. First, a set of coalescent conditional branches, which may or may not have intervening instructions, is identified. Second, a graph for the projected control flow is built to coalesce this set of conditional branches into an indirect jump. When the transformation is deemed beneficial, the original control flow is tranformed according to the graph by duplicating basic blocks when necessary.

4.1 Finding A Set of Branches to Coalesce

In order to find the largest set of coalescent branches, analysis is performed as follows. For each basic block `B`, the reachable branches from the exit point of `B` are determined. When `B` contains a conditional branch, the optimizer calculates the reachable branches that depend on the same branch variable as that of `B`. We denote such branches as being related and denote `B` as the *root* block of these branches. After detecting all sets of related branches, the optimizer selects the set with the largest number of branches. The largest set should be chosen first since branch coalescing requires the allocation of registers.

The desired reachability information is collected by calculating the following state information for each basic block `B`.

`in[B]`:	Set of blocks containing a reachable branch from the entry of `B`.
`out[B]`:	Set of blocks containing a reachable branch from the exit of `B`. This includes the conditional branch in `B`, if one exists.
`effect[B]`:	Set of blocks containing a branch instruction whose branch variable is updated by some instructions in `B`.

The above state information can be calculated with an iterative algorithm described in Figure 6. When the algorithm terminates, the `out[B]` of each basic block `B` contains the reachable branches from the exit point of `B`. This algorithm is guaranteed to terminate since, for any given control flow, (1) there exists a finite number of conditional branches, and (2) the `in[B]` and `out[B]` of each block `B` monotonically increase.

Applying the iterative algorithm described in Figure 6 to the example control flow in Figure 7 produces the dataflow information as indicated by Table I. Since block 1 has the largest set of related branches, the compiler will first attempt to coalesce these branches by placing instructions to perform an indirect jump at the root block 1. However, it is possible that related branch sets of two or more blocks have the same cardinality. In this case the optimizer will choose the block that dominates the most blocks having branches in the same related set.

```
DO
    FOR each basic block B DO
       /* Calculate out[B] from the successors of B */
       /* and its own branch                        */
       out[B] = NULL
       FOR each immediate successor block S of B DO
          out[B] = out[B] ∪ in[S]
       IF (B contains a branch instruction) THEN
          out[B] = out[B] ∪ {B}
       /* Calculate in[B] using out[B]    */
       in[B] = out[B] − effect[B]
WHILE (any changes)
```

Fig. 6. An Algorithm Calculating Reachable Branches For Each Basic Block

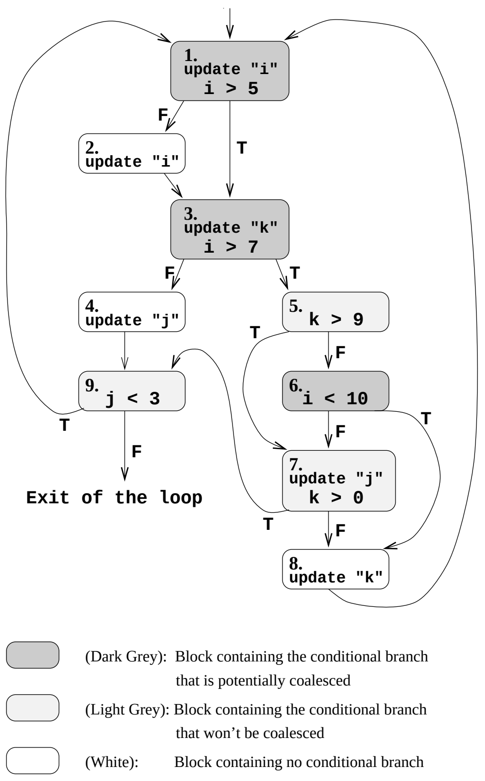

1. update "i" i > 5
F
T
2. update "i"
3. update "k" i > 7
F
T
4. update "j"
5. k > 9
T
F
6. i < 10
T
F
9. j < 3
T
F
Exit of the loop
7. update "j" k > 0
T
F
8. update "k"
(Dark Grey): Block containing the conditional branch that is potentially coalesced
(Light Grey): Block containing the conditional branch that won't be coalesced
(White): Block containing no conditional branch

Fig. 7. An Example Control Flow

Table I. Dataflow Information for the Example Control Flow

Block No.	`effect`	After The Algorithm		`Related` Branches
		`in`	`out`	
block 1	{1,3,6}	null	{1,3,6}	{1,3,6}
block 2	{1,3,6}	null	{3,6}	null
block 3	{5,7}	{3,6}	{3,5,6,7}	{3,6}
block 4	{9}	null	{9}	null
block 5	null	{5,6,7}	{5,6,7}	{5,7}
block 6	null	{6,7}	{6,7}	{6}
block 7	{9}	{7}	{7,9}	{7}
block 8	{5,7}	null	null	null
block 9	null	{9}	{9}	{9}

4.2 Projecting the Restructured Control Flow

Once a set of related branches has been selected, the optimizer projects the revised control flow to coalesce these branches into a single indirect jump. The restructured control flow is calculated by recording states in each block for these related branches. The state associated with each related branch is defined to be a set of triples, where each triple consists of the following three components,

(1) basic block number containing that branch variable,

(2) whether the conditional branch will be taken (T) or not taken (F), and

(3) the value range of the branch variable to satisfy the condition that is specified by the second component.

The projected control flow is calculated in the following manner. For a given set of related branches and its associated root block, the optimizer propagates the state (triples) of each related branch backward through the control flow (toward its root block). When the propagation completes, the optimizer determines the sequence of related branches that would be executed starting from the root block for each nonoverlapping value range of the branch variable. At this point, cost-benefit analysis is performed to determine whether or not coalescing the set of related branches into an indirect jump is worthwhile. If it is deemed beneficial, then a graph is incrementally built to project the desired restructuring at the root block. If the optimizer determines that there will be no significant code-size increase, then the graph will later be used to modify the actual control flow.

Table II. Initial States for Related Branches of Block 1

Related Branches	Initial States (Triples)
related branch in block 1	(1,T,[6..255]), (1,F,[0..5])
related branch in block 3	(3,T,[8..255]), (3,F,[0..7])
related branch in block 6	(6,T,[0..9]),(6,F,[10..255])

As an illustration, consider the example control flow in Figure 7 with one additional assumption that the branch variable `i` was detected to contain an unsigned character value [0..255]. For the set of related branches at the root block 1, Table II shows the initial states associated with these branches. In order to propagate the triples for branch 6 toward the root block 1, this state information should be

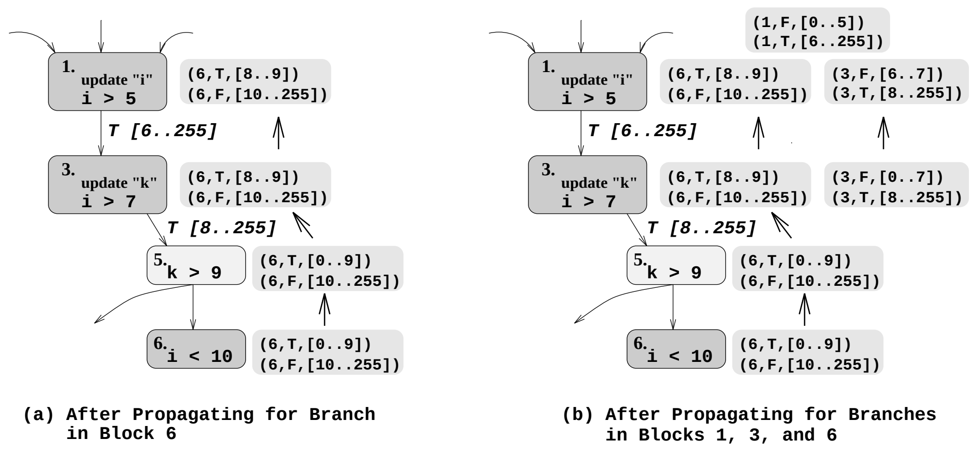

1. update "i" i > 5
T [6..255]
3. update "k" i > 7
T [8..255]
5. k > 9
6. i < 10
(6,T,[8..9]) (6,F,[10..255])
(6,T,[0..9]) (6,F,[10..255])
(1,F,[0..5]) (1,T,[6..255])
(3,F,[6..7]) (3,T,[8..255])
(3,F,[0..7]) (3,T,[8..255])
(a) After Propagating for Branch in Block 6
(b) After Propagating for Branches in Blocks 1, 3, and 6

Fig. 8. Propagating Triples toward Root Block 1

propagated through block 3. The transition from block 3 to block 6 can occur only when the value of branch variable i is in the range [8..255]. Similarly, the transition from block 1 to block 3 can occur only when the value of branch variable i is in the range [6..255]. Therefore, the value ranges of the triples for branch 6 should be properly adjusted during the propagation to reflect these two transitions. As shown in Figure 8(a), the value ranges of the triples for branch 6 are intersected with [8..255] at block 3, and the adjusted value ranges are intersected with [6..255] at the root block 1. Figure 8(b) illustrates how other triples, associated with branch 1 and 3, are propagated. Table III shows the final state information available at the root block 1 after the above value range propagation process completes.

Table III. States of Related Branches Associated with Nonoverlapping Value Ranges of i at Root Block 1

Value Range of i	States of Related Branches
[0..5]	1,F
[6..7]	1,T and 3,F
[8..9]	1,T and 3,T and 6,T
[10..255]	1,T and 3,T and 6,F

4.3 Cost-Benefit Analysis and Control-Flow Restructuring

Cost-benefit analysis is performed to determine at this point whether or not it is beneficial to coalesce the related branches of the root block into an indirect jump. The optimizer first checks if the values being compared are characters (represented in a byte). The optimizer weights the character values according to an estimated frequency of common use. For instance, values representing ASCII letters were assigned a higher weight than values representing control characters. The cost of executing the branches was calculated as a sum of products, where each product was obtained by multiplying the weights of the characters in each value range and the number of branches associated with that range. If the optimizer could not determine that the comparisons were with characters, then each value was given the same weight. The cost of executing the branches is compared to the cost of performing the indirect jump, which is described in the next section.

If the analysis determines that branch coalescing is worthwhile, then the restructuring algorithm shown in Figure 9 will produce a graph to efficiently represent the revised control flow to coalesce these related branches into an indirect jump at the root block. The central idea is that a new node will be added when no current node for that block exists with the same states for the related branches. The projected graph of the restructured control flow for Figure 7 is shown in Figure 10. The related branch in root block 1 will be replaced in the restructured code by instructions to perform an indirect jump. Note that a basic block represented with a dashed box indicates that the related branch is unnecessary and will not be placed in the restructured code.

5. EFFICIENTLY PERFORMING THE INDIRECT JUMP OPERATION

Compiler writers have long considered performing an indirect jump from a jump table as a very expensive operation. The tasks associated with performing an indirect jump includes the following:

(1) checking if the value being compared is within a bounded range,
(2) calculating the address of the jump table,
(3) calculating the offset used to index into the table,
(4) loading the target address from the table, and
(5) performing the indirect jump.

The number of instructions required to perform an indirect jump from a jump table can vary depending upon a number of factors. For the C `switch` statement shown in Figure 11(a), Figure 11(b) depicts SPARC instructions represented as RTLs that are used to implement a corresponding indirect jump (disregarding the instruction in the delay slot of the indirect jump).[6] Similar instructions are available on most RISC machines. It would appear that at least 5 pairs of conditional branches must be executed to make coalescing branches into an indirect jump operation worthwhile on the SPARC since 8 instructions are used to implement an indirect jump.

By statically analyzing the code surrounding an indirect jump operation, the optimizer can significantly reduce the cost of performing an indirect jump. Many optimizers can detect that instructions 4 and 5 in Figure 11(b) are loop invariant and therefore can move these instructions out of a loop. The authors implemented techniques that often avoid the execution of instructions 1-3 and 6 in Figure 11(b) as well.

5.1 Padding the Front of the Table

Instructions 1-3 in Figure 11(b) are used to check if the expression is in the range of possible case values. Instruction 1 can be avoided when the lowest case value is positive and relatively close to zero. The jump table can be padded with the addresses

[6]These SPARC instructions are generated by the *pcc* [Johnson 1979], *gcc* [Stallman 1990], and *vpcc* [Benitez and Davidson 1988] compilers.

```
PROCEDURE Build_Graph_From_Root(root_node,root_block)
{
   root_node = NewNode(NULL,root_block,NULL);
   FOR each non-overlapping value range VRANGE of
       the branch variable DO {
      current_states = related branch states associated with VRANGE;
      IF (current_states indicate related branch in root_block is taken)
         Build_Graph(root_node,root_block->taken,current_states,root_block);
      ELSE
         Build_Graph(root_node,root_block->not_taken,current_states,root_block);
   }
}



PROCEDURE Build_Graph(pred_node,successor_block,
                      current_states,root_block)
{
   /* Do not allow a cycle back to root_block */
   IF (successor_block == root_block)
     RETURN;
   /* Calculate new states */
   new_states = intersection between current_states and related
                branch states associated with successor block;
   IF (successor_block with new_states already
       exists in the graph) {
      Connect pred_node to the existing node;
      RETURN;
   }
   /* Create a new node for successor_block and
       append it to pred_node */
   new_node = NewNode(pred_node,successor_block,new_states);
   IF (successor_block contains related branch) {
      Mark new_node that the branch can be eliminated;
      IF (new_states indicate that successor of new node will be
          the branch target)
         Build_Graph(new_node,successor_block->taken, new_states,root_block);
      ELSE
         Build_Graph(new_node,successor_block->not_taken,new_states,root_block);
   }
   ELSE
      FOR each immediate successor block SUCC of successor_block DO {
         Build_Graph(new_node,SUCC,new_states,root_block);
      }
}
```

Fig. 9. Restructuring Algorithm

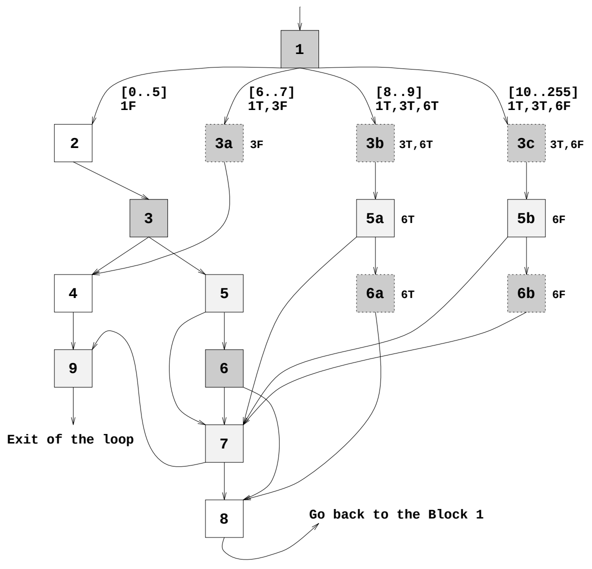


Fig. 10. Graph Representing Restructured Control Flow for Figure 2

(a)

```
switch (c) {
       case 'a':
          ...

       case 'b':
          ...

       case 'c':
          ...

       case 'd':
          ...

       case 'e':
          ...

       default:
          ...

     }
```

(b)

```
r[8]=r[8]-97;          # 1. Subtract the lowest case value
IC=r[8]?4;             # 2. Compare with (highest-lowest)
PC=ICh0,L27;           # 3. Perform unsigned > branch to
                          ensure the value is within range
                          (L27 is the default address)
r[20]=HI[L01];         # 4. Get High portion of address of
                          jump table
r[20]=r[20]|LO[L01];   # 5. Get Low portion of the address
r[8]=r[8]<<2;          # 6. Align value on a word boundary
                          so can index into jump table
r[8]=M[r[8]+r[20]];    # 7. Load target destination out of
                          jump table
PC=r[8];               # 8. Perform an indirect jump
L01:
.WORD  L22             # Target address for case 'a'
.WORD  L23             # Target address for case 'b'
.WORD  L24             # Target address for case 'c'
.WORD  L25             # Target address for case 'd'
.WORD  L26             # Target address for case 'e'
L27:
```

Fig. 11. RTLs to Perform an Indirect Jump from a Jump Table

corresponding to the default target. This technique is illustrated in Figure 12, which contains the instructions of Figure 11(b) with the modifications resulting from padding the front of the jump table. Instruction 2 in Figure 12 uses the highest case value in the comparison when padding is applied. Note also that instructions 4 and 5 in Figure 11(b) were removed in Figure 12 since it was assumed they are loop invariant for this example.

```
IC=r[8]?103;               # 2. Compare with (highest-lowest)
PC=ICh0,L27;               # 3. Perform unsigned > branch to
                                ensure the value is within range
                                (L27 is the default address)
r[8]=r[8]<<2;              # 6. Align value on a word boundary
                                so can index into jump table
r[8]=M[r[8]+r[20]];        # 7. Load target destination out of
                                jump table
PC=r[8];                   # 8. Perform an indirect jump
L01:
.word  L27                 # Target Address for 0
.word  L27                 # Target Address for 1
...                          ...
.word  L27                 # Target Address for 96 ('a'-1)
.word  L22                 # Target Address for 'a'
.word  L23                 # Target Address for 'b'
.word  L24                 # Target Address for 'c'
.word  L25                 # Target Address for 'd'
.word  L26                 # Target Address for 'e'
L27:
```

Fig. 12. RTLs after Padding the Front of the Table

5.2 Using Value-Range Analysis to Avoid the Initial Range Check

The initial range check (instructions 1-3 in Figure 11(b)) can be completely avoided if a bounded range of case values is known and an entry can be stored in the table for each value [Spuler 1994]. Assume that the value range of the variable `c` in Figure 11(a) is [0..255]. The indirect jump operation associated with the known value range of the branch variable is depicted in Figure 13.[7]

Once a set of related branches has been selected, the optimizer *vpo* uses demand-driven analysis to recursively search all the possible paths backward from the root block to determine if the range of case values is bounded. In the following subsections, a general algorithm for such range determination is depicted, and several cases that can be handled by the algorithm are illustrated.

[7]Note that 256 targets are listed in the table. Often this space is reduced by a factor of four as described in the next section.

```
r[8]=r[8]<<2;              # 6. Align value on a word boundary
                               so can index into jump table
r[8]=M[r[8]+r[20]];        # 7. Load target destination out of
                               jump table
PC=r[8];                   # 8. Perform an indirect jump
L01:
.word  L27                 # Target Address for 0
.word  L27                 # Target Address for 1
...                          ...
.word  L27                 # Target Address for 96 ('a'-1)
.word  L22                 # Target Address for 'a'
.word  L23                 # Target Address for 'b'
.word  L24                 # Target Address for 'c'
.word  L25                 # Target Address for 'd'
.word  L26                 # Target Address for 'e'
.word  L27                 # Target Address for 102 ('e'+1)
.word  L27                 # Target Address for 103
...                          ...
.word  L27                 # Target Address for 255
L27:
```

Fig. 13. SPARC Instructions with a Bounded Range of Values

5.2.1 General Algorithm to Determine Bounded Value Ranges:

A general algorithm for determining if the range of case values is bounded is shown in Figure 14. The essence of this algorithm is as follows.

(1) Expand a branch variable using previous effects on the variable by recursively searching all the possible paths backward from the root block.

(2) Whenever an expansion occurs, parse and evaluate the expanded expression to determine whether or not the range of case values can be determined.

The algorithm returns a state with a detected range of case values if one of the following conditions exists.

—`bounded`: The value ranges of a branch variable can be enumerated in a jump table.

—`unbounded`: The value ranges of the branch variable cannot be enumerated in a jump table.

—`duplicated`: The value ranges of the branch variable can be enumerated in a certain execution path. This state provides an extra opportunity for the optimizer to perform an indirect jump more efficiently in the *bounded* execution path by duplicating some blocks of instructions.

5.3 Analyzing Effects

For a given *root* block, a bounded value range of the branch variable can often be determined by examining each effect backward from the root. Consider the C code depicted in the left column of Figure 15 with an assumption that the block

```
PROCEDURE Bounded_Path(RTL_pointer, Register)
{
  current_block = basic block containing RTL_pointer.
  expanded_expr = Register.
  Value_Range_State =  None.
  Set_of_value_range = NULL.
  Set_of_duplicated_block = NULL.
  /* Expand and evaluate expanded_expr within current_block
     If the expanded_expr is determined to be BOUNDED, add
     the bounded value range to Set_of_value_range and return */
  WHILE (RTL_pointer = previous_rtl(RTL_pointer)) {
      Expand_and_Evaluate(RTL_pointer, expanded_expr,
                          Set_of_value_range,
                          Value_Range_State).
      IF (expanded_expr is either BOUNDED or UNBOUNDED)
         RETURN Value_Range_State.
      /* Alias effect such as r[9]=r[8] */
      ELSE IF (expanded_expr == Register &&
               RTL_pointer points to the instruction
               that assigns Register to New_Register)
         expanded_expr = Register = New_Register.
  }
  /* Neither BOUNDED nor UNBOUNDED state can be determined by
     evaluating expanded_expr. Expand and evaluate the expression
     by recursively looking back all predecessor blocks */
  FOR each predecessor block of current_block DO {
     temp_expr = expanded_expr.
     Recursively expand and evaluate temp_expr starting from
         the predecessor until temp_expr is determined to be
         either BOUNDED or UNBOUNDED.
     IF (temp_expr is determined to be BOUNDED)
        Add the associated value range to Set_of_value_range.
  }
  IF (Value_Range_State is both BOUNDED and UNBOUNDED &&
      there exists a single execution path along which
      the value range is BOUNDED) {
      Calculate Set_of_duplicated_block by taking intersection
        among sets of blocks along all possible execution
        paths to the root block.
      RETURN REPLICATED.
  }
  ELSE IF (Value_Range_State == BOUNDED)
      RETURN BOUNDED.
  ELSE
      RETURN UNBOUNDED.
}
```

Fig. 14. Detection Algorithm for Bounded Value Ranges

containing the condition "`c == a`" has been selected as the *root*. The bounded value range of the condition variable `c` was detected by expanding register `r[8]`, which contains the temporary value of `c`, with previous effects on that register. The right column of Figure 15 depicts the RTLs when the branch coalescing analysis was about to be performed. The expansion of `r[8]` was accomplished as follow.

```
1. r[8]             # register containing values of 'c'
2. r[8]}24          # instruction 2: right shift('}') by 24 bits
3. (b[16]{24)}24    # instruction 1: left shift('{') by 24 bits
```

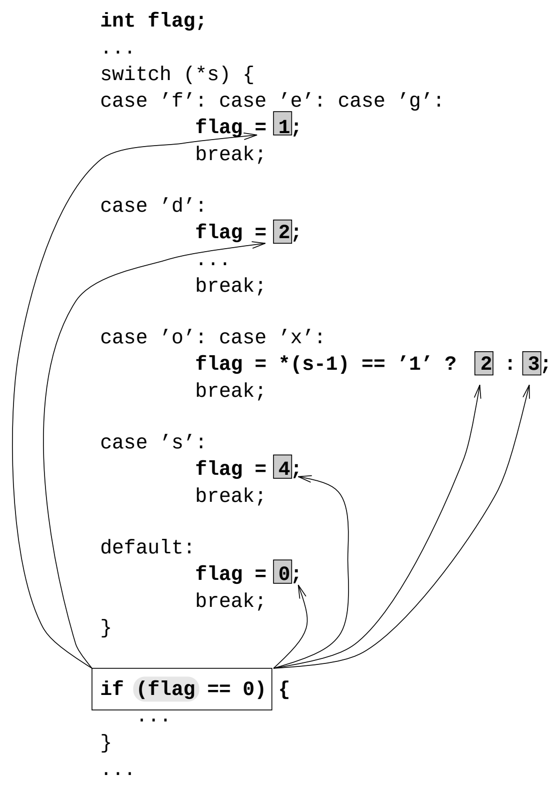
After the above expansion, the value range of `r[8]` was determined as bounded to the interval [-128..127], since the resulted effect from 24 bit left-shift followed by 24 bit right-shift is to mask the signed 8 bit value from `r[8]`. In a similar manner, the value range of a branch variable is determined as bounded to [0..255] when the variable can be expanded as the effect of unsigned byte load or conversion to an unsigned character value. Some other useful bounds were obtainable from the C mask operation, `'&'`.

Example C source	**RTLs**
	`r[8]=b[16]{24;        # 1. sll  %l0,24,%o0`
	`r[8]=r[8]}24;         # 2. sra  %o0,24,%o0`
	`...`
	# Block for c == 'a'
	`...`
	`IC=r[8]?97;           # 3. cmp  %o0,97`
	`PC=IC!0,L18;          # 4. bne  L18`
`char c;`	
`...`	*# Block for A()*
	`...`
`if (c == 'a')`	
`   A();`	*# Block for c == 'b'*
`else if (c == 'b')`	`L18:`
`   B();`	`IC=r[8]?98;           # 5. cmp  %o0,98`
`else if (c == 'c')`	`PC=IC!0,L22;          # 6. bne  L22`
`   C();`	
`...`	*# Block for B()*
	`...`
	# Block for c == 'c'
	`L22:`
	`IC=r[8]?99;           # 7. cmp  %o0,99`
	`PC=IC!0,L25;          # 8. bne  L25`
	# Block for C()
	`...`

Fig. 15. Example Case for Bounding Value Range

5.4 Analyzing Effects for All Possible Paths

For a given *root* block, a bounded value range of the branch variable was often determined by recursively searching all the possible paths backward from the *root*. Consider the C code segment shown in Figure 16 with an assumption that the block containing the condition "`flag == 0`" has been selected as the *root*. The value range of the variable `flag` was determined by recursively searching all the possible paths backward from the *root* block. The optimizer determines that the value of `flag` is bounded by the interval [0..4], since the value of `flag` is set to a certain constant in that interval for every possible path reaching the *root*.



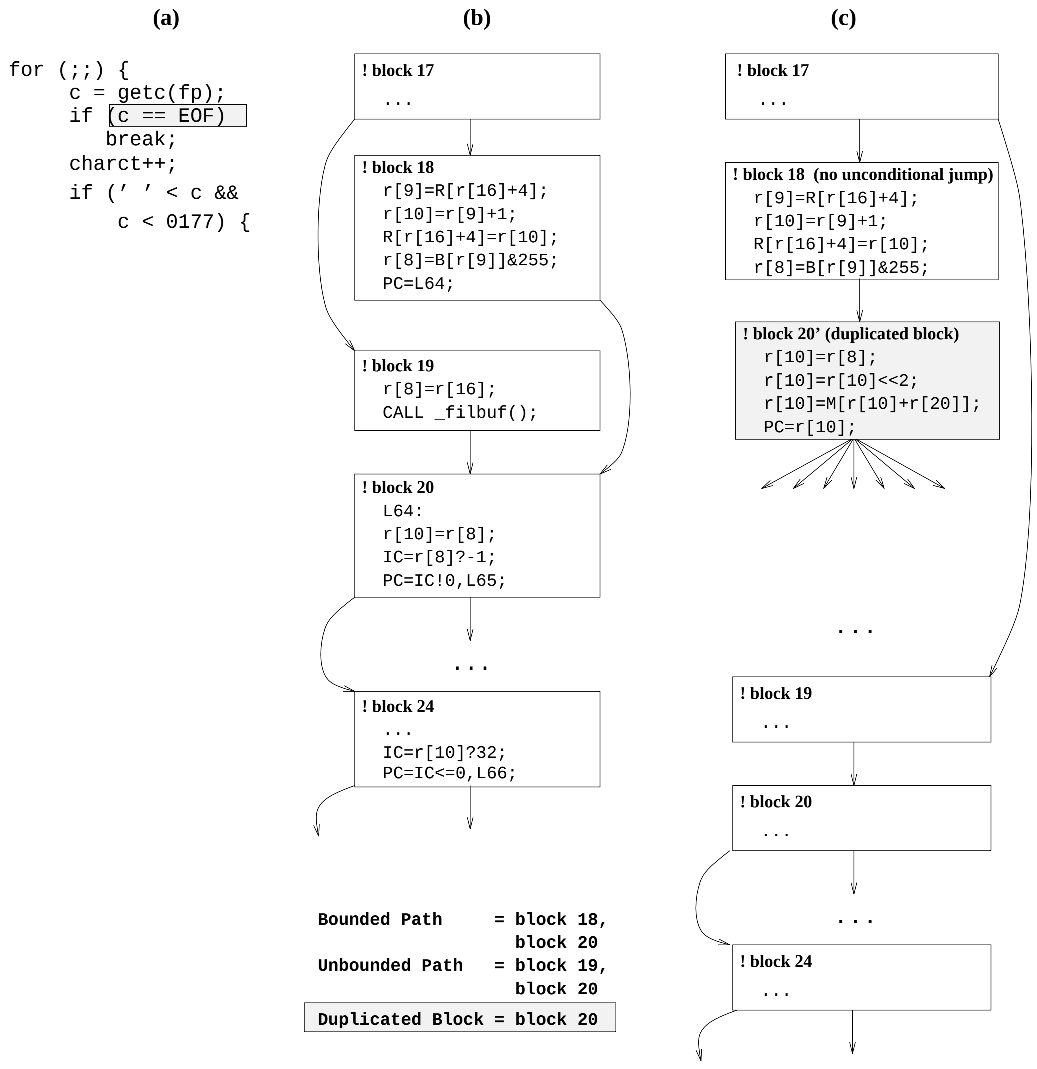
```
int flag;
...
switch (*s) {
case 'f': case 'e': case 'g':
        flag = 1;
        break;

case 'd':
        flag = 2;
        ...
        break;

case 'o': case 'x':
        flag = *(s-1) == '1' ?  2 : 3;
        break;

case 's':
        flag = 4;
        break;

default:
        flag = 0;
        break;
}

if (flag == 0) {
    ...
}
...
```

Fig. 16. Code Segment from *format()* in *awk*

5.5 Code Duplication for Bounded Value Range Path

Often a path of blocks is detected where the range of values is bounded and one or more paths are detected where the range is unbounded. Code is duplicated when deemed worthwhile to allow coalescing of branches to occur on the path with the bounded range. For example, Figure 17(a) shows a C code segment in *wc*, and the effects of the C statements in the shaded area are represented as RTLs with the control flow in Figure 17(b). The reaching algorithm in Figure 6 determined block 20 as the most beneficial *root* block. Note that the conditional branches in block 20 and block 24 were considered to be *related* since `r[10]` is an alias of `r[8]` by the RTL `r[10]=r[8]`.

Blocks 17 to 19 contain RTLs generated from invoking the `getc()` macro. Block 18 contains an RTL (`r[8]=B[r[9]]&255;`) that loads an unsigned character from a buffer and bounds the range of values from 0..255. Block 19 contains a call to `_filbuf`, which results in the value associated with `r[10]` being unbounded since no interprocedural analysis was performed. The optimizer recursively searches backwards and finds that blocks 20 and 18 are within a path back to the point where the range of values is bounded. Likewise, the compiler finds that blocks 20 and 19 are within a path where the range of values is unbounded. The intersection between the blocks in a bounded path and the blocks within any unbounded paths results in the block(s) that must be duplicated to distinguish the bounded path. Figures 17(c) shows the RTLs with the modified control flow after duplication of the block 20 and coalescing of the set of *related* branches. Coalescing can occur at the duplicated *root* (block 20') without an initial range check since the range of values is now bounded. Limits were placed on the amount of code allowed to be duplicated to prevent large code size increases.



(a)
```
for (;;) {
     c = getc(fp);
     if (c == EOF)
        break;
     charct++;
     if (' ' < c &&
         c < 0177) {
```

(b)
```
! block 17
  ...

! block 18
  r[9]=R[r[16]+4];
  r[10]=r[9]+1;
  R[r[16]+4]=r[10];
  r[8]=B[r[9]]&255;
  PC=L64;

! block 19
  r[8]=r[16];
  CALL _filbuf();

! block 20
  L64:
  r[10]=r[8];
  IC=r[8]?-1;
  PC=IC!0,L65;

...

! block 24
  ...
  IC=r[10]?32;
  PC=IC<=0,L66;
```

```
Bounded Path     = block 18,
                   block 20
Unbounded Path   = block 19,
                   block 20
Duplicated Block = block 20
```

(c)
```
! block 17
  ...

! block 18  (no unconditional jump)
  r[9]=R[r[16]+4];
  r[10]=r[9]+1;
  R[r[16]+4]=r[10];
  r[8]=B[r[9]]&255;

! block 20' (duplicated block)
  r[10]=r[8];
  r[10]=r[10]<<2;
  r[10]=M[r[10]+r[20]];
  PC=r[10];

...

! block 19
  ...

! block 20
  ...

...

! block 24
  ...
```

Fig. 17. Using Duplication to Distinguish Paths for Coalescing

5.6 Efficiently Indexing into the Jump Table

Instruction 6 in Figure 13 left shifts the value by 2 since each element of the jump table contains a complete target address requiring 4 bytes. Consider tables containing byte displacements instead of complete word addresses. For instance, Figure 18 shows how the code in Figure 13 can be transformed to use byte displacements. There are two advantages for using byte displacements. First, the left shift will no longer be necessary. Second, the table only requires one fourth the amount of space. Thus, a jump table for a value range associated with a character can be compressed from 256 to 64 words.

```
                    # r[20] is the jump table address (L01)
                    # r[22] is the base address (L02) for the displacement

r[8]=M[r[8]+r[20]];         # 7. Load target destination out of jump table
PC=r[8]+r[22];              # 8. Perform an indirect jump
.seg  ``data''
L01:
.byte  L27-L02              # Target Address for 0
.byte  L27-L02              # Target Address for 1
...                            ...
.byte  L27-L02              # Target Address for 96 ('a'-1)
.byte  L22-L02              # Target Address for 'a'
.byte  L23-L02              # Target Address for 'b'
.byte  L24-L02              # Target Address for 'c'
.byte  L25-L02              # Target Address for 'd'
.byte  L26-L02              # Target Address for 'e'
.byte  L27-L02              # Target Address for 102 ('e'+1)
.byte  L27-L02              # Target Address for 103
...                            ...
.byte  L27-L02              # Target Address for 255
.align 4
.seg  ``text''
L27:
```

Fig. 18. SPARC Instructions with Byte Displacements in the Jump Table

The disadvantages include requiring an additional register to calculate the base address for the displacements and not always having displacements small enough to fit within a byte. There are two approaches that were used to help ensure that the displacements are not too large. First, a label for the base of the displacements was placed at the instruction that was the midpoint between the first and last indirect jump targets. The jump table is always placed in the data segment so it will not cause the distance between indirect jump targets to be increased. Note this requires the calculation of the addresses of two labels (the one at the beginning of the jump table and the one used for the base address of the displacements). Before applying this approach, the compiler first ensures that the indirect jump would be in a loop and registers are available to move the calculation of both addresses out of the loop.

Second, the targets of the indirect jump may be moved to reduce the distance between targets. The instructions within a program may be divided into relocatable segments. Each segment starts with a basic block that is not fallen into from another block and ends with a block containing an unconditional transfer of control. An example of relocatable code segments is given in Figure 19. Assume each of the labels in the figure are potential targets of one indirect jump. There are three ways segments can be moved to reduce the distance between targets.

(1) A segment that does not contain any targets for a specific indirect jump can be moved when it is between segments containing such targets. For example, segment **D** can be moved to follow segment **A** since both segments contain no targets for the indirect jump.

(2) The segment containing the most instructions preceding the first target label in a segment can be moved so it will be the first segment containing targets. For example, segment **C** has blocks of instructions preceding the block containing its first target label (**L2**). By moving segment **C** to follow segment **D**, these instructions preceding **L2** will be outside the indirect jump target range.

(3) Likewise, the segment containing the most instructions following the last target label in its own segment can be moved so it will be the last positional segment containing targets. For example, segment **B** has the most instructions following its last target label (**L1**) and is moved to follow segment **E**. Jump tables are only converted to tables containing byte displacements when all targets of the indirect jump will be within the range of a byte displacement after relocating segments of code.

5.7 Filling Delay Slots for Indirect Jumps

The optimizer *vpo* previous to this work only filled delay slots of indirect jumps with instructions that precede the jump. This approach was reasonable since indirect jumps with tables occurred infrequently and filling the delay slot from one of several targets is more complicated than filling the delay slot of a branch instruction. After implementing the transformation to coalesce branches, indirect jumps occurred much more frequently. The compiler has been modified to fill the delay slot of an indirect jump with an instruction from one of the targets if it could not be filled with an instruction that preceded the jump. An instruction from a target block could only be used to fill the delay slot if it did not affect any of the live variables or registers entering any of the other target blocks.

Filling a slot for an indirect jump is less advantageous than that for a conditional branch (or unconditional jump) since more targets are associated with an indirect jump. Therefore, the optimizer *vpo* tried the following method to usefully fill slots for indirect jumps. Since each target of an indirect jump has been associated with certain range(s) of case values, the probability of the transition from an indirect jump to a certain target can be statically estimated. The optimizer *vpo* ranks the indirect jump targets based upon such estimation, and attempts to fill its slot with the instruction from the most probable target.

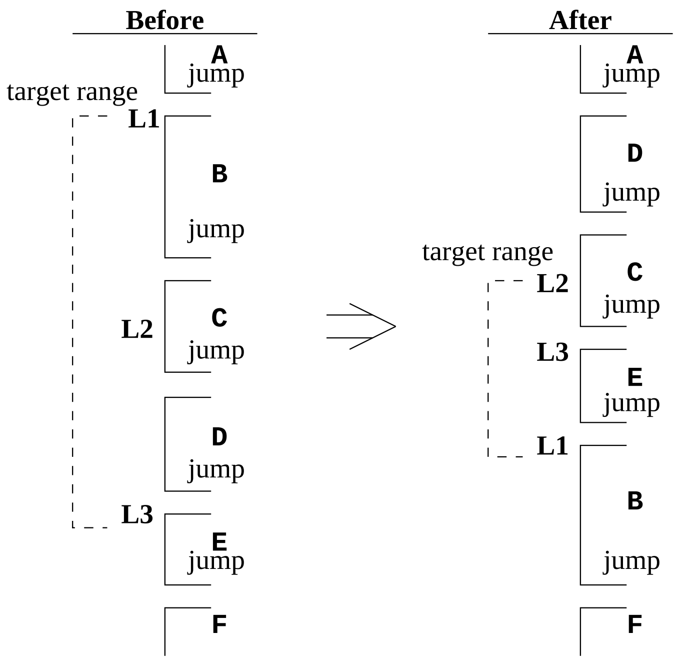


Fig. 19. Relocating Segments of Code

When a set of branches that are originally separated by some intervening instructions is selected for branch coalescing, the actual transformation is accomplished by duplicating these intervening instructions. In such a case, the usefulness of filling slots for indirect jumps can be significantly improved. For example, consider the following C code in Figure 20(a). The detection and restructuring algorithms in Chapter 4 allow the optimizer to detect all the branches in the shaded area as coalescent and transform these branches into an indirect jump by duplicating the effects of `wd++` to several destinations of the coalescent branches. The restructured code in Figure 20(b) shows the comparable C code after the transformation.

After coalescing with code duplication, most of targets of the indirect jump have identical effects of `wd++` as depicted in Figure 20(b). Thus, the code duplication from branch coalescing potentially provides extra opportunities to fill the delay slot of the indirect jump with a useful instruction. However, there is one more complication that should be resolved for successfully filling an instruction of `wd++` for the delay slot of the indirect jump. The RTLs shown in Figure 21(a) depict the restructured code after branch coalescing transformation occurs for the example C code. Note that hardware register `r[24]` contains the temporary value of `wd`. It appears that `r[24]=r[24]+1` cannot be filled for the delay slot, since `r[24]` is both set and referenced among targets of the indirect jump. However, `r[24]=r[24]+1` can be filled for the following reasons:

—`r[24]=r[24]+1` has no dependency with other instructions within the indirect jump target block containing the same instruction.

—`r[24]=r[24]+1` has no set-and-reference conflict when the analysis is performed

(a) Original Code

```
while (*wd)
  switch (*wd++) {
  case 'l':
          ipr(linect);
          break;
  case 'w':
          ipr(wordct);
          break;
  case 'c':
          ipr(charct);
          break;
  }
```

(b) Restructured Code after Duplicating "wd++"

```
again:
switch (*wd) {
case '\0':
         break;
case 'l':
        wd++;
        ipr(linect);
        goto again;
case 'w':
        wd++;
        ipr(wordct);
        goto again;
case 'c':
        wd++;
        ipr(charct);
        goto again;
default:
        wd++;
        goto again;
}
```

Fig. 20. Code Segment from *wcp()* in *wc*

by considering the targets containing that instruction as one conceptual target.

In order to fill such an indirect jump delay slot as described in the situation above, the following extra steps were added to *vpo*.

(1) For each target of the indirect jump, evaluate the probability that the target may be taken using the associated case values in the jump table. When the range of case values is bound to values representing ASCII letters, the probability is further weighted using estimated character frequency distribution of common use.

(2) Sort the indirect jump targets based on the evaluated probabilities.

(3) Starting from the most probable jump target to the least, make a list of all the instructions that can be potentially filled for the indirect jump without considering effects from other jump targets. Whenever an identical instruction is found in an other target block, add the associated probability to that of the instruction.
   (a) If an instruction does not exist in the list, then insert the RTL with its associated block address and probability.
   (b) else (the same RTL already found on other target block), add the associated probability to that of the existing RTL in the list and add the associated block address to the block address list of the existing RTL.

(4) Starting from the most probable instruction, determine if this instruction sets any variables or registers that could be live when entering any of the target blocks that do not have this instruction. If there is no conflict, then fill the delay slot with this instruction and delete it from the appropriate target blocks.

(a) Before filling delay slot for indirect jump

```
r[8]=(B[r[24]]{24)}24;
r[9]=(B[r8]+r[20]]{24)}24;
PC=r[9]+r[21];
delay slot
.seg "data"
.byte L0019-L0017
...
L0020:
.byte L82-L0017
...
.byte L0016-L0017
...
.byte L0017-L0017
...
.byte L0018-L0017
...
.align 4
.seg "text"

L0019:
r[24]=r[24]+1;
PC=L83;

L0018:
r[24]=r[24]+1;
r[8]=r[26];

L0017:
r[24]=r[24]+1;
PC=L87;

L0016:
r[24]=r[24]+1;
r[8]=r[25];

L82:
PC=RT;
NL=RS[];
```

(b) After filling delay slot for indirect jump

```
r[8]=(B[r[24]]{24)}24;
r[9]=(B[r8]+r[20]]{24)}24;
PC=r[9]+r[21];
r[24]=r[24]+1;
.seg "data"
.byte L0019-L0017
...
L0020:
.byte L82-L0017
...
.byte L0016-L0017
...
.byte L0017-L0017
...
.byte L0018-L0017
...
.align 4
.seg "text"

L0019:
! filled for the indirect jump
PC=L83;

L0018:
! filled for the indirect jump
r[8]=r[26];

L0017:
! filled for the indirect jump
PC=L87;

L0016:
! filled for the indirect jump
r[8]=r[25];

L82:
PC=RT;
NL=RS[];
```

Fig. 21. RTLs after Filling Delay Slot of the Indirect Jump for Example C Code in Figure 6.10(a)

6. OTHER ARCHITECTURAL ISSUES FOR COALESCING BRANCHES

The cost of performing an indirect jump from a jump table can vary on different machines. Not only can the number of instructions required to perform this operation vary, but indirect jump instructions (as well as conditional branches) can also result in pipeline stalls on many machines.

6.1 Dual Loop Test

To realistically estimate the pipeline impact on RISC architectures from replacing several conditional branches into an indirect jump, a dual loop test [Clapp et al. 1986; Altman and Weiderman 1987] has been conducted on a SPARCstation-IPC, SPARCstation-5, SPARCstation-20, and UltraSPARCstation-1.

—First, an optimized executable with a linear sequence of branches and with an indirect jump from a table, were generated for the C code shown in Figure 23. Let E_{linear} and $E_{indirect}$ denote such executables, respectively. Note that E_{linear} requires the execution of 2.5 branches on average for each loop iteration. Note that $E_{indirect}$ has been generated such that all the conditions in the loop body have been coalesced into an indirect jump operation requiring only two SPARC instructions as shown in Figure 18.

—Second, the authors ran each executable 20 times, and chose the shortest execution time for each executable. Let $\tau_{E_{loop}}$, $\tau_{E_{branches}}$, and $\tau_{E_{indirect}}$ represent such shortest execution times respectively. ($\tau_{E_{linear}}$ - $\tau_{E_{loop}}$) gives a relative estimate of the total time required to execute the the conditional branches over all iterations. ($\tau_{E_{indirect}}$ - $\tau_{E_{loop}}$) gives a relative estimate of the time that is required to perform an indirect jump operation as shown in Figure 18, over all iterations.

—Finally, by varying the number of conditions in the loop, the relative impact of conditional branches versus an indirect jump has been measured as shown in Table IV.

Table IV. Dual-Loop Test (10,000,000 iterations)

Machine Type	*Loop Cost*	*Linear Search*			*Indirect Jump*		
		2.5 br	4.5 br	8.5 br	2.5 br	4.5 br	8.5 br
SPARCstation-IPC	3.65s	3.82s	5.53s	8.82s	2.61s	2.71s	2.76s
SPARCstation-5	0.88s	1.03s	1.65s	2.74s	0.63s	0.76s	0.76s
SPARCstation-20	0.51s	0.93s	1.60s	2.65s	0.87s	0.93s	0.93s
UltraSPARC-1	0.40s	0.50s	1.16s	1.56	1.50s	1.51s	1.51s

From the dual loop test as described above, the authors found that an indirect jump as depicted in Figure 11(c) required about the same execution time as two pairs of compare and branch instructions for most SPARCstations except the UltraSPARC-1. Therefore, the indirect jump transformation is only applied when it is estimated that more than two coalescent branches in the set will on average be executed. For the UltraSPARC-1, an indirect jump as depicted in Figure 11 required about the same execution time as eight pairs of compare and branch instructions. The major reason is that the UltraSPARC-1 (a Superscalar architecture) provides the hardware branch target/prediction buffer support for branches, but no

```
int i;
main()
{
   long int j, k, l;
   struct timeval before,after;

   gettimeofday(&before, (struct timezone *)NULL);
   k = 0;
   l = 0;
   for (j=0; j<10000000; j++) {
      i = j & 3;
   }
   gettimeofday(&after, (struct timezone *)NULL);
   after.tv_sec -= before.tv_sec;
   after.tv_usec -= before.tv_usec;
   if (after.tv_usec < 0)
      after.tv_usec--, after.tv_usec += 1000000;
   ...
   ...
   printf(``The elapsed time: %9ld.%02ld\n'',
            after.tv_sec, after.tv_usec/10000);
}
```

Fig. 22. Code to Measure the Execution Time for Loop Overhead

hardware support for indirect jumps. In the following section, the authors argue that, with a comparable hardware branch prediction/target buffer support, such unbalanced execution time discrepency can be eliminated.

6.2 Branch Target Buffer (BTB) Support for Branches and Indirect Jumps

One characteristic feature of RISC machines is pipelining. Pipelining divides the execution of each instruction into several stages. Different stages can be overlapped in execution to increase processor throughput. However, there are several obstacles that limit the full exploitation of pipelining. One of the most serious obstacles is branch instructions. If the current instruction turns out to be a branch, then the CPU should predict in advance whether or not the branch is taken and what the target address will be in order to preserve a steady flow through the pipeline. However, the execution path of a branch cannot be easily resolved in advance. Thus, branches typically cause delays in the pipeline [Wall 1991; Perleberg and Smith 1993; Chang et al. 1997; Hennessy and Patterson 1996].

A Branch Target Buffer (BTB) can reduce these pipeline disruptions by predicting the path of the branch and caching information used by the branch. Various pieces of information can be kept in the BTB, including tags associated with the branch address, the branch target address, and branch prediction information [Perleberg and Smith 1993]. However, it has been reported that BTB-based prediction schemes perform poorly for indirect jumps, since the target of an indirect jump can change with every dynamic instance of that branch [Chang et al. 1997; Wall 1991].

```
 ...
 gettimeofday(&before, (struct timezone *)NULL);
 k = 0;
 l = 0;
 for (j=0; j<10000000; j++) {
    i = j & 3;

    /* 2.5 DYNAMIC NUMBER OF BRANCHES */
    if (i == 0) {
       k = k + 4;
       l = 4;
    }
    else if (i == 1) {
       k = k + 1;
       l = 1;
    }
    else if (i == 2) {
       k = k + 2;
       l = 2;
    }
    else if (i == 3) {
       k = k - 3;
       l = 3;
    }
}
gettimeofday(&after, (struct timezone *)NULL);
...
printf(``The elapsed time: %9ld.%02ld\n'',
         after.tv_sec, after.tv_usec/10000);
}
```

Fig. 23. Code to Measuring the Execution Time for Loop Overhead and Loop Body

In fact, some compilers provide techniques that insert extra conditional branches that check for likely targets to avoid the execution of indirect jumps from a table [Holler 1996] or indirect calls [Calder and Grunwald 1994].

Most modern architectures seldom support indirect jumps with a BTB due to poor misprediction ratios for indirect jumps. However, consider the results shown in Table IV. An UltraSPARC-1 could execute about eight pairs of compare and branch instructions in the time required to perform an indirect jump operation. One reason for the lower relative performance for indirect jumps on the UltraSPARC-1 was that this machine uses a BTB to provide architectural support for branches. There was no target buffer support on the UltraSPARC-1 for indirect jumps, which resulted in all indirect jumps being treated as mispredictions.

In the following sections, the authors claim that, with comparable BTB support for indirect jumps, the branch coalescing transformation can be beneficial in reducing the total number of dynamic branch mispredictions.[8] First, a conceptual design of BTBs is proposed that can provide comparable target buffer support for indirect jumps. Second, various branch prediction approaches will be described. Using more sophisticated branch prediction approaches as well as increasing the number of entries in BTBs is known to improve BTB performance [Perleberg and Smith 1993]. Third, issues will be presented about how to manage BTBs that support branches and indirect jumps. Finally, with comparable BTB support for

[8]We are not aware of any machines that use this exact model of BTB support for indirect jumps.

indirect jumps, the authors will provide arguments describing why the total number of branch mispredictions can be reduced by the branch coalescing transformation. In addition, another compiler technique will be introduced that can potentially reduce the number of dynamic indirect jump mispredictions.

6.2.1 A Conceptual BTB Supporting Branches and Indirect Jumps:

Branch Target buffers are available to reduce the cost of indirect jumps on some machines. These buffers are typically specialized to support indirect jumps generated from *return* statements since indirect jumps from tables are not generated frequently by most compilers [Hennessy and Patterson 1996](see page 276). However, BTBs can be easily extended to support indirect jumps from tables by considering an indirect jump as another PC-relative branch instruction [Hennessy and Patterson 1996](see page 274). For instance, Figure 24 shows one conceptual view of a BTB, which, like a cache, can have several alternative designs. If the appropriate tag is not found in the buffer, then the hardware predicts that the branch will not be taken. If the appropriate tag is found in the buffer and a branch predictor indicates the branch as taken, then the hardware predicts that the branch will be taken. Otherwise, the branch is predicted as not taken.

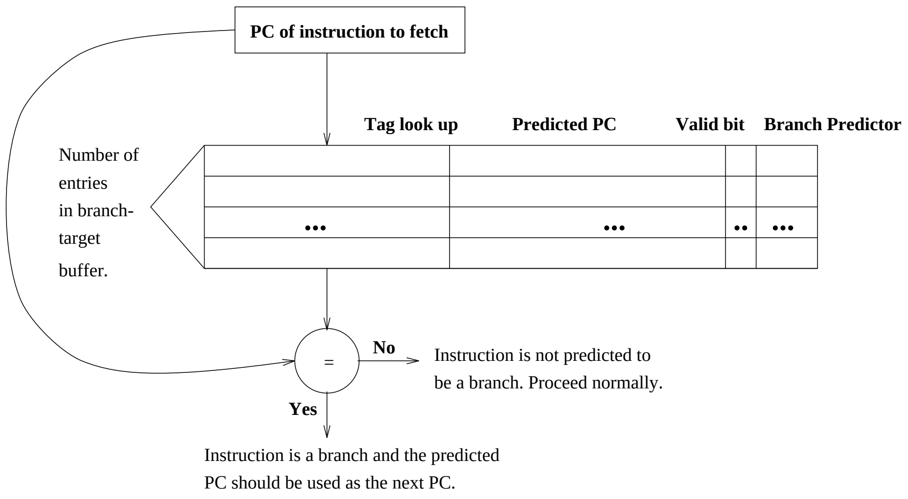

PC of instruction to fetch
Tag look up
Predicted PC
Valid bit
Branch Predictor
Number of entries in branch-target buffer.
=
No
Instruction is not predicted to be a branch. Proceed normally.
Yes
Instruction is a branch and the predicted PC should be used as the next PC.

Fig. 24. A Branch Target Buffer

6.2.2 Branch Predictors:

The n-bit predictor scheme predicts the outcome of the branch using 2^n state diagram. When n is equal to one, the predictor predicts the next execution path of a branch based upon the previous outcome of the branch. This predictor has a performance drawback such that, when a loop branch is almost taken, the same branch will likely be predicted incorrectly twice, rather than once. As an illustration for such a mispredicted branch, consider the example code fragment shown in Figure 25. Assume that one-bit prediction information is in the BTB for `branch`

```
i = 1;
while (i < 10) {        /* branch 1 */
   j = 1;
   while (j < 10) {     /* branch 2 */
      ...
      j++;
   }
   i++;
}
...
```

Fig. 25. An Example for a Mispredicted Branch

2. Mispredicting the tenth iteration of `branch 2` is inevitable since one-bit prediction information indicates that `branch 2` will be taken. However, when `branch 2` is accessed again after entering the inner loop for the second time, `branch 2` will be mispredicted as not taken. Thus, the prediction accuracy for `branch 2` that is taken in 90% of the iterations turns out to be only 80%.

In order to remedy this, two-bit predictor are often used. Consider the two-bit state diagram shown in Figure 26. By having intermediate branch prediction states, such as **State 1** and **State 2**, the above performance shortcoming of one-bit predictor can be resolved. The two-bit predictor approach has been reported to do almost as well as the more general n-bit predictors [Hennessy and Patterson 1996](see page 263), and most machines rely on the two-bit predictor instead of the more general n-bit predictor.

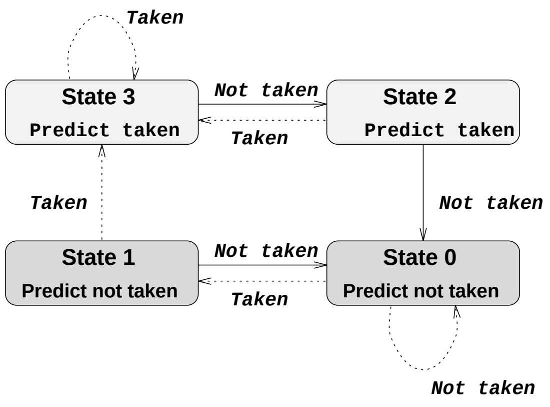


Fig. 26. The states in a two-bit predicton scheme

In many cases, the execution path of a branch can be easily determined by observing the outcomes of the previous branch executions [Pan et al. 1992]. Consider the code fragment in Figure 27. If the branch 1 and 2 are taken, then the execution path of branch 3 can be easily predicted as not taken. The n-bit predictors can be further improved to make a prediction by using the outcomes of other branches.

```
if (aa == 2)        /* branch 1 */
   aa = 0;
if (bb == 2)        /* branch 2 */
   bb = 0;
if (aa != bb) {     /* branch 3 */
  ....
}
```

Fig. 27. An Example Code Fragment for Branch Correlation

Such predictors are known as (m,n) correlation predictors. They use the outcome of the previous m branches to choose from 2^m branch predictors, each of which is a n-bit predictor for a single branch. The (m,n) predictors require one m-bit shift register to store the outcomes of the last m branch execution (0 for not taken, 1 for taken). This shift register can identify 2^m different contexts of a branch. Studies reported that (m,n) correlation predictors provide more accuracy than that of n-bit predictors [Pan et al. 1992; Hennessy and Patterson 1996].

6.2.3 BTB Management:

The target address for a branch is only placed in the buffer once the branch is taken. An indirect jump can be considered not taken (and therefore not placed in the buffer) if the target is the instruction following the indirect jump. If a branch or indirect jump is not in the buffer and it was not taken, then no delay is necessary since the not taken address is already calculated by the CPU. If the actual target of the indirect jump does not match the target in the buffer, then the branch target buffer is updated to contain the last target of the jump unless the same target is still predicted as taken. To maximize the performance of BTB, a branch, which is not in the BTB and is not taken, never replaces an entry in the buffeion of a branch in the BTB turns out to be not taken , the associated entry is immediately invalidated [Perleberg and Smith 1993]. This approach has the effect of never replacing an entry in the buffer with a branch that is not taken. Remember that a branch is predicted as not taken if it is not found in the buffer. Note that, when the BTB uses correlating information from a (m,n) correlation predictor, the m-bit shift register does not reflect the outcome of previous indirect jump executions. The major reason is that there are several targets of the indirect jump that can be considered as taken addresses [Wall 1991]. However, indirect jumps still use correlating information from the previous m executed branches.

6.2.4 Expected Benefits from Branch Coalescing Transformation with BTBs:

Indirect jumps typically have higher misprediction rates than conditional branches since an indirect jump may have many possible targets [Chang et al. 1997]. It is the authors' contention that higher misprediction rates do not necessarily mean worse performance. One must remember that several branches are being coalesced into a single indirect jump. Thus, the total number of mispredictions instead of the misprediction rate should be used when trying to measure branch target buffer performance with and without branch coalescing.

The authors argue that with comparable branch target buffer support, an indirect jump will cause no more mispredictions than the set of conditional branches it replaced. If the target of an indirect jump is mispredicted, then the target of the indirect jump changed from the last time it was executed. Likewise, at least one of the conditional branches that would have been executed instead of the indirect jump must have had different behavior and would also likely result in a misprediction. There are actually two reasons why fewer mispredictions would occur after branch coalescing. First, an indirect jump can cause at most one misprediction when executed. The execution of a sequence of the replaced conditional branches may cause multiple mispredictions. Second, there should be less contention for entries in the branch target buffer since there will be only one indirect jump as compared to the set of branches the indirect jump replaced.

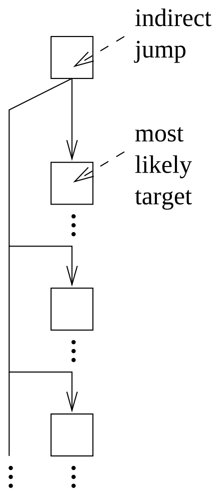


Fig. 28. Placing the Most Likely Target

Architectural and compiler support can be used to further reduce the number of mispredictions from indirect jumps. Indirect jump history and a target cache containing the targets of the indirect jump that have been encountered have been used to improve prediction accuracy [Chang et al. 1997]. The authors used compiler support to reduce the number of mispredictions. Often targets of an indirect jump have the block containing the indirect jump as their only predecessor. Value range analysis was performed to predict the most likely target for each indirect jump, which was placed immediately following the indirect jump block as shown in Figure 28. Thus, jumps to this target will result in no delay when the tag for the indirect jump is not found in the buffer since this address will be treated as the not taken address. Note that the authors do not suggest that the described approach is the best BTB design and configuration to support indirect jumps. Instead, the authors are simply showing that, with comparable BTB support for indirect jumps, aggressively coalescing branches into indirect jumps can result in improved branch prediction performance. The branch prediction simulation results from various configurations will be shown in Chapter 8.1.3. With specialized BTB support for indirect jumps [Chang et al. 1997], even better results should be obtained.

Some machines provide other special architectural support for speculative execution of instructions dependent on branches, such as boosting [Smith et al. 1990] and predicted execution [Pnevmatikatos and Sohi 1994; Mahlke et al. 1994]. The relative cost of an indirect jump versus the set of branches it replaces will be affected by such support. The compiler writer must use appropriate cost estimates based on the architectural support available for branches and indirect jumps on the target machine. An optimizer could also later convert indirect jumps into a sequence of conditional branches to exploit such architectural support.

7. RESULTS

Various measurements are given in this chapter that shows the benefits of applying the branch coalescing transformation. Several common Unix utilities were selected as benchmarks since non-numerical applications tend to have complex conditional control flow.

7.1 Dynamic Measurements by Instrumenting Code

The following measurements were collected on code generated by *vpcc* (Very Portable C Compiler)[9]using *ease* (Environment for Architectural Study and Experimentation [Davidson and Whalley 1991]) on the SPARC architecture for the selected benchmark Unix utilities. In order to isolate the benefits from the source of the branch coalescing transformation, each measurement was gathered using three different versions of *vpcc*.

1. *None*: *cfe* (C front-end) strictly translates every C `switch` statement into only the form of linear search.
2. *Original*: *cfe* translates a C `switch` statement into one of the following three alternative forms.
   (1) indirect jump using hashing (jump) tables,
   (2) binary search, or
   (3) linear search.
   Table V shows the heuristics used for the translation decision. [Yang et al. 1998; Sale 1981; Spuler 1994],

Table V. Heuristics Used for Translating C `switch` Statements

Term	Definition
n	Number of cases in a `switch` statement
m	Number of possible values between the first and last case

Indirect Jump	Binary Search	Linear Search
$n \geq 4$ && $m \leq 3n$	!indirect_jump && $n \geq 8$	!indirect_jump && !binary_search

[9]The overall structure of the compiler system is described in Figure 4 (Section 2).

3. *Coalesce*: *cfe* strictly translates every C `switch` statement into only the form of linear search. *vpo* (Very Portable Optimizer) is modified to coalese conditional branches into indirect jumps as described in Sections 4 and 5.

7.1.1 Number of Instructions Executed:

Table VI shows the number of instructions executed for each benchmark with three different versions of *vpcc* as described above. The *None* column in Table VI contains the dynamic number of instructions, which was obtained by using the *None* version. The *Original* and *Coalesce* columns respectively show the percentage changes as compared to *None* column when the *Original* and *Coalesce* versions are used in this measurement. The *Original* column[10]quantifies the benefits when indirect jumps from tables were only generated by the compiler front-end. The measurements show that a substantial benefit was obtained by conventional translation of multiway selection statements into jump tables. The *Coalesce* column quantifies the benefits when coalescing a set of branches using the techniques described in Sections 4 and 5. These frequency measurements indicate that our branch coalescing techniques (*Coalesce* column) can significantly outperform the conventional translation techniques (*Original* column) in reducing the number of instructions executed. The variance between different programs in the *Original* and *Coalesce* columns indicates that the benefit will depend on number of branches in a sequence. Note that coalescing had a negative impact on performance when performance estimates were overly optimistic or pessimistic, which occurred for *join* and *nroff*.

Table VI. Dynamic Instruction Frequency Measurements

Program	*None*	*Original*	*Coalesce*
awk	13,666,952	-0.294%	**-3.118%**
cb	19,739,127	-12.976%	**-21.204%**
cpp	30,985,306	-37.421%	**-38.538%**
ctags	81,040,455	-0.545%	**-24.160%**
deroff	15,511,056	-0.193%	**-1.153%**
grep	11,810,070	-21.620%	**-24.370%**
hyphen	19,535,372	0.000%	**-2.187%**
join	3,552,801	0.000%	**0.325%**
lex	10,052,031	-0.230%	**-0.689%**
nroff	25,118,855	-0.155%	**-0.017%**
pr	78,106,755	0.000%	**-7.760%**
ptx	20,059,920	0.000%	**-10.196%**
sdiff	17,582,760	0.000%	**-0.017%**
sed	17,321,920	-6.578%	**-7.600%**
sort	18,921,766	0.000%	**-33.053%**
wc	17,860,086	0.000%	**-27.590%**
yacc	25,658,688	-0.194%	**-0.307%**
average	25,036,387	-4.718%	**-11.861%**

[10]Note that the *Original* measurements included filling delay slots for indirect jumps from target blocks specified in jump tables to fairly compare the impact f branch coalescing.

7.1.2 Total Cache Work:

The branch coalescing impact on caching was a concern since misses from jump table loads could potentially have negative impact on performance. Table VII shows the average effect *Coalesce* had on instruction caching, data caching, and total cache work as compared to the *Original* cache measurements. The cache work cycles were calculated using the following formula, where a cache hit and a cache miss are counted as one cycle and ten cycles respectively [Smith 1982]. Note that it was assumed that data cache accesses could be performed simultaneously with instruction cache accesses.

$$\begin{aligned} \mathit{TOTAL\ CACHE\ WORK} = {} & \mathit{Instruction\ Cache\ Hits} + \\ & 10 * (\mathit{Instruction\ Cache\ Misses}) + \\ & 9 * (\mathit{Data\ Cache\ Misses}) \end{aligned}$$

The instruction cache work of *Coalesce* was reduced since the number of instructions referenced were diminished as compared to the *Original* measurements. As expected, the data cache work of *Coalesce* increased since jump table loads after branch coalescing are more frequently performed as compared to the *Original* compiler. The total cache work decreased since instruction cache accesses are more frequent than data cache accesses.

Table VII. Cache Work Improvement with a **Direct-Mapped Cache with 32 Byte Line Size**

CACHE SIZE	*Instruction*	*Data*	*TOTAL CACHE WORK*
1K	–7.095%	+6.680%	**–5.125%**
2K	–7.220%	+7.162%	**–5.614%**
4K	–4.909%	+5.066%	**–4.288%**
8K	–7.930%	+2.598%	**–7.460%**
16K	–8.231%	+3.995%	**–7.289%**
32K	–7.947%	+4.290%	**–7.328%**

7.1.3 Other Measurements:

Some other measurements not given in the tables provide useful information. There were on average about 0.901 more instructions executed between branches after *Coalesce* as compared to the *Original* measurements. Thus, the opportunities for scheduling on superscalar and superpipelined machines may be improved. In addition, coalescing only caused a 2.566% code size increase.

7.2 Execution Time Measurements

Execution time measurements were also collected on the following three models of SPARC processors.

—SPARCstation-IPC,

—SPARCstation-20, and

—Ultra-SPARC.

The first two machines did not provide any branch target/prediction buffer support. The third machine only provides target/prediction buffer support for branches, but no support for indirect jumps.

The time measurements were collected using the C run-time library function *times()* that uses the unit of time as a tick (1 second = 60 ticks). The execution times were obtained from the sum of reported *user* times of ten executions of each program.

7.2.1 Measurements on SPARCstation-IPC and SPARCstation-20:

The measurement results on these two machines are shown in Tables VIII and IX. There are a couple of reasons why the execution time decrease probably was not as significant as the reduction obtained from the number of instructions executed and total cache work. First, the execution time of an indirect jump operation required about the same time as two conditional branches. The authors anticipate that the relative cost of an indirect jump would decrease with target buffer support for branches and indirect jumps since the load delay for fetching the indirect jump target address could be avoided and fewer mispredictions would occur. Second, our compiler did not compile the C run-time library code. However, Tables VIII and IX show the execution time measurements, which included the execution of the C run-time library code. The authors anticipate that the execution time benefits can be further improved if our branch coalescing techniques were applied to the C run-time library code.

Table VIII. Execution Time Measurements for SPARCstation IPC

Program	**None**	**SPARCstation IPC**			
		Original	**Change**	*Coalesce*	**Change**
awk	2121 ts	2113 ts	–0.38%	2254 ts	5.90%
cb	1442 ts	1364 ts	–5.41%	1320 ts	–8.46%
cpp	1484 ts	1004 ts	–32.35%	1010 ts	–31.94%
ctags	4392 ts	4374 ts	–0.41%	4058 ts	–7.61%
deroff	917 ts	912 ts	–0.55%	911 ts	–0.65%
grep	442 ts	357 ts	–19.23%	340 ts	–23.08%
hyphen	741 ts	737 ts	–0.54%	736 ts	–0.68%
join	296 ts	296 ts	0.00%	303 ts	2.31%
lex	504 ts	503 ts	–0.20%	496 ts	–1.59%
nroff	1097 ts	1100 ts	0.27%	1118 ts	1.88%
pr	2854 ts	2857 ts	0.11%	2702 ts	–5.33%
ptx	3015 ts	3027 ts	0.40%	2962 ts	–1.76%
sdiff	9263 ts	9454 ts	2.01%	9280 ts	0.22%
sed	4670 ts	4449 ts	–4.76%	4403 ts	–5.72%
sort	680 ts	683 ts	0.44%	574 ts	–15.59%
wc	777 ts	778 ts	0.13%	678 ts	–12.74%
yacc	1163 ts	1281 ts	9.21%	1295 ts	10.19%
average			**–3.01%**		**–5.57%**

7.2.2 Measurements on UltraSPARC-1:

The same execution time measurements were also conducted on a UltraSPARC-1. As shown in Table X, the executables from *Coalesce* version compared to those from *None* version resulted in worse performance (even for the *Original* version). The authors strongly suspect that such disimprovement stems from no comparable target/prediction buffer support for indirect jumps. In order to properly estimate

Table IX. Execution Time Measurements for SPARCstation-20

Program	**None**	**SPARCstation-20**			
		Original	**Change**	*Coalesce*	**Change**
awk	617 ts	622 ts	0.80%	611 ts	–0.97%
cb	375 ts	363 ts	–3.20%	341 ts	–9.07%
cpp	493 ts	349 ts	–29.21%	372 ts	–29.41%
ctags	1267 ts	1208 ts	–4.66%	1109 ts	–12.47%
deroff	275 ts	274 ts	–0.36%	272 ts	–1.09%
grep	182 ts	162 ts	–10.99%	157 ts	–13.74%
hyphen	289 ts	289 ts	0.00%	282 ts	—2.42%
join	141 ts	142 ts	0.70%	144 ts	2.08%
lex	209 ts	206 ts	–1.44%	201 ts	–3.83%
nroff	381 ts	384 ts	0.78%	385 ts	1.04%
pr	874 ts	878 ts	0.46%	830 ts	–5.03%
ptx	1429 ts	1425 ts	–0.28%	1385 ts	–3.08%
sdiff	7520 ts	7479 ts	–0.55%	7475 ts	–0.60%
sed	1401 ts	1332 ts	–4.93%	1320 ts	–5.78%
sort	259 ts	258 ts	–0.39%	256 ts	–1.16%
wc	252 ts	250 ts	–0.79%	246 ts	–2.38%
yacc	414 ts	436 ts	5.05%	428 ts	3.27%
average			**–2.88%**		**–4.98%**

the execution time impact on this machine from the coalescing transformation, EASE (Environment for Architectural Study and Experimentation [Davidson and Whalley 1991]) was extended to be able to simulate branch prediction with BTB support as shown in Figure 24.

Indirect jumps from tables are generally considered to cause poorer branch prediction performance. The reason for this view is that indirect jumps typically have higher misprediction rates than conditional branches since an indirect jump may have many possible targets. However, the essence of branch coalescing transformation is to replace several conditional branches into an indirect jump. Thus, it was contended that the total number of mispredictions instead of the misprediction rate should be used when trying to measure branch target/prediction buffer performace with and without branch coalescing transformation.

Tables XI, XII, and XIII show the results from such branch simulation. As contended by the authors, even though the misprediction ratio went up after performing the branch coalescing transformation, the total number of mispredictions was decreased. Note that both the *Original* and the *Coalesce* measurements were obtained with the assumption of comparable hardware prediction support for indirect jumps.

7.3 Compile-Time Overhead

Initially, the *compile-time* overhead of branch coalescing was quite excessive. Two improvements were made to increase compile-time efficiency for the branch coalescing transformation. These improvements were decreasing the number of basic blocks used to represent jump tables and avoiding unnecessary attempts to coalesce branches.

7.3.1 Reducing the Number of Basic Blocks:

Table X. Execution Time Measurements for Ultra-SPARCstation

Program	None	Ultra-SPARCstation			
		Original	**Change**	*Coalesce*	**Change**
awk	479 ts	483 ts	0.83%	488 ts	1.84%
cb	261 ts	268 ts	2.61%	266 ts	1.88%
cpp	305 ts	286 ts	–6.23%	279 ts	–8.53%
ctags	936 ts	943 ts	0.74%	1008 ts	7.14%
deroff	199 ts	205 ts	2.93%	200 ts	0.50%
grep	123 ts	120 ts	–2.44%	118 ts	–4.07%
hyphen	206 ts	208 ts	0.96%	225 ts	8.44%
join	104 ts	104 ts	0.00%	106 ts	1.89%
lex	129 ts	135 ts	4.44%	137 ts	5.84%
nroff	239 ts	243 ts	1.65%	242 ts	1.24%
pr	529 ts	534 ts	0.94%	581 ts	8.95%
ptx	1013 ts	1016 ts	0.30%	1020 ts	0.69%
sdiff	6651 ts	6676 ts	0.37%	6660 ts	0.14%
sed	922 ts	900 ts	0.37%	893 ts	–3.15%
sort	178 ts	177 ts	–0.56%	207 ts	14.01%
wc	171 ts	170 ts	–0.59%	208 ts	17.79%
yacc	284 ts	293 ts	3.07%	285 ts	0.35%
average			**0.67%**		**3.23%**

The complexity for both data and control-flow analysis for code improving transformations is proportional to the number of basic blocks. In fact, the authors found that most of the compile-time overhead was due to the detrimental effect that additional basic blocks had on subsequent analysis and transformations.

Originally, *vpo* (Very Portable Optimizer) represented each entry in the jump table as a separate basic block. This representation scheme was a concern to the authors since most of the techniques in Chapter 6 to make indirect jumps more efficient were applied at the cost of duplicatig jump table entries. In order to avoid excessive generation of basic blocks from those techniques, an alternative scheme has been designed and implemented to compactly represent the control flow for a jump table.

As an illustration, consider the RTLs shown in Figure 29(a), which is the snapshot after eliminating the value range check instructions for the indirect jump by enumerating 256 table entries into the jump table. However, such 256 blocks for the jump table could be efficiently represented into fewer blocks when consecutive jump table entries that contain the same target address can be grouped into a single basic block. Figure 29(b) shows a compact representation of the original control flow. Note that each basic block containing a jump table entry has another field to indicate the repetition count such that the jump table entries can be restored while SPARC assembly code is being produced.

7.3.2 Avoiding Unnecessary Coalescing Attempts:

In *vpo*, several loop transformations are iteratively applied until no further improvement (change(es)) can be made, as depicted in Figure 5. Coalescing of branches was treated as a transformation for a loop since the transformation typically requires extra registers. The authors coalesced the branches from the innermost loop outward after all other transformations for a given loop have been ini-

Table XI. Branch Misprediction Ratio and Number of Mispredicted Branches with a **Direct-Mapped BTB** with (0,1) **Correlation Predictor**

Entries in BTB	**Branch Misprediction Ratio**			**Percentage Reductions in Mispredicted Branches**
	Original	*Coalesce*	**Difference**	
32	0.1182	0.1365	**0.0183**	**–5.60%**
64	0.1050	0.1152	**0.0102**	**–9.09%**
128	0.0935	0.1042	**0.0107**	**–9.52%**
256	0.0892	0.0988	**0.0096**	**–10.35%**
512	0.0871	0.0964	**0.0094**	**–10.67%**
1024	0.0811	0.0961	**0.0149**	**–4.43%**

Table XII. Branch Misprediction Ratio and Number of Mispredicted Branches with a **Direct-Mapped BTB** with (0,2) **Correlation Predictor**

Entries in BTB	**Branch Misprediction Ratio**			**Percentage Reductions in Mispredicted Branches**
	Original	*Coalesce*	**Difference**	
32	0.1118	0.1252	**0.0134**	**–8.11%**
64	0.0971	0.1014	**0.0043**	**–12.28%**
128	0.0848	0.0899	**0.0051**	**–12.81%**
256	0.0804	0.0841	**0.0038**	**–14.11%**
512	0.0779	0.0824	**0.0045**	**–14.10%**
1024	0.0720	0.0817	**0.0097**	**–7.20%**

Table XIII. Branch Misprediction Ratio and Number of Mispredicted Branches with a **Direct-Mapped BTB** with (2,2) **Correlation Predictor**

Entries in BTB	**Branch Misprediction Ratio**			**Percentage Reductions in Mispredicted Branches**
	Original	*Coalesce*	**Difference**	
32	0.1131	0.1271	**0.0140**	**–8.33%**
64	0.0969	0.1024	**0.0055**	**–12.07%**
128	0.0840	0.0902	**0.0062**	**–12.53%**
256	0.0788	0.0836	**0.0048**	**–13.56%**
512	0.0758	0.0817	**0.0059**	**–13.30%**
1024	0.0695	0.0809	**0.0114**	**–6.15%**

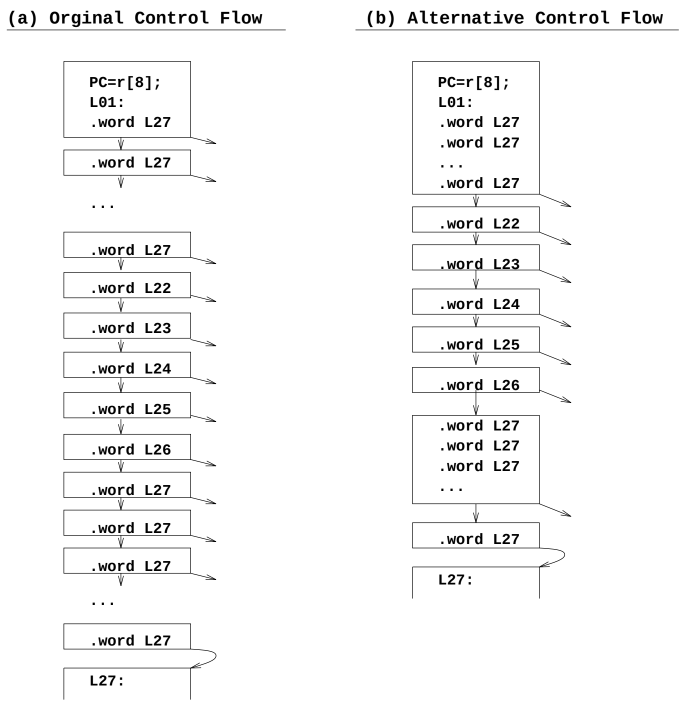


Fig. 29. Control Flow Representations for Indirect Jump Table shown in Figure 6.3

tially attempted. Within such a loop optimization framework, unnecessary branch coalescing analysis could be avoided. For a given loop, if the branch coalescing analysis has been already applied without any transformation, then there is typically no need to re-apply the analysis for the same loop. Most of transformations from the branch coalescing are typically completed during the first pass of a loop optimization process. The authors found that other improving transformations rarely provided new opportunities for branch coalescing. Therefore, the branch coalescing transformation was not applied during the second pass of the same loop optimization process.

7.3.3 Compilation Overhead:

Compile time measurements were collected on a SPARCstation-20 using the C run-time library function *times()*. The compile times were obtained from the average of the sum of the reported *user* and *system* times of 10 compilations of each benchmark. Table XIV compares the results with *None* and *Coalesce* (that is, with and without the branch coalescing transformation.) The authors suspect that the compilation overhead can be reduced with some additional tuning. Some portion of compilation overhead in system time is due to `I/O` in producing jump table entries when generating SPARC assembly code. This overhead can be avoided when an assembler supports a directive that specifies a repetition factor for consecutive values that are identical (e.g. `.word  <value><repetition factor>`).

Table XIV. Compile Time Measurements

Program	None		Coalesce		**Extra Overhead**
	user	*system*	*user*	*system*	
awk	39.40 sec	7.18 sec	76.35 sec	7.88 sec	+80.82%
cb	4.83 sec	0.72 sec	5.48 sec	0.77 sec	+12.61%
cpp	23.02 sec	3.17 sec	36.28 sec	3.40 sec	+51.56%
ctags	9.60 sec	0.72 sec	14.07 sec	0.93 sec	+45.40%
deroff	33.68 sec	1.03 sec	38.17 sec	1.10 sec	+13.11%
grep	4.68 sec	0.67 sec	6.53 sec	0.78 sec	+36.76%
hyphen	1.37 sec	0.60 sec	1.53 sec	0.60 sec	+9.32%
join	3.58 sec	0.62 sec	4.25 sec	0.67 sec	+17.06%
lex	41.40 sec	3.78 sec	49.22 sec	4.08 sec	+17.96%
nroff	43.25 sec	6.13 sec	45.83 sec	6.32 sec	+5.60%
pr	6.03 sec	0.82 sec	6.52 sec	0.85 sec	+7.54%
ptx	6.42 sec	0.78 sec	7.13 sec	0.78 sec	+9.95%
sdiff	8.37 sec	0.78 sec	12.40 sec	0.98 sec	+45.47%
sed	20.52 sec	2.27 sec	24.65 sec	2.45 sec	+18.95%
sort	9.30 sec	0.68 sec	10.38 sec	0.68 sec	+10.85%
wc	0.95 sec	0.50 sec	1.50 sec	0.53 sec	+40.23%
yacc	34.87 sec	2.67 sec	43.47 sec	2.93 sec	+23.62%
average	17.13 sec	1.95 sec	22.57 sec	2.10	**+26.28%**

8. FUTURE WORK

There are other areas that could be investigated to provide additional opportunities for coalescing conditional branches. One factor that limited the opportunities for coalescing branches into indirect jumps was not performing interprocedural analysis to more effectively determine value ranges. Often `int` arguments being compared to constants in one function are loaded from memory as a byte in a different function. Interprocedural analysis would allow the first three instructions in Figure 11(b) comprising the initial range check to be avoided more frequently.

Profiling could also be used to help determine when coalescing was worthwhile. The authors statically estimated the average number of branches that would be executed through a set of related branches. Coalescing can have a negative impact on performance when these estimates are overly optimistic or pessimistic. Profiling would provide more accurate estimates for coalescing decisions. In general, detecting bounded ranges and using an estimated frequency for character values provided good heuristics when making coalescing decisions. This approach has promising implications for conventional branch prediction.

9. CONCLUSIONS

This paper has described compiler support for effectively exploiting indirect jumps. The general improving transformation presented for coalescing branches after code generation provided benefits that otherwise would not be available.

A general approach was designed and implemented to aggressively replace a set of branches into a single indirect jump as opposed to only considering indirect jumps when translating multiway statements.

Various techniques were developed and implemented to efficiently perform the indirect jump operation by analyzing the context of the given machine instructions.

Applying these techniques often resulted in the execution of only two instructions on the SPARC. In order to provide an effective branch coalescing transformation, two cost/benefit analyses were designed and applied by estimating the average number of branches executed for the detected set of coalescent branches. In order to coalesce a set of conditional branches, which are often separated by blocks of intervening instructions, a restructuring algorithm using code duplication was designed and implemented. Furthermore, the original delay slot filling scheme was extended to usefully fill the delay slots of indirect jumps. Thus, a code-improving transformation was designed and implemented in order to essentially provide early resolution of conditional branches that may originally have been some distance from the point where the indirect jump is inserted.

BTBs (Branch Target Buffer) are available to reduce the cost of branches on many machines. The branch coalescing impact on branch mispredictions was a concern to the authors. The authors' contention was that with comparable target buffer support for indirect jumps, the total number of branch mispredictions should be reduced since several branches are being coalesced into a single indirect jump. To justify the contention, the authors accomplished the following tasks. First, the EASE environment [Davidson and Whalley 1991] was extended to be able to simulate effects on branch mispredictions with BTB support for branches and indirect jumps [Hennessy and Patterson 1996](see page 276). Second, in order to better exploit a BTB for indirect jumps, a compiler analysis technique was implemented to locate the most probable target of the indirect jump immediately after the jump as a fall-through destination. Thus, if an indirect jump is not in the buffer, then no delay is necessary since the next address of the indirect jump is already calculated by the CPU.

Finally, various measurements were collected to demonstrate the benefit of applying the branch coalescing transformation. The additional benefits from coalescing noncontiguous branches were contrasted with the simpler analysis required for only coalescing contiguous branches.

The results showed reductions in the number of instructions executed and branch mispredictions, total cache work, and execution time at the cost of tolerable compile-time overhead.

REFERENCES

Allen, F. and Cocke, J. 1971. *Design and Optimization of Compilers*. Prentice-Hall, Englewood Cliffs, NJ.

Altman, N. and Weiderman, N. 1987. Timing variation in dual-loop benchmarks. Technical report (Oct.), Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

Benitez, M. E. and Davidson, J. W. 1988. A portable global optimizer and linker. In *ACM SIGPLAN Conference on Programming Language Design and Implementation*, pp. 329–338.

Bodik, R., Gupta, R., and Soffa, M. L. 1997. Interprocedural conditional branch elimination. In *ACM SIGPLAN Conference on Programming Language Design and Implementation.*

Calder, B. and Grunwald, D. 1994. Reducing indirect function call overhead in C++ programs. In *Proceedings of the ACM Symposium on Principles and Practice of Programming Languages*, pp. 397–408.

Chang, P.-Y., Hao, E., and Patt, Y. N. 1997. Target prediction for indirect jumps. In *The 24th Annual International Symposium on Computer Architecture.*

Clapp, R. M., Duchesneau, L., Volz, R. A., Mudge, T. N., and Schultze, T. 1986. Toward real-time performace benchmarks for ADA. *Communications of the ACM 29,* 8 (Aug.), 760–778.

DAVIDSON, J. W. AND JINTURKAR, S. 1996. Aggressive loop unrolling in a retargetable, optimizing compiler. In *Proceedings of Compiler Construction Conference*, pp. 59–73.

DAVIDSON, J. W. AND WHALLEY, D. B. 1989. Quick compilers using peephole optimizations. *Software Practice & Experience 19,* 1 (Jan.), 195–203.

DAVIDSON, J. W. AND WHALLEY, D. B. 1991. A design environment for addressing architecture and compiler interactions. *Microprocessors and Microsystems 15,* 9 (Nov.), 459–472.

HENNESSY, J. AND PATTERSON, D. 1996. *Computer Architecture: A Quantitative Approach* (Second ed.). Morgan Kaufmann, San Francisco, CA.

HOLLER, A. M. 1996. Optimization for a superscalar out-of-order machine. In *Proceedings of the 29th International Symposium on Microarchitecture*, pp. 336–348.

JOHNSON, S. C. 1979. *A Tour Through the Portable C Compiler.* Unix Programmer's Manual 7th Edition Section 33.

MAHLKE, S. A., HANK, R. E., BRINGMANN, R. A., GYLLENHAAL, J. C., GALLAGHER, D. M., AND HWU, W. W. 1994. Characterizing the impact of predicated execution on branch prediction. In *Proceedings of the 27th International Symposium on Microarchitecture*, pp. 217–227.

MUELLER, F. AND WHALLEY, D. B. 1995. Avoiding conditional branches by code replication. In *ACM SIGPLAN Conference on Programming Language Design and Implementation*, pp. 56–66.

PAN, S., SO, K., AND RAHMEH, J. T. 1992. Improving the accuracy of dynamic branch prediction using branch correlation. In *Architectural Support for Programming Languages and Operating Systems*, pp. 76–84.

PERLEBERG, C. H. AND SMITH, A. J. 1993. Branch target buffer design and optimization. *IEEE Transactions on Computers 42,* 4 (April), 396–412.

PNEVMATIKATOS, D. N. AND SOHI, G. S. 1994. Guarded execution and branch prediction in dynamic ILP processors. In *Proceedings of the 21th International Symposium on Computuer Architecture*, pp. 120–129.

SALE, A. 1981. The implementation of case statements in Pascal. *Software-Practice and Experience 11*, 929–942.

SMITH, A. J. 1982. Cache memories. *Computing Surveys 14,* 3 (Sept.), 473–530.

SMITH, M. D., LAM, M. S., AND HOROWITZ, M. A. 1990. Boosting beyond static scheduling in a superscalar processor. In *Proceedings of the 17th International Symposium on Computer Architecture*, pp. 344–353.

SPULER, D. A. 1994. Compiler code generation for multiway branch statements as a static search problem. Technical report, Dept. of Computer Science, James Cook University, Townsville, 4811, Australia.

STALLMAN, R. M. 1990. *Using and Porting GNU CC (version 1.37.1).* Cambridge, MA: Free Software Foundation, Inc.

WALL, D. W. 1991. Limits of instruction-level parallelism. In *Architectural Support for Programming Languages and Operating Systems*, pp. 176–188.

YANG, M., UH, G.-R., AND WHALLEY, D. B. 1998. Improving performance by branch reordering. In *ACM SIGPLAN Conference on Programming Language Design and Implementation.*