A Tour through the UNIX† C Compiler

D. M. Ritchie

Bell Laboratories
Murray Hill, New Jersey 07974

The Intermediate Language

Communication between the two phases of the compiler proper is carried out by means of a pair of intermediate files. These files are treated as having identical structure, although the second file contains only the code generated for strings. It is convenient to write strings out separately to reduce the need for multiple location counters in a later assembly phase.

The intermediate language is not machine-independent; its structure in a number of ways reflects the fact that C was originally a one-pass compiler chopped in two to reduce the maximum memory requirement. In fact, only the latest version of the compiler has a complete intermediate language at all. Until recently, the first phase of the compiler generated assembly code for those constructions it could deal with, and passed expression parse trees, in absolute binary form, to the second phase for code generation. Now, at least, all inter-phase information is passed in a describable form, and there are no absolute pointers involved, so the coupling between the phases is not so strong.

The areas in which the machine (and system) dependencies are most noticeable are

1. Storage allocation for automatic variables and arguments has already been performed, and nodes for such variables refer to them by offset from a display pointer. Type conversion (for example, from integer to pointer) has already occurred using the assumption of byte addressing and 2-byte words.
2. Data representations suitable to the PDP-11 are assumed; in particular, floating point constants are passed as four words in the machine representation.

As it happens, each intermediate file is represented as a sequence of binary numbers without any explicit demarcations. It consists of a sequence of conceptual lines, each headed by an operator, and possibly containing various operands. The operators are small numbers; to assist in recognizing failure in synchronization, the high-order byte of each operator word is always the octal number 376. Operands are either 16-bit binary numbers or strings of characters representing names. Each name is terminated by a null character. There is no alignment requirement for numerical operands and so there is no padding after a name string.

The binary representation was chosen to avoid the necessity of converting to and from character form and to minimize the size of the files. It would be very easy to make each operator-operand 'line' in the file be a genuine, printable line, with the numbers in octal or decimal; this in fact was the representation originally used.

The operators fall naturally into two classes: those which represent part of an expression, and all others. Expressions are transmitted in a reverse-Polish notation; as they are being read, a tree is built which is isomorphic to the tree constructed in the first phase. Expressions are passed as a whole, with no non-expression operators intervening. The reader maintains a stack; each leaf of the expression tree (name, constant) is pushed on the stack; each unary operator replaces the top of the stack by a node whose operand is the old top-of-stack; each binary operator replaces the top pair on the stack with a single entry. When the expression is complete there is exactly one item on the stack. Following each expression is a special

†UNIX is a Trademark of Bell Laboratories.

operator which passes the unique previous expression to the ‘optimizer’ described below and then to the code generator.

Here is the list of operators not themselves part of expressions.

EOF

marks the end of an input file.

BDATA *flag data ...*

specifies a sequence of bytes to be assembled as static data. It is followed by pairs of words; the first member of the pair is non-zero to indicate that the data continue; a zero flag is not followed by data and terminates the operator. The data bytes occupy the low-order part of a word.

WDATA *flag data ...*

specifies a sequence of words to be assembled as static data; it is identical to the BDATA operator except that entire words, not just bytes, are passed.

PROG

means that subsequent information is to be compiled as program text.

DATA

means that subsequent information is to be compiled as static data.

BSS

means that subsequent information is to be compiled as unitialized static data.

SYMDEF *name*

means that the symbol *name* is an external name defined in the current program. It is produced for each external data or function definition.

CSPACE *name size*

indicates that the name refers to a data area whose size is the specified number of bytes. It is produced for external data definitions without explicit initialization.

SSPACE *size*

indicates that *size* bytes should be set aside for data storage. It is used to pad out short initializations of external data and to reserve space for static (internal) data. It will be preceded by an appropriate label.

EVEN

is produced after each external data definition whose size is not an integral number of words. It is not produced after strings except when they initialize a character array.

NLABEL *name*

is produced just before a BDATA or WDATA initializing external data, and serves as a label for the data.

RLABEL *name*

is produced just before each function definition, and labels its entry point.

SNAME *name number*

is produced at the start of each function for each static variable or label declared therein. Subsequent uses of the variable will be in terms of the given number. The code generator uses this only to produce a debugging symbol table.

ANAME *name number*

Likewise, each automatic variable's name and stack offset is specified by this operator. Arguments count as automatics.

RNAME *name number*

Each register variable is similarly named, with its register number.

SAVE *number*

produces a register-save sequence at the start of each function, just after its label (RLABEL).

SETREG *number*

is used to indicate the number of registers used for register variables. It actually gives the register number of the lowest free register; it is redundant because the RNAME operators could be counted instead.

PROFIL

is produced before the save sequence for functions when the profile option is turned on. It produces code to count the number of times the function is called.

SWIT *deflab line label value ...*

is produced for switches. When control flows into it, the value being switched on is in the register forced by RFORCE (below). The switch statement occurred on the indicated line of the source, and the label number of the default location is *deflab.* Then the operator is followed by a sequence of label-number and value pairs; the list is terminated by a 0 label.

LABEL *number*

generates an internal label. It is referred to elsewhere using the given number.

BRANCH *number*

indicates an unconditional transfer to the internal label number given.

RETRN

produces the return sequence for a function. It occurs only once, at the end of each function.

EXPR *line*

causes the expression just preceding to be compiled. The argument is the line number in the source where the expression occurred.

NAME *class type name*

NAME *class type number*

indicates a name occurring in an expression. The first form is used when the name is external; the second when the name is automatic, static, or a register. Then the number indicates the stack offset, the label number, or the register number as appropriate. Class and type encoding is described elsewhere.

CON *type value*

transmits an integer constant. This and the next two operators occur as part of expressions.

FCON *type 4-word-value*

transmits a floating constant as four words in PDP-11 notation.

SFCON *type value*

transmits a floating-point constant whose value is correctly represented by its high-order word in PDP-11 notation.

NULL

indicates a null argument list of a function call in an expression; call is a binary operator whose second operand is the argument list.

CBRANCH *label cond*

produces a conditional branch. It is an expression operator, and will be followed by an EXPR. The branch to the label number takes place if the expression's truth value is the same as that of *cond.* That is, if *cond=1* and the expression evaluates to true, the branch is taken.

binary-operator *type*

There are binary operators corresponding to each such source-language operator; the type of the result of each is passed as well. Some perhaps-unexpected ones are: COMMA, which is a right-associative operator designed to simplify right-to-left evaluation of function arguments; prefix and postfix ++ and ––, whose second operand is the increment amount, as a CON; QUEST and COLON, to express the conditional expression as 'a?(b:c)'; and a sequence of special operators for expressing relations between pointers, in case pointer comparison is different from integer comparison (e.g. unsigned).

unary-operator *type*

There are also numerous unary operators. These include ITOF, FTOI, FTOL, LTOF, ITOL, LTOI which convert among floating, long, and integer; JUMP which branches indirectly through a label expression; INIT, which compiles the value of a constant expression used as an initializer; RFORCE, which is used before a return sequence or a switch to place a value in an agreed-upon register.

Expression Optimization

Each expression tree, as it is read in, is subjected to a fairly comprehensive analysis. This is performed by the *optim* routine and a number of subroutines; the major things done are

1. Modifications and simplifications of the tree so its value may be computed more efficiently and conveniently by the code generator.
2. Marking each interior node with an estimate of the number of registers required to evaluate it. This register count is needed to guide the code generation algorithm.

One thing that is definitely not done is discovery or exploitation of common subexpressions, nor is this done anywhere in the compiler.

The basic organization is simple: a depth-first scan of the tree. *Optim* does nothing for leaf nodes (except for automatics; see below), and calls *unoptim* to handle unary operators. For binary operators, it calls itself to process the operands, then treats each operator separately. One important case is commutative and associative operators, which are handled by *acommute.*

Here is a brief catalog of the transformations carried out by by *optim* itself. It is not intended to be complete. Some of the transformations are machine-dependent, although they may well be useful on machines other than the PDP-11.

1. As indicated in the discussion of *unoptim* below, the optimizer can create a node type corresponding to the location addressed by a register plus a constant offset. Since this is precisely the implementation of automatic variables and arguments, where the register is fixed by convention, such variables are changed to the new form to simplify later processing.
2. Associative and commutative operators are processed by the special routine *acommute.*
3. After processing by *acommute,* the bitwise & operator is turned into a new *andn* operator; ‘a & b’ becomes ‘a *andn* ˜b’. This is done because the PDP-11 provides no *and* operator, but only *andn.* A similar transformation takes place for ‘=&’.
4. Relationals are turned around so the more complicated expression is on the left. (So that ‘2 > f(x)’ becomes ‘f(x) < 2’). This improves code generation since the algorithm prefers to have the right operand require fewer registers than the left.
5. An expression minus a constant is turned into the expression plus the negative constant, and the *acommute* routine is called to take advantage of the properties of addition.
6. Operators with constant operands are evaluated.
7. Right shifts (unless by 1) are turned into left shifts with a negated right operand, since the PDP-11 lacks a general right-shift operator.
8. A number of special cases are simplified, such as division or multiplication by 1, and shifts by 0.

The *unoptim* routine performs the same sort of processing for unary operators.

1. ‘*&x’ and ‘&*x’ are simplified to ‘x’.
2. If *r* is a register and *c* is a constant or the address of a static or external variable, the expressions ‘*(r+c)’ and ‘*r’ are turned into a special kind of name node which expresses the name itself and the offset. This simplifies subsequent processing because such constructions can appear as the the address of a PDP-11 instruction.
3. When the unary ‘&’ operator is applied to a name node of the special kind just discussed, it is reworked to make the addition explicit again; this is done because the PDP-11 has no ‘load address’ instruction.
4. Constructions like ‘*r++’ and ‘*−−r’ where *r* is a register are discovered and marked as being implementable using the PDP-11 auto-increment and -decrement modes.
5. If ‘!’ is applied to a relational, the ‘!’ is discarded and the sense of the relational is reversed.
6. Special cases involving reflexive use of negation and complementation are discovered.
7. Operations applying to constants are evaluated.

The *acommute* routine, called for associative and commutative operators, discovers clusters of the same operator at the top levels of the current tree, and arranges them in a list: for ‘a+((b+c)+(d+f))’ the list would be‘a,b,c,d,e,f’. After each subtree is optimized, the list is sorted in decreasing difficulty of computation; as mentioned above, the code generation algorithm works best when left operands are the difficult ones. The ‘degree of difficulty’ computed is actually finer than the mere number of registers required; a constant is considered simpler than the address of a static or external, which is simpler than reference to a variable. This makes it easy to fold all the constants together, and also to merge together the sum of a constant and the address of a static or external (since in such nodes there is space for an ‘offset’ value). There are also special cases, like multiplication by 1 and addition of 0.

A special routine is invoked to handle sums of products. *Distrib* is based on the fact that it is better to compute ‘c1*c2*x + c1*y’ as ‘c1*(c2*x + y)’ and makes the divisibility tests required to assure the correctness of the transformation. This transformation is rarely possible with code directly written by the user, but it invariably occurs as a result of the implementation of multi-dimensional arrays.

Finally, *acommute* reconstructs a tree from the list of expressions which result.

Code Generation

The grand plan for code-generation is independent of any particular machine; it depends largely on a set of tables. But this fact does not necessarily make it very easy to modify the compiler to produce code for other machines, both because there is a good deal of machine-dependent structure in the tables, and because in any event such tables are non-trivial to prepare.

The arguments to the basic code generation routine *rcexpr* are a pointer to a tree representing an expression, the name of a code-generation table, and the number of a register in which the value of the expression should be placed. *Rcexpr* returns the number of the register in which the value actually ended up; its caller may need to produce a *mov* instruction if the value really needs to be in the given register. There are four code generation tables.

Regtab is the basic one, which actually does the job described above: namely, compile code which places the value represented by the expression tree in a register.

Cctab is used when the value of the expression is not actually needed, but instead the value of the condition codes resulting from evaluation of the expression. This table is used, for example, to evaluate the expression after *if.* It is clearly silly to calculate the value (0 or 1) of the expression ‘a==b’ in the context ‘if (a==b) ... ’

The *sptab* table is used when the value of an expression is to be pushed on the stack, for example when it is an actual argument. For example in the function call ‘f(a)’ it is a bad idea to load *a* into a register which is then pushed on the stack, when there is a single instruction which does the job.

The *efftab* table is used when an expression is to be evaluated for its side effects, not its value. This occurs mostly for expressions which are statements, which have no value. Thus the code for the statement ‘a = b’ need produce only the approoriate *mov* instruction, and need not leave the value of *b* in a register, while in the expression ‘a + (b = c)’ the value of ‘b = c’ will appear in a register.

All of the tables besides *regtab* are rather small, and handle only a relatively few special cases. If one of these subsidiary tables does not contain an entry applicable to the given expression tree, *rcexpr* uses *regtab* to put the value of the expression into a register and then fixes things up; nothing need be done when the table was *efftab,* but a *tst* instruction is produced when the table called for was *cctab,* and a *mov* instruction, pushing the register on the stack, when the table was *sptab.*

The *rcexpr* routine itself picks off some special cases, then calls *cexpr* to do the real work. *Cexpr* tries to find an entry applicable to the given tree in the given table, and returns –1 if no such entry is found, letting *rcexpr* try again with a different table. A successful match yields a string containing both literal characters which are written out and pseudo-operations, or macros, which are expanded. Before studying the contents of these strings we will consider how table entries are matched against trees.

Recall that most non-leaf nodes in an expression tree contain the name of the operator, the type of the value represented, and pointers to the subtrees (operands). They also contain an estimate of the number of registers required to evaluate the expression, placed there by the expression-optimizer routines. The register counts are used to guide the code generation process, which is based on the Sethi-Ullman algorithm.

The main code generation tables consist of entries each containing an operator number and a pointer to a subtable for the corresponding operator. A subtable consists of a sequence of entries, each with a key describing certain properties of the operands of the operator involved; associated with the key is a code string. Once the subtable corresponding to the operator is found, the subtable is searched linearly until a key is found such that the properties demanded by the key are compatible with the operands of the tree node. A successful match returns the code string; an unsuccessful search, either for the operator in the main table or a compatble key in the subtable, returns a failure indication.

The tables are all contained in a file which must be processed to obtain an assembly language program. Thus they are written in a special-purpose language. To provided definiteness to the following discussion, here is an example of a subtable entry.

```
%n,aw
        F
        add     A2,R
```

The ‘%’ indicates the key; the information following (up to a blank line) specifies the code string. Very briefly, this entry is in the subtable for ‘+’ of *regtab;* the key specifies that the left operand is any integer, character, or pointer expression, and the right operand is any word quantity which is directly addressible (e.g. a variable or constant). The code string calls for the generation of the code to compile the left (first) operand into the current register (‘F’) and then to produce an ‘add’ instruction which adds the second operand (‘A2’) to the register (‘R’). All of the notation will be explained below.

Only three features of the operands are used in deciding whether a match has occurred. They are:

1. Is the type of the operand compatible with that demanded?
2. Is the ‘degree of difficulty’ (in a sense described below) compatible?
3. The table may demand that the operand have a ‘*’ (indirection operator) as its highest operator.

As suggested above, the key for a subtable entry is indicated by a ‘%,’ and a comma-separated pair of specifications for the operands. (The second specification is ignored for unary operators). A specification indicates a type requirement by including one of the following letters. If no type letter is present, any integer, character, or pointer operand will satisfy the requirement (not float, double, or long).

b   A byte (character) operand is required.

w   A word (integer or pointer) operand is required.

f   A float or double operand is required.

d   A double operand is required.

l   A long (32-bit integer) operand is required.

Before discussing the ‘degree of difficulty’ specification, the algorithm has to be explained more completely. *Rcexpr* (and *cexpr)* are called with a register number in which to place their result. Registers 0, 1, ... are used during evaluation of expressions; the maximum register which can be used in this way depends on the number of register variables, but in any event only registers 0 through 4 are available since r5 is used as a stack frame header and r6 (sp) and r7 (pc) have special hardware properties. The code generation routines assume that when called with register *n* as argument, they may use *n+1*, ... (up to the first register variable) as temporaries. Consider the expression ‘X+Y’, where both X and Y are expressions. As a first approximation, there are three ways of compiling code to put this expression in register *n.*

1. If Y is an addressible cell, (recursively) put X into register *n* and add Y to it.
2. If Y is an expression that can be calculated in *k* registers, where *k* smaller than the number of registers available, compile X into register *n,* Y into register *n+1,* and add register *n+1* to *n.*
3. Otherwise, compile Y into register *n*, save the result in a temporary (actually, on the stack) compile X into register *n,* then add in the temporary.

The distinction between cases 2 and 3 therefore depends on whether the right operand can be compiled in fewer than *k* registers, where *k* is the number of free registers left after registers 0 through *n* are taken: 0 through *n–1* are presumed to contain already computed temporary results; *n* will, in case 2, contain the value of the left operand while the right is being evaluated.

These considerations should make clear the specification codes for the degree of difficulty, bearing in mind that a number of special cases are also present:

z   is satisfied when the operand is zero, so that special code can be produced for expressions like ‘x = 0’.

1   is satisfied when the operand is the constant 1, to optimize cases like left and right shift by 1, which can be done efficiently on the PDP-11.

c   is satisfied when the operand is a positive (16-bit) constant; this takes care of some special cases in long arithmetic.

a   is satisfied when the operand is addressible; this occurs not only for variables and constants, but also for some more complicated constructions, such as indirection through a simple variable, '*p++' where *p* is a register variable (because of the PDP-11's auto-increment address mode), and '*(p+c)' where *p* is a register and *c* is a constant. Precisely, the requirement is that the operand refers to a cell whose address can be written as a source or destination of a PDP-11 instruction.

e is satisfied by an operand whose value can be generated in a register using no more than *k* registers, where *k* is the number of registers left (not counting the current register). The 'e' stands for 'easy.'

n is satisfied by any operand. The 'n' stands for 'anything.'

These degrees of difficulty are considered to lie in a linear ordering and any operand which satisfies an earlier-mentioned requirement will satisfy a later one. Since the subtables are searched linearly, if a '1' specification is included, almost certainly a 'z' must be written first to prevent expressions containing the constant 0 to be compiled as if the 0 were 1.

Finally, a key specification may contain a '*' which requires the operand to have an indirection as its leading operator. Examples below should clarify the utility of this specification.

Now let us consider the contents of the code string associated with each subtable entry. Conventionally, lower-case letters in this string represent literal information which is copied directly to the output. Upper-case letters generally introduce specific macro-operations, some of which may be followed by modifying information. The code strings in the tables are written with tabs and new-lines used freely to suggest instructions which will be generated; the table-compiling program compresses tabs (using the 0200 bit of the next character) and throws away some of the new-lines. For example the macro 'F' is ordinarily written on a line by itself; but since its expansion will end with a new-line, the new-line after 'F' itself is dispensable. This is all to reduce the size of the stored tables.

The first set of macro-operations is concerned with compiling subtrees. Recall that this is done by the *cexpr* routine. In the following discussion the 'current register' is generally the argument register to *cexpr;* that is, the place where the result is desired. The 'next register' is numbered one higher than the current register. (This explanation isn't fully true because of complications, described below, involving operations which require even-odd register pairs.)

F causes a recursive call to the *rcexpr* routine to compile code which places the value of the first (left) operand of the operator in the current register.

F1 generates code which places the value of the first operand in the next register. It is incorrectly used if there might be no next register; that is, if the degree of difficulty of the first operand is not 'easy;' if not, another register might not be available.

FS generates code which pushes the value of the first operand on the stack, by calling *rcexpr* specifying *sptab* as the table.

Analogously,

S, S1, SScompile the second (right) operand into the current register, the next register, or onto the stack.

To deal with registers, there are

R which expands into the name of the current register.

R1 which expands into the name of the next register.

R+ which expands into the the name of the current register plus 1. It was suggested above that this is the same as the next register, except for complications; here is one of them. Long integer variables have 32 bits and require 2 registers; in such cases the next register is the current register plus 2. The code would like to talk about both halves of the long quantity, so R refers to the register with the high-order part and R+ to the low-order part.

R− This is another complication, involving division and mod. These operations involve a pair of registers of which the odd-numbered contains the left operand. *Cexpr* arranges that the current register is odd; the R− notation allows the code to refer to the next lower, even-numbered register.

To refer to addressible quantities, there are the notations:

A1 causes generation of the address specified by the first operand. For this to be legal, the operand must be addressible; its key must contain an 'a' or a more restrictive specification.

A2 correspondingly generates the address of the second operand providing it has one.

We now have enough mechanism to show a complete, if suboptimal, table for the + operator on word or byte operands.

```
%n,z
        F

%n,1
        F
        inc     R

%n,aw
        F
        add     A2,R

%n,e
        F
        S1
        add     R1,R

%n,n
        SS
        F
        add     (sp)+,R
```

The first two sequences handle some special cases. Actually it turns out that handling a right operand of 0 is unnecessary since the expression-optimizer throws out adds of 0. Adding 1 by using the ‘increment’ instruction is done next, and then the case where the right operand is addressible. It must be a word quantity, since the PDP-11 lacks an ‘add byte’ instruction. Finally the cases where the right operand either can, or cannot, be done in the available registers are treated.

The next macro-instructions are conveniently introduced by noticing that the above table is suitable for subtraction as well as addition, since no use is made of the commutativity of addition. All that is needed is substitution of ‘sub’ for ‘add’ and ‘dec’ for ’inc.’ Considerable saving of space is achieved by factoring out several similar operations.

I is replaced by a string from another table indexed by the operator in the node being expanded. This secondary table actually contains two strings per operator.

I′ is replaced by the second string in the side table entry for the current operator.

Thus, given that the entries for ‘+’ and ‘–’ in the side table (which is called *instab)* are ‘add’ and ‘inc,’ ‘sub’ and ‘dec’ respectively, the middle of of the above addition table can be written

```
%n,1
        F
        I’      R

%n,aw
        F
        I       A2,R
```

and it will be suitable for subtraction, and several other operators, as well.

Next, there is the question of character and floating-point operations.

B1 generates the letter ‘b’ if the first operand is a character, ‘f’ if it is float or double, and nothing otherwise. It is used in a context like ‘movB1’ which generates a ‘mov’, ‘movb’, or ‘movf’ instruction according to the type of the operand.

B2 is just like B1 but applies to the second operand.

BE generates ‘b’ if either operand is a character and null otherwise.

BF generates ‘f’ if the type of the operator node itself is float or double, otherwise null.

For example, there is an entry in *efftab* for the ‘=’ operator

```
%a,aw
%ab,a
        IBE     A2,A1
```

Note first that two key specifications can be applied to the same code string. Next, observe that when a word is assigned to a byte or to a word, or a word is assigned to a byte, a single instruction, a *mov* or *movb* as appropriate, does the job. However, when a byte is assigned to a word, it must pass through a register to implement the sign-extension rules:

```
%a,n
        S
        IB1     R,A1
```

Next, there is the question of handling indirection properly. Consider the expression ‘X + *Y’, where X and Y are expressions, Assuming that Y is more complicated than just a variable, but on the other hand qualifies as ‘easy’ in the context, the expression would be compiled by placing the value of X in a register, that of *Y in the next register, and adding the registers. It is easy to see that a better job can be done by compiling X, then Y (into the next register), and producing the instruction symbolized by ‘add (R1),R’. This scheme avoids generating the instruction ‘mov (R1),R1’ required actually to place the value of *Y in a register. A related situation occurs with the expression ‘X + *(p+6)’, which exemplifies a construction frequent in structure and array references. The addition table shown above would produce

```
[put X in register R]
mov     p,R1
add     $6,R1
mov     (R1),R1
add     R1,R
```

when the best code is

```
[put X in R]
mov     p,R1
add     6(R1),R
```

As we said above, a key specification for a code table entry may require an operand to have an indirection as its highest operator. To make use of the requirement, the following macros are provided.

F* the first operand must have the form *X. If in particular it has the form *(Y + c), for some constant *c*, then code is produced which places the value of Y in the current register. Otherwise, code is produced which loads X into the current register.

F1* resembles F* except that the next register is loaded.

S* resembles F* except that the second operand is loaded.

S1* resembles S* except that the next register is loaded.

FS* The first operand must have the form ‘*X’. Push the value of X on the stack.

SS* resembles FS* except that it applies to the second operand.

To capture the constant that may have been skipped over in the above macros, there are

#1 The first operand must have the form *X; if in particular it has the form *(Y + c) for *c* a constant, then the constant is written out, otherwise a null string.

#2 is the same as #1 except that the second operand is used.

Now we can improve the addition table above. Just before the ‘%n,e’ entry, put

```
%n,ew*
        F
        S1*
        add     #2(R1),R
```

and just before the '%n,n' put

```
%n,nw*
        SS*
        F
        add     *(sp)+,R
```

When using the stacking macros there is no place to use the constant as an index word, so that particular special case doesn't occur.

The constant mentioned above can actually be more general than a number. Any quantity acceptable to the assembler as an expression will do, in particular the address of a static cell, perhaps with a numeric offset. If *x* is an external character array, the expression 'x[i+5] = 0' will generate the code

```
mov     i,r0
clrb    x+5(r0)
```

via the table entry (in the '=' part of *efftab)*

```
%e*,z
        F
        I'B1    #1(R)
```

Some machine operations place restrictions on the registers used. The divide instruction, used to implement the divide and mod operations, requires the dividend to be placed in the odd member of an even-odd pair; other peculiarities of multiplication make it simplest to put the multiplicand in an odd-numbered register. There is no theory which optimally accounts for this kind of requirement. *Cexpr* handles it by checking for a multiply, divide, or mod operation; in these cases, its argument register number is incremented by one or two so that it is odd, and if the operation was divide or mod, so that it is a member of a free even-odd pair. The routine which determines the number of registers required estimates, conservatively, that at least two registers are required for a multiplication and three for the other peculiar operators. After the expression is compiled, the register where the result actually ended up is returned. (Divide and mod are actually the same operation except for the location of the result).

These operations are the ones which cause results to end up in unexpected places, and this possibility adds a further level of complexity. The simplest way of handling the problem is always to move the result to the place where the caller expected it, but this will produce unnecessary register moves in many simple cases; 'a = b*c' would generate

```
mov     b,r1
mul     c,r1
mov     r1,r0
mov     r0,a
```

The next thought is used the passed-back information as to where the result landed to change the notion of the current register. While compiling the '=' operation above, which comes from a table entry like

```
%a,e
        S
        mov     R,A1
```

it is sufficient to redefine the meaning of 'R' after processing the 'S' which does the multiply. This technique is in fact used; the tables are written in such a way that correct code is produced. The trouble is that the technique cannot be used in general, because it invalidates the count of the number of registers required for an expression. Consider just 'a*b + X' where X is some expression. The algorithm assumes that the value of a*b, once computed, requires just one register. If there are three registers available, and X requires two registers to compute, then this expression will match a key specifying '%n,e'. If a*b is computed and left in register 1, then there are, contrary to expectations, no longer two registers available to compute X, but only one, and bad code will be produced. To guard against this possibility, *cexpr* checks the result returned by recursive calls which implement F, S and their relatives. If the result is not in the expected register, then the number of registers required by the other operand is checked; if it can be done using those registers which remain even after making unavailable the unexpectedly-occupied register, then the notions of the 'next register' and possibly the 'current register' are redefined. Otherwise a register-copy instruction is produced. A register-copy is also always produced when the current operator is one of those which have odd-even requirements.

Finally, there are a few loose-end macro operations and facts about the tables. The operators:

V is used for long operations. It is written with an address like a machine instruction; it expands into 'adc' (add carry) if the operation is an additive operator, 'sbc' (subtract carry) if the operation is a subtractive operator, and disappears, along with the rest of the line, otherwise. Its purpose is to allow common treatment of logical operations, which have no carries, and additive and subtractive operations, which generate carries.

T generates a 'tst' instruction if the first operand of the tree does not set the condition codes correctly. It is used with divide and mod operations, which require a sign-extended 32-bit operand. The code table for the operations contains an 'sxt' (sign-extend) instruction to generate the high-order part of the dividend.

H is analogous to the 'F' and 'S' macros, except that it calls for the generation of code for the current tree (not one of its operands) using *regtab.* It is used in *cctab* for all the operators which, when executed normally, set the condition codes properly according to the result. It prevents a 'tst' instruction from being generated for constructions like 'if (a+b) ...' since after calculation of the value of 'a+b' a conditional branch can be written immediately.

All of the discussion above is in terms of operators with operands. Leaves of the expression tree (variables and constants), however, are peculiar in that they have no operands. In order to regularize the matching process, *cexpr* examines its operand to determine if it is a leaf; if so, it creates a special 'load' operator whose operand is the leaf, and substitutes it for the argument tree; this allows the table entry for the created operator to use the 'A1' notation to load the leaf into a register.

Purely to save space in the tables, pieces of subtables can be labelled and referred to later. It turns out, for example, that rather large portions of the the *efftab* table for the '=' and '=+' operators are identical. Thus '=' has an entry

```
%[move3:]
%a,aw
%ab,a
        IBE     A2,A1
```

while part of the '=+' table is

```
%aw,aw
%       [move3]
```

Labels are written as '%[... :]', before the key specifications; references are written with '% [...]' after the key. Peculiarities in the implementation make it necessary that labels appear before references to them.

The example illustrates the utility of allowing separate keys to point to the same code string. The assignment code works properly if either the right operand is a word, or the left operand is a byte; but since there is no 'add byte' instruction the addition code has to be restricted to word operands.

Delaying and reordering

Intertwined with the code generation routines are two other, interrelated processes. The first, implemented by a routine called *delay*, is based on the observation that naive code generation for the expression 'a = b++' would produce

```
mov    b,r0
inc    b
mov    r0,a
```

The point is that the table for postfix ++ has to preserve the value of *b* before incrementing it; the general way to do this is to preserve its value in a register. A cleverer scheme would generate

```
mov    b,a
inc    b
```

Delay is called for each expression input to *rcexpr,* and it searches for postfix ++ and −− operators. If one is found applied to a variable, the tree is patched to bypass the operator and compiled as it stands; then the increment or decrement itself is done. The effect is as if 'a = b; b++' had been written. In this example, of course, the user himself could have done the same job, but more complicated examples are easily constructed, for example 'switch (x++)'. An essential restriction is that the condition codes not be required. It would be incorrect to compile 'if (a++) ...' as

```
tst    a
inc    a
beq    ...
```

because the 'inc' destroys the required setting of the condition codes.

Reordering is a similar sort of optimization. Many cases which it detects are useful mainly with register variables. If *r* is a register variable, the expression 'r = x+y' is best compiled as

```
mov    x,r
add    y,r
```

but the codes tables would produce

```
mov    x,r0
add    y,r0
mov    r0,r
```

which is in fact preferred if *r* is not a register. (If *r* is not a register, the two sequences are the same size, but the second is slightly faster.) The scheme is to compile the expression as if it had been written 'r = x; r =+ y'. The *reorder* routine is called with a pointer to each tree that *rcexpr* is about to compile; if it has the right characteristics, the 'r = x' tree is constructed and passed recursively to *rcexpr;* then the original tree is modified to read 'r =+ y' and the calling instance of *rcexpr* compiles that instead. Of course the whole business is itself recursive so that more extended forms of the same phenomenon are handled, like 'r = x + y | z'.

Care does have to be taken to avoid 'optimizing' an expression like 'r = x + r' into 'r = x; r =+ r'. It is required that the right operand of the expression on the right of the '=' be a ', distinct from the register variable.

The second case that *reorder* handles is expressions of the form 'r = X' used as a subexpression. Again, the code out of the tables for 'x = r = y' would be

```
mov    y,r0
mov    r0,r
mov    r0,x
```

whereas if *r* were a register it would be better to produce

```
mov    y,r
mov    r,x
```

When *reorder* discovers that a register variable is being assigned to in a subexpression, it calls *rcexpr* recursively to compile the subexpression, then fiddles the tree passed to it so that the register variable itself appears as the operand instead of the whole subexpression. Here care has to be taken to avoid an infinite regress, with *rcexpr* and *reorder* calling each other forever to handle assignments to registers.

A third set of cases treated by *reorder* comes up when any name, not necessarily a register, occurs as a left operand of an assignment operator other than '=' or as an operand of prefix '++' or '−−'. Unless condition-code tests are involved, when a subexpression like '(a =+ b)' is seen, the assignment is performed and the argument tree modified so that *a* is its operand; effectively 'x + (y =+ z)' is compiled as 'y =+ z; x + y'. Similarly, prefix increment and decrement are pulled out and performed first, then the remainder of the expression.

Throughout code generation, the expression optimizer is called whenever *delay* or *reorder* change the expression tree. This allows some special cases to be found that otherwise would not be seen.

A New Input-Output Package

D. M. Ritchie

A new package of IO routines is available under the Unix system. It was designed with the following goals in mind.

1. It should be similar in spirit to the earlier Portable Library, and, to the extent possible, be compatible with it. At the same time a few dubious design choices in the Portable Library will be corrected.
2. It must be as efficient as possible, both in time and in space, so that there will be no hesitation in using it no matter how critical the application.
3. It must be simple to use, and also free of the magic numbers and mysterious calls the use of which mars the understandability and portability of many programs using older packages.
4. The interface provided should be applicable on all machines, whether or not the programs which implement it are directly portable to other systems, or to machines other than the PDP11 running a version of Unix.

It is intended that this package replace the Portable Library. Although it is not directly compatible, as discussed below, it is sufficiently similar that a set of relatively small, inexpensive adaptor routines exist which make it appear identical to the current Portable Library except in some very obscure details.

The most crucial difference between this package and the Portable Library is that the current offering names streams in terms of pointers rather than by the integers known as 'file descriptors.' Thus, for example, the routine which opens a named file returns a pointer to a certain structure rather than a number; the routine which reads an open file takes as an argument the pointer returned from the open call.

General Usage

Each program using the library must have the line

```
#include <stdio.h>
```

which defines certain macros and variables. The library containing the routines is '/usr/lib/libS.a,' so the command to compile is

```
cc . . . −lS
```

All names in the include file intended only for internal use begin with an underscore '_' to reduce the possibility of collision with a user name. The names intended to be visible outside the package are

stdin The name of the standard input file

stdout The name of the standard output file

stderr The name of the standard error file

EOF is actually −1, and is the value returned by the read routines on end-of-file or error.

NULL is a notation for the null pointer, returned by pointer-valued functions to indicate an error

FILE expands to 'struct _iob' and is a useful shorthand when declaring pointers to streams.

BUFSIZ is a number (viz. 512) of the size suitable for an IO buffer supplied by the user. See *setbuf*, below.

getc, getchar, putc, putchar, feof, ferror, fileno are defined as macros. Their actions are described below; they are mentioned here to point out that it is not possible to redeclare them and that they are not actually functions; thus, for example, they may not have breakpoints set on them.

The routines in this package, like the current Portable Library, offer the convenience of automatic buffer allocation and output flushing where appropriate. Absent, however, is the facility of changing the default input and output streams by assigning to 'cin' and 'cout.' The names 'stdin,' stdout,' and 'stderr' are in effect constants and may not be assigned to.

Calls

The routines in the library are in nearly one-to-one correspondence with those in the Portable Library. In several cases the name has been changed. This is an attempt to reduce confusion. If the attempt is judged to fail the names may be made identical even though the arguments may be different. The order of this list generally follows the order used in the Portable Library document.

*FILE *fopen(filename, type)*

Fopen opens the file and, if needed, allocates a buffer for it. *Filename* is a character string specifying the name. *Type* is a character string (not a single character). It may be '"r",' '"w",' or '"a"' to indicate intent to read, write, or append. The value returned is a file pointer. If it is null the attempt to open failed.

int getc(ioptr)

returns the next character from the stream named by *ioptr,* which is a pointer to a file such as returned by *fopen,* or the name *stdin.* The integer EOF is returned on end-of-file or when an error occurs. The null character is a legal character.

putc(c, ioptr)

Putc writes the character *c* on the output stream named by *ioptr,* which is a value returned from *fopen* or perhaps *stdout* or *stderr.* The character is returned as value, but EOF is returned on error.

fclose(ioptr)

The file corresponding to *ioptr* is closed after any buffers are emptied. A buffer allocated by the IO system is freed. *Fclose* is automatic on normal termination of the program.

fflush(ioptr)

Any buffered information on the (output) stream named by *ioptr* is written out. Output files are normally buffered if and only if they are not directed to the terminal, but *stderr* is unbuffered unless *setbuf* is used.

exit(errcode)

Exit terminates the process and returns its argument as status to the parent. This is a special version of the routine which calls *fflush* for each output file. To terminate without flushing, use *_exit.*

feof(ioptr)

returns non-zero when end-of-file has occurred on the specified input stream.

ferror(ioptr)

returns non-zero when an error has occurred while reading or writing the named stream. The error indication lasts until the file has been closed.

getchar( )

is identical to 'getc(stdin)'.

putchar(c)
is identical to 'putc(c, stdout)'.

*char *gets(s)*
reads characters up to a new-line from the standard input. The new-line character is replaced by a null character. It is the user's responsibility to make sure that the character array *s* is large enough. *Gets* returns its argument, or null if end-of-file or error occurred.

*char *fgets(s, n, ioptr)*
reads up to *n* characters from the stream *ioptr* into the character pointer *s.* The read terminates with a new-line character. The new-line character is placed in the buffer followed by a null pointer. The first argument, or a null pointer if error or end-of-file occurred, is returned.

puts(s)
writes the null-terminated string (character array) *s* on the standard output. A new-line is appended. No value is returned.

fputs(s, ioptr)
writes the null-terminated string (character array) on the stream *s.* No new-line is appended. No value is returned.

ungetc(c, ioptr)
The argument character *c* is pushed back on the input stream named by *ioptr.* Only one character may be pushed back.

printf(format, a1, . . .)

fprintf(ioptr, format, a1, . . .)

sprintf(s, format, a1, . . .)
Printf writes on the standard output. *Fprintf* writes on the named output stream. *Sprintf* puts characters in the character array (string) named by *s.* The specifications are as usual.

scanf(format, a1, . . .)

fscanf(ioptr, format, a1, . . .)

sscanf(s, format, a1, . . .)
Scanf reads from the standard input. *Fscanf* reads from the named input stream. *Sscanf* reads from the character string supplied as *s.* The specifications are identical to those of the Portable Library.

*fread(ptr, sizeof(*ptr), nitems, ioptr)*
writes *nitems* of data beginning at *ptr* on file *ioptr.* It behaves identically to the Portable Library's *cread.* No advance notification that binary IO is being done is required; when, for portability reasons, it becomes required, it will be done by adding an additional character to the mode-string on the fopen call.

*fwrite(ptr, sizeof(*ptr), nitems, ioptr)*
Like *fread,* but in the other direction.

rewind(ioptr)
rewinds the stream named by *ioptr.* It is not very useful except on input, since a rewound output file is still open only for output.

system(string)

atof(s)

tmpnam(s)

abort(code)

intss( )

cfree(ptr)

wdleng( )
are available with specifications identical to those described for the Portable Library.

*char *calloc(n, sizeof(object))*
returns null when no space is available. The space is guaranteed to be 0.

ftoa
is not implemented but there are plausible alternatives.

nargs( )
is not implemented.

getw(ioptr)
returns the next word from the input stream named by *ioptr*. EOF is returned on end-of-file or error, but since this a perfectly good integer *feof* and *ferror* should be used.

putw(w, ioptr)
writes the integer *w* on the named output stream.

setbuf(ioptr, buf)
Setbuf may be used after a stream has been opened but before IO has started. If *buf* is null, the stream will be unbuffered. Otherwise the buffer supplied will be used. It is a character array of sufficient size:

```
char     buf[BUFSIZ];
```

fileno(ioptr)
returns the integer file descriptor associated with the file.

Several additional routines are available.

fseek(ioptr, offset, ptrname)
The location of the next byte in the stream named by *ioptr* is adjusted. *Offset* is a long integer. If *ptrname* is 0, the offset is measured from the beginning of the file; if *ptrname* is 1, the offset is measured from the current read or write pointer; if *ptrname* is 2, the offset is measured from the end of the file. The routine accounts properly for any buffering.

long ftell(iop)
The byte offset, measured from the beginning of the file, associated with the named stream is returned. Any buffering is properly accounted for.

getpw(uid, buf)

The password file is searched for the given integer user ID. If an appropriate line is found, it is copied into the character array *buf*, and 0 is returned. If no line is found corresponding to the user ID then 1 is returned.

strcat(s1, s2)

S1 and *s2* are character pointers. The end (null byte) of the *s1* string is found and *s2* is copied to *s1* starting there. The space pointed to by *s1* must be large enough.

strcmp(s1, s2)

The character strings *s1* and *s2* are compared. The result is positive, zero, or negative according as *s1* is greater than, equal to, or less than *s2* in ASCII collating sequence.

strcpy(s1, s2)

The null-terminated character string *s2* is copied to the location pointed to by *s1*.

strlen(s)

The number of bytes in s up to a null byte is returned. *S* is a character pointer.

gcvt(num, ndig, buf)

Num is a floating or double quantity. *Ndig* significant digits are converted to ASCII and placed into the character array *buf*. The conversion is in Fortran *e* or *f* style, whichever yields the shorter string. Insignificant trailing zeros are eliminated.

C Changes

1. Long integers

The compiler implements 32-bit integers. The associated type keyword is ‘long’. The word can act rather like an adjective in that ‘long int’ means a 32-bit integer and ‘long float’ means the same as ‘double.’ But plain ‘long’ is a long integer. Essentially all operations on longs are implemented except that assignment-type operators do not have values, so l1+(l2=+l3) won’t work. Neither will l1 = l2 = 0.

Long constants are written with a terminating ‘l’ or ‘L’. E.g. "123L" or "0177777777L" or "0X56789abcdL". The latter is a hex constant, which could also have been short; it is marked by starting with "0X". Every fixed decimal constant larger than 32767 is taken to be long, and so are octal or hex constants larger than 0177777 (0Xffff, or 0xFFFF if you like). A warning is given in such a case since this is actually an incompatibility with the older compiler. Where the constant is just used as an initializer or assigned to something it doesn’t matter. If it is passed to a subroutine then the routine will not get what it expected.

When a short and a long integer are operands of an arithmetic operator, the short is converted to long (with sign extension). This is true also when a short is assigned to a long. When a long is assigned to a short integer it is truncated at the high order end with no notice of possible loss of significant digits. This is true as well when a long is added to a pointer (which includes its usage as a subscript). The conversion rules for expressions involving doubles and floats mixed with longs are the same as those for short integers, *mutatis mutandis.*

A point to note is that constant expressions involving longs are not evaluated at compile time, and may not be used where constants are expected. Thus

```
long x {5000L*5000L};
```

is illegal;

```
long x {5000*5000};
```

is legal but wrong because the high-order part is lost; but both

```
long x 25000000L;
```

and

```
long x 25.e6;
```

are correct and have the same meaning because the double constant is converted to long at compile time.

2. Unsigned integers

A new fundamental data type with keyword ‘unsigned,’ is available. It may be used alone:

```
unsigned u;
```

or as an adjective with ‘int’

```
unsigned int u;
```

with the same meaning. There are not yet (or possibly ever) unsigned longs or chars. The meaning of an unsigned variable is that of an integer modulo 2ˆn, where n is 16 on the PDP-11. All operators whose operands are unsigned produce results consistent with this interpretation except division and remainder where the divisor is larger than 32767; then the result is incorrect. The dividend in an unsigned division may however have any value (i.e. up to 65535) with correct results. Right shifts of unsigned quantities are guaranteed to be logical shifts.

When an ordinary integer and an unsigned integer are combined then the ordinary integer is mapped into an integer mod 2ˆ16 and the result is unsigned. Thus, for example ‘u = -1’ results in assigning 65535 to u. This is mathematically reasonable, and also happens to involve no run-time overhead.

When an unsigned integer is assigned to a plain integer, an (undiagnosed) overflow occurs when the unsigned integer exceeds 2ˆ15-1.

It is intended that unsigned integers be used in contexts where previously character pointers were used (artificially and nonportably) to represent unsigned integers.

3. Block structure.

A sequence of declarations may now appear at the beginning of any compound statement in {}. The variables declared thereby are local to the compound statement. Any declarations of the same name existing before the block was entered are pushed down for the duration of the block. Just as in functions, as before, auto variables disappear and lose their values when the block is left; static variables retain their values. Also according to the same rules as for the declarations previously allowed at the start of functions, if no storage class is mentioned in a declaration the default is automatic.

Implementation of inner-block declarations is such that there is no run-time cost associated with using them.

4. Initialization (part 1)

This compiler properly handles initialization of structures so the construction

```
struct { char name[8]; char type; float val; } x
        { "abc", 'a', 123.4 };
```

compiles correctly. In particular it is recognized that the string is supposed to fill an 8-character array, the ‘a’ goes into a character, and that the 123.4 must be rounded and placed in a single-precision cell. Structures of arrays, arrays of structures, and the like all work; a more formal description of what is done follows.

```
<initializer> ::= <element>

<element> ::= <expression> | <element> , <element> |
        { <element> } | { <element> , }
```

An element is an expression or a comma-separated sequence of elements possibly enclosed in braces. In a brace-enclosed sequence, a comma is optional after the last element. This very ambiguous definition is parsed as described below. "Expression" must of course be a constant expression within the previous meaning of the Act.

An initializer for a non-structured scalar is an element with exactly one expression in it.

An "aggregate" is a structure or an array. If the initializer for an aggregate begins with a left brace, then the succeeding comma-separated sequence of elements initialize the members of the aggregate. It is erroneous for the number of members in the sequence to exceed the number of elements in the aggregate. If the sequence has too few members the aggregate is padded.

If the initializer for an aggregate does not begin with a left brace, then the members of the aggregate are initialized with successive elements from the succeeding comma-separated sequence. If the sequence terminates before the aggregate is filled the aggregate is padded.

The "top level" initializer is the object which initializes an external object itself, as opposed to one of its members. The top level initializer for an aggregate must begin with a left brace.

If the top-level object being initialized is an array and if its size is omitted in the declaration, e.g. "int a[]", then the size is calculated from the number of elements which initialized it.

Short of complete assimilation of this description, there are two simple approaches to the initialization of complicated objects. First, observe that it is always legal to initialize any object with a comma-separated sequence of expressions. The members of every structure and array are stored in a specified order, so the expressions which initialize these members may if desired be laid out in a row to successively, and recursively, initialize the members.

Alternatively, the sequences of expressions which initialize arrays or structures may uniformly be enclosed in braces.

5. Initialization (part 2)

Declarations, whether external, at the head of functions, or in inner blocks may have initializations whose syntax is the same as previous external declarations with initializations. The only restrictions are that automatic structures and arrays may not be initialized (they can’t be assigned either); nor, for the moment at least, may external variables when declared inside a function.

The declarations and initializations should be thought of as occurring in lexical order so that forward references in initializations are unlikely to work. E.g.,

```
{ int a a;
  int b c;
  int c 5;
  ...
}
```

Here a is initialized by itself (and its value is thus undefined); b is initialized with the old value of c (which is either undefined or any c declared in an outer block).

6. Bit fields

A declarator inside a structure may have the form

<declarator> : <constant>

which specifies that the object declared is stored in a field the number of bits in which is specified by the constant. If several such things are stacked up next to each other then the compiler allocates the fields from right to left, going to the next word when the new field will not fit. The declarator may also have the form

: <constant>

which allocates an unnamed field to simplify accurate modelling of things like hardware formats where there are unused fields. Finally,

: 0

means to force the next field to start on a word boundary.

The types of bit fields can be only "int" or "char". The only difference between the two is in the alignment and length restrictions: no int field can be longer than 16 bits, nor any char longer than 8 bits. If a char field will not fit into the current character, then it is moved up to the next character boundary.

Both int and char fields are taken to be unsigned (non-negative) integers.

Bit-field variables are not quite full-class citizens. Although most operators can be applied to them, including assignment operators, they do not have addresses (i.e. there are no bit pointers) so the unary & operator cannot be applied to them. For essentially this reason there are no arrays of bit field variables.

There are three twoes in the implementation: addition (=+) applied to fields can result in an overflow into the next field; it is not possible to initialize bit fields.

7. Macro preprocessor

The proprocessor handles ‘define’ statements with formal arguments. The line

```
#define macro(a1,...,an) ...a1...an...
```

is recognized by the presence of a left parenthesis following the defined name. When the form

```
macro(b1,...,bn)
```

is recognized in normal C program text, it is replaced by the definition, with the corresponding *bi* actual argument string substituted for the corresponding *ai* formal arguments. Both actual and formal arguments are separated by commas not included in parentheses; the formal arguments have the syntax of names.

Macro expansions are no longer surrounded by spaces. Lines in which a replacement has taken place are rescanned until no macros remain.

The preprocessor has a rudimentary conditional facility. A line of the form

```
#ifdef name
```

is ignored if ‘name’ is defined to the preprocessor (i.e. was the subject of a ‘define’ line). If name is not defined then all lines through a line of the form

```
#endif
```

are ignored. A corresponding form is

```
#ifndef name
...
#endif
```

which ignores the intervening lines unless ‘name’ is defined. The name ‘unix’ is predefined and replaced by itself to aid writers of C programs which are expected to be transported to other machines with C compilers.

In connection with this, there is a new option to the cc command:

```
cc -Dname
```

which causes ‘name’ to be defined to the preprocessor (and replaced by itself). This can be used together with conditional preprocessor statements to select variant versions of a program at compile time.

The previous two facilities (macros with arguments, conditional compilation) were actually available in the 6th Edition system, but undocumented. New in this release of the cc command is the ability to nest ‘include’ files. Preprocessor include lines may have the new form

```
#include <file>
```

where the angle brackets replace double quotes. In this case, the file name is prepended with a standard prefix, namely ‘/usr/include’. In is intended that commonly-used include files be placed in this directory; the convention reduces the dependence on system-specific naming conventions. The standard prefix can be replaced by the cc command option ‘-I’:

```
cc -Iotherdirectory
```

8. Registers

A formal argument may be given the storage class ‘register.’ When this occurs the save sequence copies it from the place the caller left it into a fast register; all usual restrictions on its use are the same as for ordinary register variables.

Now any variable inside a function may be declared ‘register;’ if the type is unsuitable, or if there are more than three register declarations, then the compiler makes it ‘auto’ instead. The restriction that the & operator may not be applied to a register remains.

9. Mode declarations

A declaration of the form

typedef type-specifier declarator ;

makes the name given in the declarator into the equivalent of a keyword specifying the type which the name would have in an ordinary declaration. Thus

```
typedef int *iptr;
```

makes 'iptr' usable in declarations of pointers to integers; subsequently the declarations

```
iptr ip;
int *ip;
```

would mean the same thing. Type names introduced in this way obey the same scope rules as ordinary variables. The facility is new, experimental, and probably buggy.

10. Restrictions

The compiler is somewhat stickier about some constructions that used to be accepted.

One difference is that external declarations made inside functions are remembered to the end of the file, that is even past the end of the function. The most frequent problem that this causes is that implicit declaration of a function as an integer in one routine, and subsequent explicit declaration of it as another type, is not allowed. This turned out to affect several source programs distributed with the system.

It is now required that all forward references to labels inside a function be the subject of a 'goto.' This has turned out to affect mainly people who pass a label to the routine 'setexit.' In fact a routine is supposed to be passed here, and why a label worked I do not know.

In general this compiler makes it more difficult to use label variables. Think of this as a contribution to structured programming.

The compiler now checks multiple declarations of the same name more carefully for consistency. It used to be possible to declare the same name to be a pointer to different structures; this is caught. So too are declarations of the same array as having different sizes. The exception is that array declarations with empty brackets may be used in conjunction with a declaration with a specified size. Thus

```
int a[];     int a[50];
```

is acceptable (in either order).

An external array all of whose definitions involve empty brackets is diagnosed as 'undefined' by the loader; it used to be taken as having 1 element.