

Chapter 8

Code Generation

The final phase in our compiler model is the code generator. It takes as input the intermediate representation (IR) produced by the front end of the compiler, along with relevant symbol table information, and produces as output a semantically equivalent target program, as shown in Fig. 8.1.

The requirements imposed on a code generator are severe. The target program must preserve the semantic meaning of the source program and be of high quality—that is, it must make effective use of the available resources of the target machine. Moreover, the code generator itself must run efficiently.

The challenge is that, mathematically, the problem of generating an optimal target program for a given source program is undecidable. Many of the subproblems encountered in code generation, such as register allocation, are computationally intractable. In practice, we must be content with heuristic techniques that generate good, but not necessarily optimal, code. Fortunately, heuristics have matured enough that a carefully designed code generator can produce code that is several times faster than code produced by a naive one.

Compilers that need to produce efficient target programs include an optimization phase prior to code generation. The optimizer maps the IR into IR from which more efficient code can be generated. In general, the code optimization and code generation phases of a compiler, often referred to as the *back end*, may make multiple passes over the IR before generating the target program. Code optimization is discussed in detail in Chapter 9. The techniques presented in this chapter can be used whether or not an optimization phase occurs before code generation.

A code generator has three primary tasks: instruction selection, register

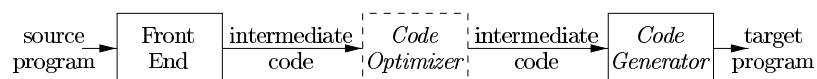


Figure 8.1 Position of code generator

CHAPTER 8 CODE GENERATION

allocation and assignment and instruction ordering. The importance of these tasks is outlined in Section 8.1. Instruction selection involves choosing appropriate target machine instructions to implement the IR statements. Register allocation and assignment involves deciding what values to keep in which registers. Instruction ordering involves deciding in what order to schedule the execution of instructions.

This chapter presents algorithms that code generators can use to translate the IR into a sequence of target language instructions for simple register machines. The algorithms will be illustrated by using the machine model in Section 8.2. Chapter 10 covers the problem of code generation for complex modern machines that support a great deal of parallelism within a single instruction.

After discussing the broad issues in the design of a code generator, we show what kind of target code a compiler needs to generate to support the abstractions embodied in a typical source language. In Section 8.3 we outline implementations of static and stack allocation of data areas and show how names in the IR can be converted into addresses in the target code.

Many code generators partition IR instructions into basic blocks, which consist of sequences of instructions that are always executed together. The partitioning of the IR into basic blocks is the subject of Section 8.4. The following section presents simple local transformations that can be used to transform basic blocks into modified basic blocks from which more efficient code can be generated. These transformations are a rudimentary form of code optimization, although the deeper theory of code optimization will not be taken up until Chapter 9. An example of a useful local transformation is the discovery of common subexpressions at the level of intermediate code and the resultant replacement of arithmetic operations by simpler copy operations.

Section 8.6 presents a simple code generation algorithm that generates code for each statement in turn, keeping operands in registers as long as possible. The output of this kind of code generator can be readily improved by peephole optimization techniques such as those discussed in the following Section 8.7.

The remaining sections explore instruction selection and register allocation.

8.1 Issues in the Design of a Code Generator

While the details are dependent on the specifics of the intermediate representation, the target language, and the run-time system, tasks such as instruction selection, register allocation and assignment, and instruction ordering are encountered in the design of almost all code generators.

The most important criterion for a code generator is that it produce correct code. Correctness takes on special significance because of the number of special cases that a code generator might face. Given the premium on correctness, designing a code generator so it can be easily implemented, tested, and maintained is an important design goal.

8.1 ISSUES IN THE DESIGN OF A CODE GENERATOR

8.1.1 Input to the Code Generator

The input to the code generator is the intermediate representation of the source program produced by the front end along with information in the symbol table that is used to determine the run time addresses of the data objects denoted by the names in the IR.

The many choices for the IR include three address representations such as quadruples, triples, indirect triples, virtual machine representations such as bytecodes and stack machine code, linear representations such as post-fix notation, and graphical representations such as syntax trees and DAGs. Many of the algorithms in this chapter are couched in terms of the representations considered in Chapter 6: three address code, trees, and DAGs. The techniques we discuss can be applied, however, to the other intermediate representations as well.

In this chapter, we assume that the front end has scanned, parsed, and translated the source program into a relatively low level IR, so that the values of the names appearing in the IR can be represented by quantities that the target machine can directly manipulate, such as integers and floating point numbers. We also assume that all syntactic and static semantic errors have been detected, that the necessary type checking has taken place, and that type conversion operators have been inserted wherever necessary. The code generator can therefore proceed on the assumption that its input is free of these kinds of errors.

8.1.2 The Target Program

The instruction set architecture of the target machine has a significant impact on the difficulty of constructing a good code generator that produces high quality machine code. The most common target machine architectures are RISC (reduced instruction set computer), CISC (complex instruction set computer), and stack based.

A RISC machine typically has many registers, three address instructions, simple addressing modes, and a relatively simple instruction set architecture. In contrast, a CISC machine typically has few registers, two address instructions, a variety of addressing modes, several register classes, variable length instructions, and instructions with side effects.

In a stack based machine, operations are done by pushing operands onto a stack and then performing the operations on the operands at the top of the stack. To achieve high performance, the top of the stack is typically kept in registers. Stack based machines almost disappeared because it was felt that the stack organization was too limiting and required too many swap and copy operations.

However, stack based architectures were revived with the introduction of the Java Virtual Machine (JVM). The JVM is a software interpreter for Java bytecodes, an intermediate language produced by Java compilers. The inter-

preter provides software compatibility across multiple platforms a major factor in the success of Java

To overcome the high performance penalty of interpretation which can be on the order of a factor of 10 *just in time* JIT Java compilers have been created These JIT compilers translate bytecodes during run time to the native hardware instruction set of the target machine Another approach to improving Java performance is to build a compiler that compiles directly into the machine instructions of the target machine bypassing the Java bytecodes entirely

Producing an absolute machine language program as output has the advantage that it can be placed in a fixed location in memory and immediately executed Programs can be compiled and executed quickly

Producing a relocatable machine language program often called an *object module* as output allows subprograms to be compiled separately A set of relocatable object modules can be linked together and loaded for execution by a linking loader Although we must pay the added expense of linking and loading if we produce relocatable object modules we gain a great deal of flexibility in being able to compile subroutines separately and to call other previously compiled programs from an object module If the target machine does not handle relocation automatically the compiler must provide explicit relocation information to the loader to link the separately compiled program modules

Producing an assembly language program as output makes the process of code generation somewhat easier We can generate symbolic instructions and use the macro facilities of the assembler to help generate code The price paid is the assembly step after code generation

In this chapter we shall use a very simple RISC like computer as our target machine We add to it some CISC like addressing modes so that we can also discuss code generation techniques for CISC machines For readability we use assembly code as the target language As long as addresses can be calculated from offsets and other information stored in the symbol table the code generator can produce relocatable or absolute addresses for names just as easily as symbolic addresses

8.1.3 Instruction Selection

The code generator must map the IR program into a code sequence that can be executed by the target machine The complexity of performing this mapping is determined by factors such as

- the level of the IR
- the nature of the instruction set architecture
- the desired quality of the generated code

If the IR is high level the code generator may translate each IR statement into a sequence of machine instructions using code templates Such statement by statement code generation however often produces poor code that needs

8.1 ISSUES IN THE DESIGN OF A CODE GENERATOR

further optimization. If the IR reflects some of the low level details of the underlying machine, then the code generator can use this information to generate more efficient code sequences.

The nature of the instruction set of the target machine has a strong effect on the difficulty of instruction selection. For example, the uniformity and completeness of the instruction set are important factors. If the target machine does not support each data type in a uniform manner, then each exception to the general rule requires special handling. On some machines, for example, floating point operations are done using separate registers.

Instruction speeds and machine idioms are other important factors. If we do not care about the efficiency of the target program, instruction selection is straightforward. For each type of three address statement, we can design a code skeleton that defines the target code to be generated for that construct. For example, every three address statement of the form $x = y + z$ where x , y , and z are statically allocated, can be translated into the code sequence

LD R0, y	R0, y	load y into register R0
ADD R0, R0, z	R0, R0, z	add z to R0
ST x, R0	x, R0	store R0 into x

This strategy often produces redundant loads and stores. For example, the sequence of three address statements

a	b	c
d	a	e

would be translated into

LD R0, b	R0, b
ADD R0, R0, c	R0, R0, c
ST a, R0	a, R0
LD R0, a	R0, a
ADD R0, R0, e	R0, R0, e
ST d, R0	d, R0

Here, the fourth statement is redundant since it loads a value that has just been stored, and so is the third if a is not subsequently used.

The quality of the generated code is usually determined by its speed and size. On most machines, a given IR program can be implemented by many different code sequences, with significant cost differences between the different implementations. A naive translation of the intermediate code may therefore lead to correct but unacceptably inefficient target code.

For example, if the target machine has an increment instruction `INC`, then the three address statement $a = a + 1$ may be implemented more efficiently by the single instruction `INC a` rather than by a more obvious sequence that loads a into a register, adds one to the register, and then stores the result back into a .

LD	R0	a		R0	a
ADD	R0	R0	1	R0	R0 1
ST	a	R0		a	R0

We need to know instruction costs in order to design good code sequences but unfortunately accurate cost information is often difficult to obtain. Deciding which machine code sequence is best for a given three address construct may also require knowledge about the context in which that construct appears.

In Section 8.9 we shall see that instruction selection can be modeled as a tree pattern matching process in which we represent the IR and the machine instructions as trees. We then attempt to tile an IR tree with a set of subtrees that correspond to machine instructions. If we associate a cost with each machine instruction subtree we can use dynamic programming to generate optimal code sequences. Dynamic programming is discussed in Section 8.11.

8.1.4 Register Allocation

A key problem in code generation is deciding what values to hold in what registers. Registers are the fastest computational unit on the target machine but we usually do not have enough of them to hold all values. Values not held in registers need to reside in memory. Instructions involving register operands are invariably shorter and faster than those involving operands in memory so efficient utilization of registers is particularly important.

The use of registers is often subdivided into two subproblems:

1. *Register allocation* during which we select the set of variables that will reside in registers at each point in the program.
2. *Register assignment* during which we pick the specific register that a variable will reside in.

Finding an optimal assignment of registers to variables is difficult even with single register machines. Mathematically the problem is NP complete. The problem is further complicated because the hardware and/or the operating system of the target machine may require that certain register usage conventions be observed.

Example 8.1 Certain machines require *register pairs*—an even and next odd numbered register—for some operands and results. For example, on some machines integer multiplication and integer division involve register pairs. The multiplication instruction is of the form

$$M \ x \ y$$

where x the multiplicand is the odd register of an even-odd register pair and y the multiplier can be anywhere. The product occupies the entire even-odd register pair. The division instruction is of the form

8.1 ISSUES IN THE DESIGN OF A CODE GENERATOR

$D \ x \ y$

where the dividend occupies an even-odd register pair whose even register is x the divisor is y . After division, the even register holds the remainder and the odd register the quotient.

Now consider the two three-address code sequences in Fig. 8.2 in which the only difference in a and b is the operator in the second statement. The shortest assembly code sequences for a and b are given in Fig. 8.3.

t	a	b	t	a	b
t	t	c	t	t	c
t	t	d	t	t	d
a			b		

Figure 8.2 Two three-address code sequences

L	$R1$	a	L	$R0$	a
A	$R1$	b	A	$R0$	b
M	$R0$	c	A	$R0$	c
D	$R0$	d	$SRDA$	$R0$	32
ST	$R1$	t	D	$R0$	d
			ST	$R1$	t
a			b		

Figure 8.3 Optimal machine code sequences

R_i stands for register i . $SRDA$ stands for Shift Right Double Arithmetic and $SRDA \ R0 \ 32$ shifts the dividend into $R1$ and clears $R0$ so all bits equal its sign bit. L , ST , and A stand for load, store, and add, respectively. Note that the optimal choice for the register into which a is to be loaded depends on what will ultimately happen to t . \square

Strategies for register allocation and assignment are discussed in Section 8.8. Section 8.10 shows that for certain classes of machines we can construct code sequences that evaluate expressions using as few registers as possible.

8.1.5 Evaluation Order

The order in which computations are performed can affect the efficiency of the target code. As we shall see, some computation orders require fewer registers to hold intermediate results than others. However, picking a best order in the general case is a difficult NP-complete problem. Initially, we shall avoid

the problem by generating code for the three address statements in the order in which they have been produced by the intermediate code generator. In Chapter 10 we shall study code scheduling for pipelined machines that can execute several operations in a single clock cycle.

8.2 The Target Language

Familiarity with the target machine and its instruction set is a prerequisite for designing a good code generator. Unfortunately, in a general discussion of code generation it is not possible to describe any target machine in sufficient detail to generate good code for a complete language on that machine. In this chapter we shall use as a target language assembly code for a simple computer that is representative of many register machines. However, the code generation techniques presented in this chapter can be used on many other classes of machines as well.

8.2.1 A Simple Target Machine Model

Our target computer models a three address machine with load and store operations, computation operations, jump operations, and conditional jumps. The underlying computer is a byte addressable machine with n general purpose registers R_0, R_1, \dots, R_{n-1} . A full fledged assembly language would have scores of instructions. To avoid hiding the concepts in a myriad of details we shall use a very limited set of instructions and assume that all operands are integers. Most instructions consists of an operator followed by a target followed by a list of source operands. A label may precede an instruction. We assume the following kinds of instructions are available.

Load operations The instruction `LD dst $addr$` loads the value in location $addr$ into location dst . This instruction denotes the assignment $dst = addr$. The most common form of this instruction is `LD r x` which loads the value in location x into register r . An instruction of the form `LD r_1 r_2` is a *register to register copy* in which the contents of register r_2 are copied into register r_1 .

Store operations The instruction `ST x r` stores the value in register r into the location x . This instruction denotes the assignment $x = r$.

Computation operations of the form `OP dst src_1 src_2` where OP is a operator like ADD or SUB and dst , src_1 and src_2 are locations, not necessarily distinct. The effect of this machine instruction is to apply the operation represented by OP to the values in locations src_1 and src_2 and place the result of this operation in location dst . For example `SUB r_1 r_2 r_3` computes $r_1 = r_2 - r_3$. Any value formerly stored in r_1 is lost, but if r_1 is r_2 or r_3 the old value is read first. Unary operators that take only one operand do not have a src_2 .

8.2 THE TARGET LANGUAGE

Unconditional jumps The instruction `BR L` causes control to branch to the machine instruction with label L . `BR` stands for *branch*.

Conditional jumps of the form `Bcond r L` where r is a register, L is a label and *cond* stands for any of the common tests on values in the register r . For example `BLTZ r L` causes a jump to label L if the value in register r is less than zero and allows control to pass to the next machine instruction if not.

We assume our target machine has a variety of addressing modes

In instructions, a location can be a variable name x referring to the memory location that is reserved for x that is, the l value of x .

A location can also be an indexed address of the form $a[r]$ where a is a variable and r is a register. The memory location denoted by $a[r]$ is computed by taking the l value of a and adding to it the value in register r . For example, the instruction `LD R1, a[R2]` has the effect of setting $R1 \leftarrow contents\ a + contents\ R2$ where *contents* x denotes the contents of the register or memory location represented by x . This addressing mode is useful for accessing arrays where a is the base address of the array (that is, the address of the first element) and r holds the number of bytes past that address we wish to go to reach one of the elements of array a .

A memory location can be an integer indexed by a register. For example, `LD R1, 100[R2]` has the effect of setting $R1 \leftarrow contents\ 100 + contents\ R2$ that is, of loading into $R1$ the value in the memory location obtained by adding 100 to the contents of register $R2$. This feature is useful for following pointers as we shall see in the example below.

We also allow two indirect addressing modes. (r) means the memory location found in the location represented by the contents of register r and $100(r)$ means the memory location found in the location obtained by adding 100 to the contents of r . For example, `LD R1, 100(R2)` has the effect of setting $R1 \leftarrow contents\ contents\ 100 + contents\ R2$ that is, of loading into $R1$ the value in the memory location stored in the memory location obtained by adding 100 to the contents of register $R2$.

Finally, we allow an immediate constant addressing mode. The constant is prefixed by `#`. The instruction `LD R1, #100` loads the integer 100 into register $R1$ and `ADD R1, R1, #100` adds the integer 100 into register $R1$.

Comments at the end of instructions are preceded by `;`

Example 8.2 The three address statement $x \leftarrow y + z$ can be implemented by the machine instructions

CHAPTER 8 CODE GENERATION

LD R1 y	R1 y
LD R2 z	R2 z
SUB R1 R1 R2	R1 R1 R2
ST x R1	x R1

We can do better perhaps. One of the goals of a good code generation algorithm is to avoid using all four of these instructions whenever possible. For example y and z may have been computed in a register and if so we can avoid the LD step s . Likewise we might be able to avoid ever storing x if its value is used within the register set and is not subsequently needed.

Suppose a is an array whose elements are 8 byte values perhaps real numbers. Also assume elements of a are indexed starting at 0. We may execute the three address instruction $b = a[i]$ by the machine instructions

LD R1 i	R1 i
MUL R1 R1 8	R1 R1 8
LD R2 a R1	R2 contents a contents R1
ST b R2	b R2

That is the second step computes $8i$ and the third step places in register R2 the value in the i th element of a the one found in the location that is $8i$ bytes past the base address of the array a .

Similarly the assignment into the array a represented by three address instruction $a[j] = c$ is implemented by

LD R1 c	R1 c
LD R2 j	R2 j
MUL R2 R2 8	R2 R2 8
ST a R2 R1	contents a contents R2 R1

To implement a simple pointer indirection such as the three address statement $x = *p$ we can use machine instructions like

LD R1 p	R1 p
LD R2 0 R1	R2 contents 0 contents R1
ST x R2	x R2

The assignment through a pointer $*p = y$ is similarly implemented in machine code by

LD R1 p	R1 p
LD R2 y	R2 y
ST 0 R1 R2	contents 0 contents R1 R2

Finally consider a conditional jump three address instruction like

if $x < y$ goto L

8.2 THE TARGET LANGUAGE

The machine code equivalent would be something like

LD	R1	x		R1	x
LD	R2	y		R2	y
SUB	R1	R1	R2	R1	R1 R2
BLTZ	R1	M		if R1 < 0	jump to M

Here *M* is the label that represents the first machine instruction generated from the three address instruction that has label *L*. As for any three address instruction we hope that we can save some of these machine instructions because the needed operands are already in registers or because the result need never be stored. □

8.2.2 Program and Instruction Costs

We often associate a cost with compiling and running a program. Depending on what aspect of a program we are interested in optimizing, some common cost measures are the length of compilation time and the size, running time and power consumption of the target program.

Determining the actual cost of compiling and running a program is a complex problem. Finding an optimal target program for a given source program is an undecidable problem in general, and many of the subproblems involved are NP hard. As we have indicated, in code generation we must often be content with heuristic techniques that produce good but not necessarily optimal target programs.

For the remainder of this chapter we shall assume each target language instruction has an associated cost. For simplicity we take the cost of an instruction to be one plus the costs associated with the addressing modes of the operands. This cost corresponds to the length in words of the instruction. Addressing modes involving registers have zero additional cost, while those involving a memory location or constant in them have an additional cost of one because such operands have to be stored in the words following the instruction. Some examples:

The instruction `LD R0, R1` copies the contents of register *R1* into register *R0*. This instruction has a cost of one because no additional memory words are required.

The instruction `LD R0, M` loads the contents of memory location *M* into register *R0*. The cost is two since the address of memory location *M* is in the word following the instruction.

The instruction `LD R1, #100, R2` loads into register *R1* the value given by *contents[contents[100]]*. The cost is two because the constant 100 is stored in the word following the instruction.

CHAPTER 8 CODE GENERATION

In this chapter we assume the cost of a target language program on a given input is the sum of costs of the individual instructions executed when the program is run on that input. Good code generation algorithms seek to minimize the sum of the costs of the instructions executed by the generated target program on typical inputs. We shall see that in some situations we can actually generate optimal code for expressions on certain classes of register machines.

8.2.3 Exercises for Section 8.2

Exercise 8.2.1 Generate code for the following three address statements assuming all variables are stored in memory locations.

- a $x \leftarrow 1$
- b $x \leftarrow a$
- c $x \leftarrow a + 1$
- d $x \leftarrow a + b$
- e The two statements

```
x ← b + c
y ← a + x
```

Exercise 8.2.2 Generate code for the following three address statements assuming a and b are arrays whose elements are 4 byte values.

- a The four statement sequence

```
x ← a[i]
y ← b[j]
a[i] ← y
b[j] ← x
```

- b The three statement sequence

```
x ← a[i]
y ← b[i]
z ← x + y
```

- c The three statement sequence

```
x ← a[i]
y ← b[x]
a[i] ← y
```

8.2 THE TARGET LANGUAGE

Exercise 8.2.3 Generate code for the following three address sequence assuming that *p* and *q* are in memory locations

```
y    q
q    q    4
    p    y
p    p    4
```

Exercise 8.2.4 Generate code for the following sequence assuming that *x*, *y* and *z* are in memory locations

```
    if x    y goto L1
    z    0
    goto L2
L1  z    1
```

Exercise 8.2.5 Generate code for the following sequence assuming that *n* is in a memory location

```
    s    0
    i    0
L1  if i    n goto L2
    s    s    i
    i    i    1
    goto L1
L2
```

Exercise 8.2.6 Determine the costs of the following instruction sequences

```
a      LD R0 y
        LD R1 z
        ADD R0 R0 R1
        ST x R0

b      LD R0 i
        MUL R0 R0 8
        LD R1 a R0
        ST b R1

c      LD R0 c
        LD R1 i
        MUL R1 R1 8
        ST a R1 R0

d      LD R0 p
        LD R1 0 R0
        ST x R1
```

```

e      LD R0  p
        LD R1  x
        ST 0 R0  R1

f      LD  R0  x
        LD  R1  y
        SUB R0  R0  R1
        BLTZ R3  R0

```

8.3 Addresses in the Target Code

In this section we show how names in the IR can be converted into addresses in the target code by looking at code generation for simple procedure calls and returns using static and stack allocation. In Section 7.1 we described how each executing program runs in its own logical address space that was partitioned into four code and data areas.

1. A statically determined area *Code* that holds the executable target code. The size of the target code can be determined at compile time.
2. A statically determined data area *Static* for holding global constants and other data generated by the compiler. The size of the global constants and compiler data can also be determined at compile time.
3. A dynamically managed area *Heap* for holding data objects that are allocated and freed during program execution. The size of the *Heap* cannot be determined at compile time.
4. A dynamically managed area *Stack* for holding activation records as they are created and destroyed during procedure calls and returns. Like the *Heap*, the size of the *Stack* cannot be determined at compile time.

8.3.1 Static Allocation

To illustrate code generation for simplified procedure calls and returns, we shall focus on the following three address statements:

```

call callee
return
halt
action which is a placeholder for other three address statements

```

The size and layout of activation records are determined by the code generator via the information about names stored in the symbol table. We shall first illustrate how to store the return address in an activation record on a procedure

8.3 ADDRESSES IN THE TARGET CODE

call and how to return control to it after the procedure call. For convenience we assume the first location in the activation record holds the return address.

Let us first consider the code needed to implement the simplest case: static allocation. Here a `call callee` statement in the intermediate code can be implemented by a sequence of two target machine instructions:

```
ST  callee staticArea  here  20
BR  callee codeArea
```

The ST instruction saves the return address at the beginning of the activation record for *callee* and the BR transfers control to the target code for the called procedure *callee*. The attribute *callee staticArea* is a constant that gives the address of the beginning of the activation record for *callee* and the attribute *callee codeArea* is a constant referring to the address of the first instruction of the called procedure *callee* in the *Code* area of the run time memory.

The operand *here 20* in the ST instruction is the literal return address: it is the address of the instruction following the BR instruction. We assume that *here* is the address of the current instruction and that the three constants plus the two instructions in the calling sequence have a length of 5 words or 20 bytes.

The code for a procedure ends with a return to the calling procedure, except that the first procedure has no caller, so its final instruction is HALT, which returns control to the operating system. A `return` statement can be implemented by a simple jump instruction:

```
BR  callee staticArea
```

which transfers control to the address saved at the beginning of the activation record for *callee*.

Example 8.3 Suppose we have the following three address code:

```
                                code for c
action1
call p
action2
halt

                                code for p
action3
return
```

Figure 8.4 shows the target program for this three address code. We use the pseudoinstruction ACTION to represent the sequence of machine instructions to execute the statement `action`, which represents three address code that is not relevant for this discussion. We arbitrarily start the code for procedure *c* at address 100 and for procedure *p* at address 200. We assume that each ACTION instruction takes 20 bytes. We further assume that the activation records for these procedures are statically allocated starting at locations 300 and 364 respectively.

The instructions starting at address 100 implement the statements

CHAPTER 8 CODE GENERATION

`action1 call p action2 halt`

of the first procedure *c*. Execution therefore starts with the instruction `ACTION1` at address 100. The `ST` instruction at address 120 saves the return address 140 in the machine status field which is the first word in the activation record of *p*. The `BR` instruction at address 132 transfers control the first instruction in the target code of the called procedure *p*.

100	<code>ACTION₁</code>	code for <i>c</i>
120	<code>ST 364 140</code>	code for <code>action₁</code> save return address 140 in location 364
132	<code>BR 200</code>	call <i>p</i>
140	<code>ACTION₂</code>	
160	<code>HALT</code>	return to operating system
200	<code>ACTION₃</code>	code for <i>p</i>
220	<code>BR 364</code>	return to address saved in location 364
300		300-363 hold activation record for <i>c</i>
304		return address
		local data for <i>c</i>
364		364-451 hold activation record for <i>p</i>
368		return address
		local data for <i>p</i>

Figure 8.4 Target code for static allocation

After executing `ACTION3` the jump instruction at location 220 is executed. Since location 140 was saved at address 364 by the call sequence above, 364 represents 140 when the `BR` statement at address 220 is executed. Therefore when procedure *p* terminates, control returns to address 140 and execution of procedure *c* resumes. □

8.3.2 Stack Allocation

Static allocation can become stack allocation by using relative addresses for storage in activation records. In stack allocation, however, the position of an activation record for a procedure is not known until run time. This position is usually stored in a register, so words in the activation record can be accessed as offsets from the value in this register. The indexed address mode of our target machine is convenient for this purpose.

Relative addresses in an activation record can be taken as offsets from any known position in the activation record, as we saw in Chapter 7. For convenience,

8.3 ADDRESSES IN THE TARGET CODE

nience we shall use positive offsets by maintaining in a register SP a pointer to the beginning of the activation record on top of the stack. When a procedure call occurs, the calling procedure increments SP and transfers control to the called procedure. After control returns to the caller, we decrement SP, thereby deallocating the activation record of the called procedure.

The code for the `rst` procedure initializes the stack by setting SP to the start of the stack area in memory.

LD	SP	<i>stackStart</i>	initialize the stack
code for the <code>rst</code> procedure			
HALT			terminate execution

A procedure call sequence increments SP, saves the return address, and transfers control to the called procedure.

ADD	SP	SP	<i>caller recordSize</i>	increment stack pointer
ST	0	SP	<i>here</i> 16	save return address
BR		<i>callee codeArea</i>		jump to the callee

The operand *caller recordSize* represents the size of an activation record, so the ADD instruction makes SP point to the next activation record. The operand *here* 16 in the ST instruction is the address of the instruction following BR; it is saved in the address pointed to by SP.

The return sequence consists of two parts. The called procedure transfers control to the return address using

BR	0	SP	return to caller
----	---	----	------------------

The reason for using 0 SP in the BR instruction is that we need two levels of indirection. 0 SP is the address of the first word in the activation record, and 0 SP is the return address saved there.

The second part of the return sequence is in the caller, which decrements SP, thereby restoring SP to its previous value. That is, after the subtraction, SP points to the beginning of the activation record of the caller.

SUB	SP	SP	<i>caller recordSize</i>	decrement stack pointer
-----	----	----	--------------------------	-------------------------

Chapter 7 contains a broader discussion of calling sequences and the trade-offs in the division of labor between the calling and called procedures.

Example 8.4 The program in Fig. 8.5 is an abstraction of the quicksort program in the previous chapter. Procedure *q* is recursive, so more than one activation of *q* can be alive at the same time.

Suppose that the sizes of the activation records for procedures *m*, *p*, and *q* have been determined to be *m**size*, *p**size*, and *q**size* respectively. The first word in each activation record will hold a return address. We arbitrarily assume that the code for these procedures starts at addresses 100, 200, and 300, respectively.

CHAPTER 8 CODE GENERATION

```
                                code for m
action1
call q
action2
halt

                                code for p
action3
return

                                code for q
action4
call p
action5
call q
action6
call q
return
```

Figure 8.5 Code for Example 8.4

and that the stack starts at address 600. The target program is shown in Figure 8.6.

We assume that ACTION₄ contains a conditional jump to the address 456 of the return sequence from q; otherwise, the recursive procedure q is condemned to call itself forever.

Let *msize*, *psize*, and *qsize* be 20, 40, and 60, respectively. The first instruction at address 100 initializes the SP to 600, the starting address of the stack. SP holds 620 just before control transfers from m to q, because *msize* is 20. Subsequently, when q calls p, the instruction at address 320 increments SP to 680, where the activation record for p begins. SP reverts to 620 after control returns to q. If the next two recursive calls of q return immediately, the maximum value of SP during this execution is 680. Note, however, that the last stack location used is 739, since the activation record of q starting at location 680 extends for 60 bytes. □

8.3.3 Run Time Addresses for Names

The storage allocation strategy and the layout of local data in an activation record for a procedure determine how the storage for names is accessed. In Chapter 6, we assumed that a name in a three address statement is really a pointer to a symbol table entry for that name. This approach has a significant advantage: it makes the compiler more portable, since the front end need not be changed even when the compiler is moved to a different machine where a different run time organization is needed. On the other hand, generating the specific sequence of access steps while generating intermediate code can be of

8.3 ADDRESSES IN THE TARGET CODE

100	LD SP	600	code for m initialize the stack
108	ACTION ₁		code for action ₁
128	ADD SP	SP <i>msize</i>	call sequence begins
136	ST 0 SP	152	push return address
144	BR	300	call q
152	SUB SP	SP <i>msize</i>	restore SP
160	ACTION ₂		
180	HALT		
200	ACTION ₃		code for p
220	BR 0 SP		return
300	ACTION ₄		code for q contains a conditional jump to 456
320	ADD SP	SP <i>qsize</i>	
328	ST 0 SP	344	push return address
336	BR	200	call p
344	SUB SP	SP <i>qsize</i>	
352	ACTION ₅		
372	ADD SP	SP <i>qsize</i>	
380	ST 0 SP	396	push return address
388	BR	300	call q
396	SUB SP	SP <i>qsize</i>	
404	ACTION ₆		
424	ADD SP	SP <i>qsize</i>	
432	ST 0 SP	440	push return address
440	BR	300	call q
448	SUB SP	SP <i>qsize</i>	
456	BR 0 SP		return
600			stack starts here

Figure 8.6 Target code for stack allocation

significant advantage in an optimizing compiler since it lets the optimizer take advantage of details it would not see in the simple three address statement

In either case names must eventually be replaced by code to access storage locations. We thus consider some elaborations of the simple three address copy statement $x = 0$. After the declarations in a procedure are processed, suppose the symbol table entry for x contains a relative address 12 for x . For example, consider the case in which x is in a statically allocated area beginning at address *static*. Then the actual run time address of x is *static* + 12. Although the compiler can eventually determine the value of *static* + 12 at compile time, the position of the static area may not be known when intermediate code to access the name is generated. In that case, it makes sense to generate three address code to compute *static* + 12 with the understanding that this computation will be carried out during the code generation phase or possibly by the loader before the program runs. The assignment $x = 0$ then translates into

```
static 12  0
```

If the static area starts at address 100, the target code for this statement is

```
LD 112  0
```

8.3.4 Exercises for Section 8.3

Exercise 8.3.1 Generate code for the following three address statements assuming stack allocation where register SP points to the top of the stack

```
call p
call q
return
call r
return
return
```

Exercise 8.3.2 Generate code for the following three address statements assuming stack allocation where register SP points to the top of the stack

```
a x  1
b x  a
c x  a  1
d x  a  b
e The two statements
```

```
x  b  c
y  a  x
```

8.4 BASIC BLOCKS AND FLOW GRAPHS

Exercise 8.3.3 Generate code for the following three address statements again assuming stack allocation and assuming **a** and **b** are arrays whose elements are 4 byte values

- a The four statement sequence

```
x  a i
y  b j
a i  y
b j  x
```

- b The three statement sequence

```
x  a i
y  b i
z  x  y
```

- c The three statement sequence

```
x  a i
y  b x
a i  y
```

8.4 Basic Blocks and Flow Graphs

This section introduces a graph representation of intermediate code that is helpful for discussing code generation even if the graph is not constructed explicitly by a code generation algorithm. Code generation benefits from context. We can do a better job of register allocation if we know how values are defined and used, as we shall see in Section 8.8. We can do a better job of instruction selection by looking at sequences of three address statements, as we shall see in Section 8.9.

The representation is constructed as follows

- 1 Partition the intermediate code into *basic blocks* which are maximal sequences of consecutive three address instructions with the properties that
 - a The flow of control can only enter the basic block through the first instruction in the block. That is, there are no jumps into the middle of the block.
 - b Control will leave the block without halting or branching, except possibly at the last instruction in the block.
- 2 The basic blocks become the nodes of a *flow graph* whose edges indicate which blocks can follow which other blocks.

The Effect of Interrupts

The notion that control once it reaches the beginning of a basic block is certain to continue through to the end requires a bit of thought. There are many reasons why an interrupt not reflected explicitly in the code could cause control to leave the block perhaps never to return. For example, an instruction like $x \leftarrow y \cdot z$ appears not to affect control flow, but if z is 0 it could actually cause the program to abort.

We shall not worry about such possibilities. The reason is as follows. The purpose of constructing basic blocks is to optimize the code. Generally, when an interrupt occurs, either it will be handled and control will come back to the instruction that caused the interrupt, as if control had never deviated, or the program will halt with an error. In the latter case, it doesn't matter how we optimized the code, even if we depended on control reaching the end of the basic block, because the program didn't produce its intended result anyway.

Starting in Chapter 9, we discuss transformations on flow graphs that turn the original intermediate code into optimized intermediate code from which better target code can be generated. The optimized intermediate code is turned into machine code using the code generation techniques in this chapter.

8.4.1 Basic Blocks

Our first job is to partition a sequence of three address instructions into basic blocks. We begin a new basic block with the first instruction and keep adding instructions until we meet either a jump, a conditional jump, or a label on the following instruction. In the absence of jumps and labels, control proceeds sequentially from one instruction to the next. This idea is formalized in the following algorithm.

Algorithm 8.5 Partitioning three address instructions into basic blocks

INPUT A sequence of three address instructions

OUTPUT A list of the basic blocks for that sequence in which each instruction is assigned to exactly one basic block

METHOD First, we determine those instructions in the intermediate code that are *leaders*, that is, the first instructions in some basic block. The instruction just past the end of the intermediate program is not included as a leader. The rules for finding leaders are

1. The first three address instruction in the intermediate code is a leader.

8.4 BASIC BLOCKS AND FLOW GRAPHS

- 2 Any instruction that is the target of a conditional or unconditional jump is a leader
- 3 Any instruction that immediately follows a conditional or unconditional jump is a leader

Then for each leader its basic block consists of itself and all instructions up to but not including the next leader or the end of the intermediate program \square

```

1  i  1
2  j  1
3  t1 10  i
4  t2 t1  j
5  t3 8   t2
6  t4 t3  88
7  a t4  0 0
8  j  j  1
9  if j  10 goto 3
10 i  i  1
11 if i  10 goto 2
12 i  1
13 t5  i  1
14 t6 88  t5
15 a t6  1 0
16 i  i  1
17 if i  10 goto 13

```

Figure 8.7 Intermediate code to set a 10 × 10 matrix to an identity matrix

Example 8.6 The intermediate code in Fig. 8.7 turns a 10 × 10 matrix **a** into an identity matrix. Although it is not important where this code comes from, it might be the translation of the pseudocode in Fig. 8.8. In generating the intermediate code, we have assumed that the real valued array elements take 8 bytes each, and that the matrix **a** is stored in row major form.

```

for i from 1 to 10 do
    for j from 1 to 10 do
        a[i][j] = 0.0
for i from 1 to 10 do
    a[i][i] = 1.0

```

Figure 8.8 Source code for Fig. 8.7

First instruction 1 is a leader by rule 1 of Algorithm 8.5. To find the other leaders we first need to find the jumps. In this example there are three jumps all conditional at instructions 9, 11, and 17. By rule 2 the targets of these jumps are leaders: they are instructions 3, 2, and 13 respectively. Then by rule 3 each instruction following a jump is a leader: those are instructions 10 and 12. Note that no instruction follows 17 in this code, but if there were code following the 18th instruction would also be a leader.

We conclude that the leaders are instructions 1, 2, 3, 10, 12, and 13. The basic block of each leader contains all the instructions from itself until just before the next leader. Thus the basic block of 1 is just 1; for leader 2 the block is just 2; Leader 3, however, has a basic block consisting of instructions 3 through 9 inclusive. Instruction 10's block is 10 and 11; instruction 12's block is just 12; and instruction 13's block is 13 through 17. \square

8.4.2 Next Use Information

Knowing when the value of a variable will be used next is essential for generating good code. If the value of a variable that is currently in a register will never be referenced subsequently, then that register can be assigned to another variable.

The *use* of a name in a three address statement is defined as follows. Suppose three address statement i assigns a value to x . If statement j has x as an operand, and control can flow from statement i to j along a path that has no intervening assignments to x , then we say statement j *uses* the value of x computed at statement i . We further say that x is *live* at statement i .

We wish to determine for each three address statement $x = y \pm z$ what the next uses of x , y , and z are. For the present we do not concern ourselves with uses outside the basic block containing this three address statement.

Our algorithm to determine liveness and next use information makes a backward pass over each basic block. We store the information in the symbol table. We can easily scan a stream of three address statements to find the ends of basic blocks as in Algorithm 8.5. Since procedures can have arbitrary side effects we assume for convenience that each procedure call starts a new basic block.

Algorithm 8.7 Determining the liveness and next use information for each statement in a basic block

INPUT A basic block B of three address statements. We assume that the symbol table initially shows all nontemporary variables in B as being live on exit.

OUTPUT At each statement $i = x = y \pm z$ in B we attach to i the liveness and next use information of x , y , and z .

METHOD We start at the last statement in B and scan backwards to the beginning of B . At each statement $i = x = y \pm z$ in B we do the following:

1. Attach to statement i the information currently found in the symbol table regarding the next use and liveness of x , y , and z .

8.4 BASIC BLOCKS AND FLOW GRAPHS

- 2 In the symbol table set x to “not live” and “no next use”
- 3 In the symbol table set y and z to “live” and the next uses of y and z to i

Here we have used \square as a symbol representing any operator. If the three address statement i is of the form $x \square y$ or $x \square y \square z$ the steps are the same as above ignoring z . Note that the order of steps 2 and 3 may not be interchanged because x may be y or z . \square

8.4.3 Flow Graphs

Once an intermediate code program is partitioned into basic blocks we represent the flow of control between them by a flow graph. The nodes of the flow graph are the basic blocks. There is an edge from block B to block C if and only if it is possible for the first instruction in block C to immediately follow the last instruction in block B . There are two ways that such an edge could be justified.

There is a conditional or unconditional jump from the end of B to the beginning of C .

C immediately follows B in the original order of the three address instructions and B does not end in an unconditional jump.

We say that B is a *predecessor* of C and C is a *successor* of B .

Often we add two nodes called the *entry* and *exit* that do not correspond to executable intermediate instructions. There is an edge from the entry to the first executable node of the flow graph that is to the basic block that comes from the first instruction of the intermediate code. There is an edge to the exit from any basic block that contains an instruction that could be the last executed instruction of the program. If the final instruction of the program is not an unconditional jump then the block containing the final instruction of the program is one predecessor of the exit but so is any basic block that has a jump to code that is not part of the program.

Example 8.8 The set of basic blocks constructed in Example 8.6 yields the flow graph of Fig. 8.9. The entry points to basic block B_1 since B_1 contains the first instruction of the program. The only successor of B_1 is B_2 because B_1 does not end in an unconditional jump and the leader of B_2 immediately follows the end of B_1 .

Block B_3 has two successors. One is itself because the leader of B_3 instruction 3 is the target of the conditional jump at the end of B_3 instruction 9. The other successor is B_4 because control can fall through the conditional jump at the end of B_3 and next enter the leader of B_4 .

Only B_6 points to the exit of the flow graph since the only way to get to code that follows the program from which we constructed the flow graph is to fall through the conditional jump that ends B_6 . \square

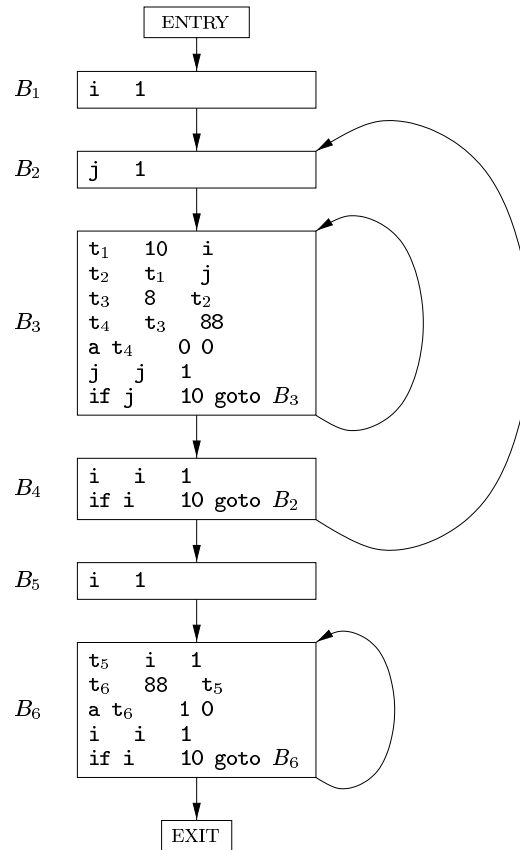


Figure 8.9 Flow graph from Fig. 8.7

8.4.4 Representation of Flow Graphs

First note from Fig. 8.9 that in the flow graph it is normal to replace the jumps to instruction numbers or labels by jumps to basic blocks. Recall that every conditional or unconditional jump is to the leader of some basic block, and it is to this block that the jump will now refer. The reason for this change is that after constructing the flow graph it is common to make substantial changes to the instructions in the various basic blocks. If jumps were to instructions we would have to fix the targets of the jumps every time one of the target instructions was changed.

Flow graphs, being quite ordinary graphs, can be represented by any of the data structures appropriate for graphs. The content of nodes (basic blocks) need their own representation. We might represent the content of a node by a

8.4 BASIC BLOCKS AND FLOW GRAPHS

pointer to the leader in the array of three address instructions together with a count of the number of instructions or a second pointer to the last instruction. However, since we may be changing the number of instructions in a basic block frequently, it is likely to be more efficient to create a linked list of instructions for each basic block.

8.4.5 Loops

Programming language constructs like while statements, do while statements, and for statements naturally give rise to loops in programs. Since virtually every program spends most of its time in executing its loops, it is especially important for a compiler to generate good code for loops. Many code transformations depend upon the identification of loops in a flow graph. We say that a set of nodes L in a flow graph is a *loop* if L contains a node e called the *loop entry* such that

1. e is not ENTRY, the entry of the entire flow graph.
2. No node in L besides e has a predecessor outside L . That is, every path from ENTRY to any node in L goes through e .
3. Every node in L has a nonempty path, completely within L , to e .

Example 8.9 The flow graph of Fig. 8.9 has three loops.

1. B_3 by itself
2. B_6 by itself
3. $\{B_2, B_3, B_4\}$

The first two are single nodes with an edge to the node itself. For instance, B_3 forms a loop with B_3 as its entry. Note that the last requirement for a loop is that there be a nonempty path from B_3 to itself. Thus, a single node like B_2 , which does not have an edge $B_2 \rightarrow B_2$, is not a loop, since there is no nonempty path from B_2 to itself within $\{B_2\}$.

The third loop $L = \{B_2, B_3, B_4\}$ has B_2 as its loop entry. Note that among these three nodes, only B_2 has a predecessor B_1 that is not in L . Further, each of the three nodes has a nonempty path to B_2 staying within L . For instance, B_2 has the path $B_2 \rightarrow B_3 \rightarrow B_4 \rightarrow B_2$. \square

8.4.6 Exercises for Section 8.4

Exercise 8.4.1 Figure 8.10 is a simple matrix multiplication program.

- a. Translate the program into three address statements of the type we have been using in this section. Assume the matrix entries are numbers that require 8 bytes, and that matrices are stored in row major order.

- b Construct the flow graph for your code from a
- c Identify the loops in your flow graph from b

```

for i 0 i n i
  for j 0 j n j
    c i j 0 0
for i 0 i n i
  for j 0 j n j
    for k 0 k n k
      c i j c i j a i k b k j

```

Figure 8.10 A matrix multiplication algorithm

Exercise 8.4.2 Figure 8.11 is code to count the number of primes from 2 to n using the sieve method on a suitably large array a . That is $a[i]$ is TRUE at the end only if there is no prime \sqrt{i} or less that evenly divides i . We initialize all $a[i]$ to TRUE and then set $a[j]$ to FALSE if we find a divisor of j .

- a Translate the program into three address statements of the type we have been using in this section. Assume integers require 4 bytes
- b Construct the flow graph for your code from a
- c Identify the loops in your flow graph from b

```

for i 2 i n i
  a i TRUE
count 0
s sqrt n
for i 2 i s i
  if a i i has been found to be a prime
    count
  for j 2 i j n j j i
    a j FALSE no multiple of i is a prime

```

Figure 8.11 Code to sieve for primes

8.5 Optimization of Basic Blocks

We can often obtain a substantial improvement in the running time of code merely by performing *local* optimization within each basic block by itself. More thorough *global* optimization, which looks at how information flows among the basic blocks of a program, is covered in later chapters, starting with Chapter 9. It is a complex subject, with many different techniques to consider.

8.5.1 The DAG Representation of Basic Blocks

Many important techniques for local optimization begin by transforming a basic block into a DAG (directed acyclic graph). In Section 6.1.1, we introduced the DAG as a representation for single expressions. The idea extends naturally to the collection of expressions that are created within one basic block. We construct a DAG for a basic block as follows:

1. There is a node in the DAG for each of the initial values of the variables appearing in the basic block.
2. There is a node N associated with each statement s within the block. The children of N are those nodes corresponding to statements that are the last definitions prior to s of the operands used by s .
3. Node N is labeled by the operator applied at s , and also attached to N is the list of variables for which it is the last definition within the block.
4. Certain nodes are designated *output nodes*. These are the nodes whose variables are *live on exit* from the block, that is, their values may be used later in another block of the flow graph. Calculation of these *live variables* is a matter for global flow analysis, discussed in Section 9.2.5.

The DAG representation of a basic block lets us perform several code-improving transformations on the code represented by the block:

- a. We can eliminate *local common subexpressions*, that is, instructions that compute a value that has already been computed.
- b. We can eliminate *dead code*, that is, instructions that compute a value that is never used.
- c. We can reorder statements that do not depend on one another; such reordering may reduce the time a temporary value needs to be preserved in a register.
- d. We can apply algebraic laws to reorder operands of three-address instructions, and sometimes thereby simplify the computation.

8.5.2 Finding Local Common Subexpressions

Common subexpressions can be detected by noticing, as a new node M is about to be added, whether there is an existing node N with the same children, in the same order, and with the same operator. If so, N computes the same value as M and may be used in its place. This technique was introduced as the “value number” method of detecting common subexpressions in Section 6.1.1.

Example 8.10 A DAG for the block

```

a  b  c
b  a  d
c  b  c
d  a  d

```

is shown in Fig. 8.12. When we construct the node for the third statement $c = b + c$, we know that the use of b in $b + c$ refers to the node of Fig. 8.12 labeled b , because that is the most recent definition of b . Thus, we do not confuse the values computed at statements one and three.

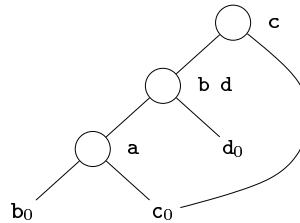


Figure 8.12 DAG for basic block in Example 8.10

However, the node corresponding to the fourth statement $d = a + d$ has the operator $+$ and the nodes with attached variables a and d_0 as children. Since the operator and the children are the same as those for the node corresponding to statement two, we do not create this node, but add d to the list of definitions for the node labeled b . □

It might appear that, since there are only three nonleaf nodes in the DAG of Fig. 8.12, the basic block in Example 8.10 can be replaced by a block with only three statements. In fact, if b is not live on exit from the block, then we do not need to compute that variable, and can use d to receive the value represented by the node labeled b in Fig. 8.12. The block then becomes

```

a  b  c
d  a  d
c  d  c

```

8.5 OPTIMIZATION OF BASIC BLOCKS

However, if both **b** and **d** are live on exit, then a fourth statement must be used to copy the value from one to the other¹.

Example 8.11 When we look for common subexpressions, we really are looking for expressions that are guaranteed to compute the same value, no matter how that value is computed. Thus, the DAG method will miss the fact that the expression computed by the first and fourth statements in the sequence

a	b	c
b	b	d
c	c	d
e	b	c

is the same, namely $b_0 + c_0$. That is, even though **b** and **c** both change between the first and last statements, their sum remains the same, because $b = c = b + d - c = d$. The DAG for this sequence is shown in Fig. 8.13, but does not exhibit any common subexpressions. However, algebraic identities applied to the DAG, as discussed in Section 8.5.4, may expose the equivalence. \square

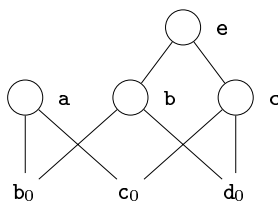


Figure 8.13 DAG for basic block in Example 8.11

8.5.3 Dead Code Elimination

The operation on DAGs that corresponds to dead code elimination can be implemented as follows. We delete from a DAG any root node with no ancestors that has no live variables attached. Repeated application of this transformation will remove all nodes from the DAG that correspond to dead code.

Example 8.12 If in Fig. 8.13, **a** and **b** are live but **c** and **e** are not, we can immediately remove the root labeled **e**. Then, the node labeled **c** becomes a root and can be removed. The roots labeled **a** and **b** remain, since they each have live variables attached. \square

¹In general, we must be careful when reconstructing code from DAGs: how we choose the names of variables. If a variable is defined twice, or if it is assigned once and the initial value v_0 is also used, then we must make sure that we do not change the value of v_0 until we have made all uses of the node whose value v_0 previously held.

8.5.4 The Use of Algebraic Identities

Algebraic identities represent another important class of optimizations on basic blocks. For example, we may apply arithmetic identities such as

$$\begin{array}{rcl} x & - & 0 \\ x & + & 1 \end{array} \quad \begin{array}{rcl} 0 & + & x \\ 1 & + & x \end{array} \quad \begin{array}{rcl} x & - & x \\ x & + & x \end{array} \quad \begin{array}{rcl} x & - & 0 \\ x & + & 1 \end{array} \quad \begin{array}{rcl} x & & x \end{array}$$

to eliminate computations from a basic block.

Another class of algebraic optimizations includes local *reduction in strength*, that is, replacing a more expensive operator by a cheaper one as in

EXPENSIVE	CHEAPER
x^2	$x \times x$
$2 \times x$	$x + x$
$x \times 2$	$x \times 0.5$

A third class of related optimizations is *constant folding*. Here we evaluate constant expressions at compile time and replace the constant expressions by their values.² Thus the expression 2×3.14 would be replaced by 6.28 . Many constant expressions arise in practice because of the frequent use of symbolic constants in programs.

The DAG construction process can help us apply these and other more general algebraic transformations such as commutativity and associativity. For example, suppose the language reference manual specifies that \times is commutative, that is, $x \times y = y \times x$. Before we create a new node labeled \times with left child M and right child N , we always check whether such a node already exists. However, because \times is commutative, we should then check for a node having operator \times , left child N , and right child M .

The relational operators such as $<$ and $>$ sometimes generate unexpected common subexpressions. For example, the condition $x < y$ can also be tested by subtracting the arguments and performing a test on the condition code set by the subtraction.³ Thus, only one node of the DAG may need to be generated for $x < y$ and $x > y$.

Associative laws might also be applicable to expose common subexpressions. For example, if the source code has the assignments

```
a = b + c
e = c + d + b
```

the following intermediate code might be generated

²Arithmetic expressions should be evaluated the same way at compile time as they are at run time. K. Thompson has suggested an elegant solution to constant folding: compile the constant expression, execute the target code on the spot, and replace the expression with the result. Thus, the compiler does not need to contain an interpreter.

³The subtraction can, however, introduce overflows and underflows while a compare instruction would not.

8.5 OPTIMIZATION OF BASIC BLOCKS

```

a   b   c
t   c   d
e   t   b

```

If `t` is not needed outside this block we can change this sequence to

```

a   b   c
e   a   d

```

using both the associativity and commutativity of

The compiler writer should examine the language reference manual carefully to determine what rearrangements of computations are permitted since because of possible overflows or underflows computer arithmetic does not always obey the algebraic identities of mathematics. For example the Fortran standard states that a compiler may evaluate any mathematically equivalent expression provided that the integrity of parentheses is not violated. Thus a compiler may evaluate $x \ y \ x \ z$ as $x \ y \ z$ but it may not evaluate $a \ b \ c$ as $a \ b \ c$. A Fortran compiler must therefore keep track of where parentheses were present in the source language expressions if it is to optimize programs in accordance with the language definition.

8.5.5 Representation of Array References

At first glance it might appear that the array indexing instructions can be treated like any other operator. Consider for instance the sequence of three address statements

```

x   a[i]
a[j]   y
z   a[i]

```

If we think of `a[i]` as an operation involving `a` and `i` similar to $a \cdot i$ then it might appear as if the two uses of `a[i]` were a common subexpression. In that case we might be tempted to optimize by replacing the third instruction `z = a[i]` by the simpler `z = x`. However since `j` could equal `i` the middle statement may in fact change the value of `a[i]` thus it is not legal to make this change.

The proper way to represent array accesses in a DAG is as follows

1. An assignment from an array like `x = a[i]` is represented by creating a node with operator `=` and two children representing the initial value of the array `a0` in this case and the index `i`. Variable `x` becomes a label of this new node.
2. An assignment to an array like `a[j] = y` is represented by a new node with operator `=` and three children representing `a0`, `j` and `y`. There is no variable labeling this node. What is different is that the creation of

this node *kills* all currently constructed nodes whose value depends on a_0 . A node that has been killed cannot receive any more labels—that is, it cannot become a common subexpression.

Example 8 13 The DAG for the basic block

```

x ← a[i]
a[j] ← y
z ← a[i]

```

is shown in Fig. 8 14. The node N for x is created first, but when the node labeled z is created, N is killed. Thus, when the node for z is created, it cannot be identified with N , and a new node with the same operands a_0 and i_0 must be created instead. \square

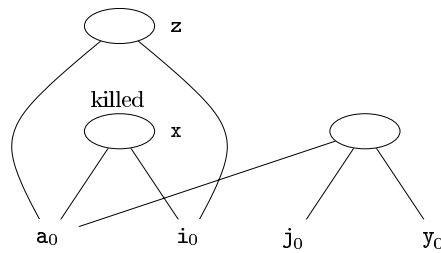


Figure 8 14 The DAG for a sequence of array assignments

Example 8 14 Sometimes a node must be killed even though none of its children have an array like a_0 in Example 8 13 as attached variable. Likewise a node can kill if it has a descendant that is an array, even though none of its children are array nodes. For instance, consider the three address code

```

b ← 12 * a
x ← b[i]
b[j] ← y

```

What is happening here is that, for efficiency reasons, b has been defined to be a position in an array a . For example, if the elements of a are four bytes long, then b represents the fourth element of a . If j and i represent the same value, then $b[i]$ and $b[j]$ represent the same location. Therefore it is important to have the third instruction $b[j] \leftarrow y$ kill the node with x as its attached variable. However, as we see in Fig. 8 15, both the killed node and the node that does the killing have a_0 as a grandchild, not as a child. \square

8.5 OPTIMIZATION OF BASIC BLOCKS

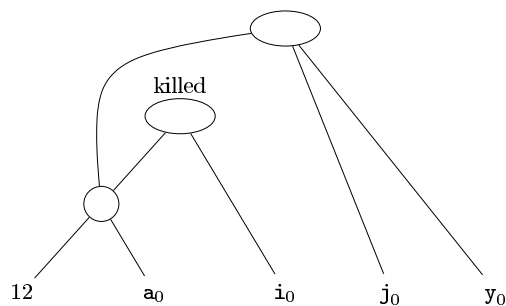


Figure 8.15 A node that kills a use of an array need not have that array as a child

8.5.6 Pointer Assignments and Procedure Calls

When we assign indirectly through a pointer as in the assignments

$$\begin{array}{ll} x & \leftarrow p \\ q & \leftarrow y \end{array}$$

we do not know what p or q point to. In effect $x \leftarrow p$ is a use of every variable whatsoever and $q \leftarrow y$ is a possible assignment to every variable. As a consequence the operator must take all nodes that are currently associated with identifiers as arguments which is relevant for dead code elimination. More importantly the operator kills all other nodes so far constructed in the DAG.

There are global pointer analyses one could perform that might limit the set of variables a pointer could reference at a given place in the code. Even local analysis could restrict the scope of a pointer. For instance in the sequence

$$\begin{array}{ll} p & \leftarrow x \\ p & \leftarrow y \end{array}$$

we know that x and no other variable is given the value of y so we don't need to kill any node but the node to which x was attached.

Procedure calls behave much like assignments through pointers. In the absence of global data flow information we must assume that a procedure uses and changes any data to which it has access. Thus if procedure P is in the scope of variable x a call to P both uses the node with attached variable x and kills that node.

8.5.7 Reassembling Basic Blocks From DAG's

After we perform whatever optimizations are possible while constructing the DAG or by manipulating the DAG once constructed we may reconstitute the three address code for the basic block from which we built the DAG. For each

node that has one or more attached variables we construct a three address statement that computes the value of one of those variables. We prefer to compute the result into a variable that is live on exit from the block. However, if we do not have global live variable information to work from we need to assume that every variable of the program but not temporaries that are generated by the compiler to process expressions is live on exit from the block.

If the node has more than one live variable attached then we have to introduce copy statements to give the correct value to each of those variables. Sometimes global optimization can eliminate those copies if we can arrange to use one of two variables in place of the other.

Example 8 15 Recall the DAG of Fig. 8 12. In the discussion following Example 8 10 we decided that if *b* is not live on exit from the block then the three statements

```
a  b  c
d  a  d
c  d  c
```

suffice to reconstruct the basic block. The third instruction *c = d - c* must use *d* as an operand rather than *b* because the optimized block never computes *b*.

If both *b* and *d* are live on exit or if we are not sure whether or not they are live on exit then we need to compute *b* as well as *d*. We can do so with the sequence

```
a  b  c
d  a  d
b  d
c  d  c
```

This basic block is still more efficient than the original. Although the number of instructions is the same we have replaced a subtraction by a copy which tends to be less expensive on most machines. Further, it may be that by doing a global analysis we can eliminate the use of this computation of *b* outside the block by replacing it by uses of *d*. In that case we can come back to this basic block and eliminate *b = d* later. Intuitively we can eliminate this copy if wherever this value of *b* is used *d* is still holding the same value. That situation may or may not be true depending on how the program recomputes *d*. □

When reconstructing the basic block from a DAG we not only need to worry about what variables are used to hold the values of the DAG's nodes but we also need to worry about the order in which we list the instructions computing the values of the various nodes. The rules to remember are

1. The order of instructions must respect the order of nodes in the DAG. That is we cannot compute a node's value until we have computed a value for each of its children.

8.5 OPTIMIZATION OF BASIC BLOCKS

- 2 Assignments to an array must follow all previous assignments to or evaluations from the same array according to the order of these instructions in the original basic block
- 3 Evaluations of array elements must follow any previous assignments to the same array. The only permutation allowed is that two evaluations from the same array may be done in either order as long as neither crosses over an assignment to that array
- 4 Any use of a variable must follow all previous assignments to the original block, procedure calls or indirect assignments through a pointer
- 5 Any procedure call or indirect assignment through a pointer must follow all previous evaluations of any variable according to the original block

That is, when reordering code, no statement may cross a procedure call or assignment through a pointer, and uses of the same array may cross each other only if both are array accesses, but not assignments to elements of the array.

8.5.8 Exercises for Section 8.5

Exercise 8.5.1 Construct the DAG for the basic block

```

d  b  c
e  a  b
b  b  c
a  e  d

```

Exercise 8.5.2 Simplify the three address code of Exercise 8.5.1 assuming

- a Only a is live on exit from the block
- b a , b and c are live on exit from the block

Exercise 8.5.3 Construct the DAG for the code in block B_6 of Fig. 8.9. Do not forget to include the comparison $i = 10$.

Exercise 8.5.4 Construct the DAG for the code in block B_3 of Fig. 8.9

Exercise 8.5.5 Extend Algorithm 8.7 to process three statements of the form

```

a  a[i]  b
b  a  b[i]
c  a  b
c  a  b

```

Exercise 8.5.6 Construct the DAG for the basic block

```

a i    b
p      c
d      a j
e      p
p      a i

```

on the assumption that

- a p can point anywhere
- b p can point only to b or d

Exercise 8 5 7 If a pointer or array expression such as `a i` or `p` is assigned and then used without the possibility of being changed in the interim we can take advantage of the situation to simplify the DAG. For example in the code of Exercise 8 5 6 if `p` cannot point to `d` then the fourth statement `e = p` can be replaced by `e = c`. Revise the DAG construction algorithm to take advantage of such situations and apply your algorithm to the code of Exercise 8 5 6.

Exercise 8 5 8 Suppose a basic block is formed from the C assignment statements

```

x = a + b + c + d + e + f
y = a + c + e

```

- a Give the three address statements (only one addition per statement) for this block.
- b Use the associative and commutative laws to modify the block to use the fewest possible number of instructions assuming both `x` and `y` are live on exit from the block.

8 6 A Simple Code Generator

In this section we shall consider an algorithm that generates code for a single basic block. It considers each three address instruction in turn and keeps track of what values are in what registers so it can avoid generating unnecessary loads and stores.

One of the primary issues during code generation is deciding how to use registers to best advantage. There are four principal uses of registers.

In most machine architectures some or all of the operands of an operation must be in registers in order to perform the operation.

Registers make good temporaries — places to hold the result of a subexpression while a larger expression is being evaluated — or more generally a place to hold a variable that is used only within a single basic block.

8.6 A SIMPLE CODE GENERATOR

Registers are used to hold *global* values that are computed in one basic block and used in other blocks—for example, a loop index that is incremented going around the loop and is used several times within the loop.

Registers are often used to help with run-time storage management—for example, to manage the run-time stack, including the maintenance of stack pointers and possibly the top elements of the stack itself.

These are competing needs, since the number of registers available is limited.

The algorithm in this section assumes that some set of registers is available to hold the values that are used within the block. Typically, this set of registers does not include all the registers of the machine, since some registers are reserved for global variables and managing the stack. We assume that the basic block has already been transformed into a preferred sequence of three-address instructions by transformations such as combining common subexpressions. We further assume that for each operator, there is exactly one machine instruction that takes the necessary operands in registers and performs that operation, leaving the result in a register. The machine instructions are of the form

LD *reg mem*

ST *mem reg*

OP *reg reg reg*

8.6.1 Register and Address Descriptors

Our code generation algorithm considers each three-address instruction in turn and decides what loads are necessary to get the needed operands into registers. After generating the loads, it generates the operation itself. Then, if there is a need to store the result into a memory location, it also generates that store.

In order to make the needed decisions, we require a data structure that tells us what program variables currently have their value in a register, and which register or registers, if so. We also need to know whether the memory location for a given variable currently has the proper value for that variable, since a new value for the variable may have been computed in a register and not yet stored. The desired data structure has the following descriptors:

- 1 For each available register, a *register descriptor* keeps track of the variable names whose current value is in that register. Since we shall use only those registers that are available for local use within a basic block, we assume that initially all register descriptors are empty. As the code generation progresses, each register will hold the value of zero or more names.
- 2 For each program variable, an *address descriptor* keeps track of the location or locations where the current value of that variable can be found. The location might be a register, a memory address, a stack location, or some set of more than one of these. The information can be stored in the symbol table entry for that variable name.

8.6.2 The Code Generation Algorithm

An essential part of the algorithm is a function *getReg* I which selects registers for each memory location associated with the three address instruction I . Function *getReg* has access to the register and address descriptors for all the variables of the basic block and may also have access to certain useful data flow information such as the variables that are live on exit from the block. We shall discuss *getReg* after presenting the basic algorithm. While we do not know the total number of registers available for local data belonging to a basic block we assume that there are enough registers so that after freeing all available registers by storing their values in memory there are enough registers to accomplish any three address operation.

In a three address instruction such as $x \leftarrow y \oplus z$ we shall treat \oplus as a generic operator and ADD as the equivalent machine instruction. We do not therefore take advantage of commutativity of \oplus . Thus when we implement the operation the value of y must be in the second register mentioned in the ADD instruction never the third. A possible improvement to the algorithm is to generate code for both $x \leftarrow y \oplus z$ and $x \leftarrow z \oplus y$ whenever \oplus is a commutative operator and pick the better code sequence.

Machine Instructions for Operations

For a three address instruction such as $x \leftarrow y \oplus z$ do the following

- 1 Use *getReg* $x \leftarrow y \oplus z$ to select registers for x , y and z . Call these R_x , R_y and R_z .
- 2 If y is not in R_y according to the register descriptor for R_y then issue an instruction LD $R_y \leftarrow y'$ where y' is one of the memory locations for y according to the address descriptor for y .
- 3 Similarly if z is not in R_z issue an instruction LD $R_z \leftarrow z'$ where z' is a location for z .
- 4 Issue the instruction ADD $R_x \leftarrow R_y \oplus R_z$.

Machine Instructions for Copy Statements

There is an important special case a three address copy statement of the form $x \leftarrow y$. We assume that *getReg* will always choose the same register for both x and y . If y is not already in that register R_y then generate the machine instruction LD $R_y \leftarrow y$. If y was already in R_y we do nothing. It is only necessary that we adjust the register descriptor for R_y so that it includes x as one of the values found there.

Ending the Basic Block

As we have described the algorithm, variables used by the block may wind up with their only location being a register. If the variable is a temporary used only within the block—that is, one when the block ends, we can forget about the value of the temporary and assume its register is empty. However, if the variable is live on exit from the block, or if we don't know which variables are live on exit, then we need to assume that the value of the variable is needed later. In that case, for each variable x whose address descriptor does not say that its value is located in the memory location for x , we must generate the instruction $ST\ x\ R$, where R is a register in which x 's value exists at the end of the block.

Managing Register and Address Descriptors

As the code generation algorithm issues load, store, and other machine instructions, it needs to update the register and address descriptors. The rules are as follows:

1. For the instruction $LD\ R\ x$
 - a. Change the register descriptor for register R so it holds only x .
 - b. Change the address descriptor for x by adding register R as an additional location.
2. For the instruction $ST\ x\ R$, change the address descriptor for x to include its own memory location.
3. For an operation such as $ADD\ R_x\ R_y\ R_z$ implementing a three-address instruction $x = y + z$
 - a. Change the register descriptor for R_x so that it holds only x .
 - b. Change the address descriptor for x so that its only location is R_x . Note that the memory location for x is *not* now in the address descriptor for x .
 - c. Remove R_x from the address descriptor of any variable other than x .
4. When we process a copy statement $x = y$ after generating the load for y into register R_y , if needed, and after managing descriptors as for all load statements, per rule 1
 - a. Add x to the register descriptor for R_y .
 - b. Change the address descriptor for x so that its only location is R_y .

Example 8.16 Let us translate the basic block consisting of the three-address statements

```

t  a  b
u  a  c
v  t  u
a  d
d  v  u

```

Here we assume that `t`, `u` and `v` are temporaries local to the block while `a`, `b`, `c` and `d` are variables that are live on exit from the block. Since we have not yet discussed how the function *getReg* might work, we shall simply assume that there are as many registers as we need, but that when a register's value is no longer needed, for example, it holds only a temporary, all of whose uses have been passed, then we reuse its register.

A summary of all the machine code instructions generated is in Fig. 8.16. The figure also shows the register and address descriptors before and after the translation of each three address instruction.

		R1	R2	R3	a	b	c	d	t	u	v
					a	b	c	d			
t	a b										
	LD R1 a										
	LD R2 b										
	SUB R2 R1 R2										
		a	t		a R1	b	c	d	R2		
u	a c										
	LD R3 c										
	SUB R1 R1 R3										
		u	t	c	a	b	c R3	d	R2	R1	
v	t u										
	ADD R3 R2 R1										
		u	t	v	a	b	c	d	R2	R1	R3
a	d										
	LD R2 d										
		u	a d	v	R2	b	c	d R2		R1	R3
d	v u										
	ADD R1 R3 R1										
		d	a	v	R2	b	c	R1			R3
exit											
	ST a R2										
	ST d R1										
		d	a	v	a R2	b	c	d R1			R3

Figure 8.16 Instructions generated and the changes in the register and address descriptors

For the first three address instruction `t = a + b` we need to issue three instructions, since nothing is in a register initially. Thus, we see `a` and `b` loaded

8.6 A SIMPLE CODE GENERATOR

into registers R1 and R2 and the value t produced in register R2. Notice that we can use R2 for t because the value b previously in R2 is not needed within the block. Since b is presumably live on exit from the block had it not been in its own memory location—as indicated by its address descriptor—we would have had to store R2 into b first. The decision to do so had we needed R2 would be taken by *getReg*.

The second instruction $u = a + c$ does not require a load of a since it is already in register R1. Further, we can reuse R1 for the result u since the value of a previously in that register is no longer needed within the block and its value is in its own memory location if a is needed outside the block. Note that we change the address descriptor for a to indicate that it is no longer in R1 but is in the memory location called a .

The third instruction $v = t + u$ requires only the addition. Further, we can use R3 for the result v since the value of c in that register is no longer needed within the block and c has its value in its own memory location.

The copy instruction $a = d$ requires a load of d since it is not in a register. We show register R2's descriptor holding both a and d . The addition of a to the register descriptor is the result of our processing the copy statement and is not the result of any machine instruction.

The fifth instruction $d = v + u$ uses two values that are in registers. Since u is a temporary whose value is no longer needed, we have chosen to reuse its register R1 for the new value of d . Notice that d is now in only R1 and is not in its own memory location. The same holds for a which is in R2 and not in the memory location called a . As a result, we need a coda to the machine code for the basic block that stores the live on exit variables a and d into their memory locations. We show these as the last two instructions. \square

8.6.3 Design of the Function *getReg*

Lastly, let us consider how to implement *getReg I* for a three address instruction I . There are many options, although there are also some absolute prohibitions against choices that lead to incorrect code due to the loss of the value of one or more live variables. We begin our examination with the case of an operation step for which we again use $x = y + z$ as the generic example. First, we must pick a register for y and a register for z . The issues are the same so we shall concentrate on picking register R_y for y . The rules are as follows:

- 1 If y is currently in a register, pick a register already containing y as R_y . Do not issue a machine instruction to load this register, as none is needed.
- 2 If y is not in a register but there is a register that is currently empty, pick one such register as R_y .
- 3 The difficult case occurs when y is not in a register and there is no register that is currently empty. We need to pick one of the allowable registers anyway and we need to make it safe to reuse. Let R be a candidate