Chapter 14 ::
Building a Runnable Program

Programming Language Pragmatics

Michael L. Scott

ELSEVIER

Introduction

- The phases of compilation are commonly grouped in
    - *front end* responsible for the analysis of source code
    - *back end* responsible for the synthesis of target code.
- Chapters 2 and 4 discussed the work of the front end, culminating in the construction of a syntax tree
- The current chapter turns to the work of the back-end, and specifically to code generation, assembly, and linking
- We continue with code improvement in Chapter 15
- Now we look at how the compiler produces that code from a syntax tree, generating a runnable program

Back-End Compiler Structure

- A plausible seven-phase structure for a conventional compiler is shown in Figure 14.1:
  - First three phases (scanning, parsing, and semantic analysis) are language-dependent
  - Last two (target code generation and machine-specific code improvement) are machine dependent
  - The middle two (intermediate code generation and machine-independent code improvement) are dependent on neither the language nor the machine
- The scanner and parser drive a set of action routines that build a syntax tree
- The semantic analyzer traverses the tree, performing all static semantic checks and initializing various attributes of use to the back end
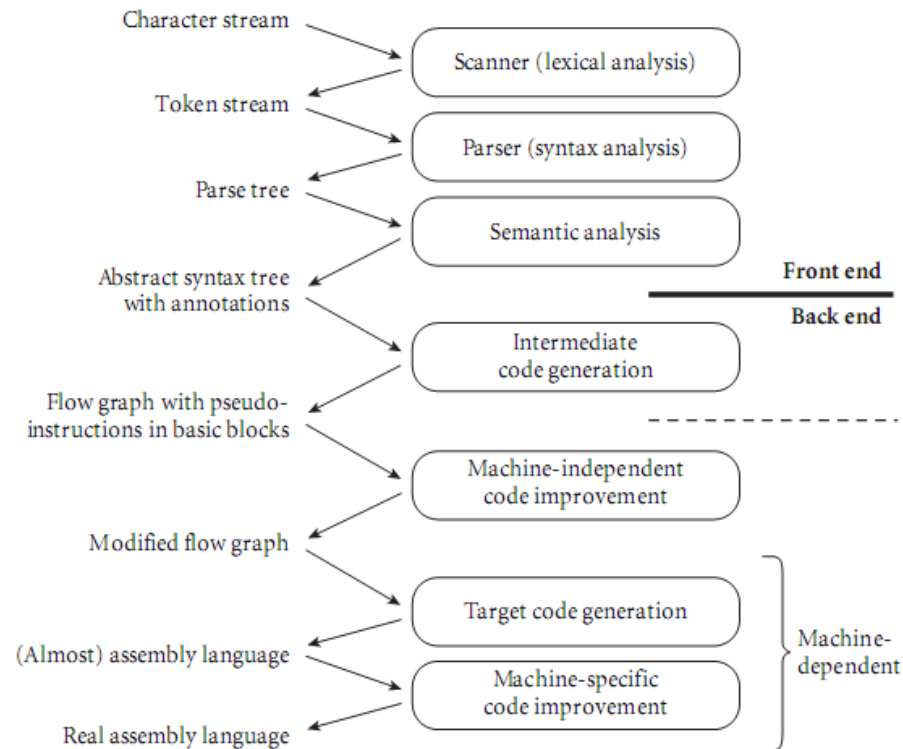
Back-End Compiler Structure

Figure 14.1 A plausible structure for the compiler back end. Here we have shown a sharper separation between semantic analysis and intermediate code generation than we considered in Chapter 1 (see Figure 1.3, page 24). Machine-independent code improvement employs an intermediate form that resembles the assembly language for an idealized machine with an unlimited number of registers. Machine-specific code improvement—register allocation and instruction scheduling in particular—employs the assembly language of the target machine. The dashed line shows a common alternative "break point" between the front end and back end of a two-pass compiler. In some implementations, machine-independent code improvement may be located in a separate "middle end" pass.

Back-End Compiler Structure

- Certain code improvements can be performed on syntax trees - a less hierarchical representation of the program makes most code improvement easier

- Our example compiler therefore includes an explicit phase for intermediate code generation
  - The code generator groups the nodes into *basic blocks*
  - It then creates a *control flow graph* in which the nodes are basic blocks and the arcs represent interblock control flow
    - Within each basic block, operations are represented as instructions for an idealized RISC machine with an unlimited number of registers - we will call these *virtual registers*
    - By allocating a new one for every computed value, the compiler can avoid creating artificial connections between otherwise independent computations too early in the compilation process

Back-End Compiler Structure

- The second phase of the back end machine-independent code performs transformations on the control flow graph.
  - *local code improvement* - it modifies the instruction sequence within each basic block to eliminate redundant loads, stores, and arithmetic computations
  - *global code improvement* - it also identifies and removes a variety of redundancies across the boundaries between basic blocks within a subroutine
  - an expression whose value is computed immediately before an if statement need not be recomputed after else
  - An expression that appears within the body of a loop need only be evaluated once if its value will not change in subsequent iterations
- Some global improvements change the number of basic blocks and/or the arcs among them
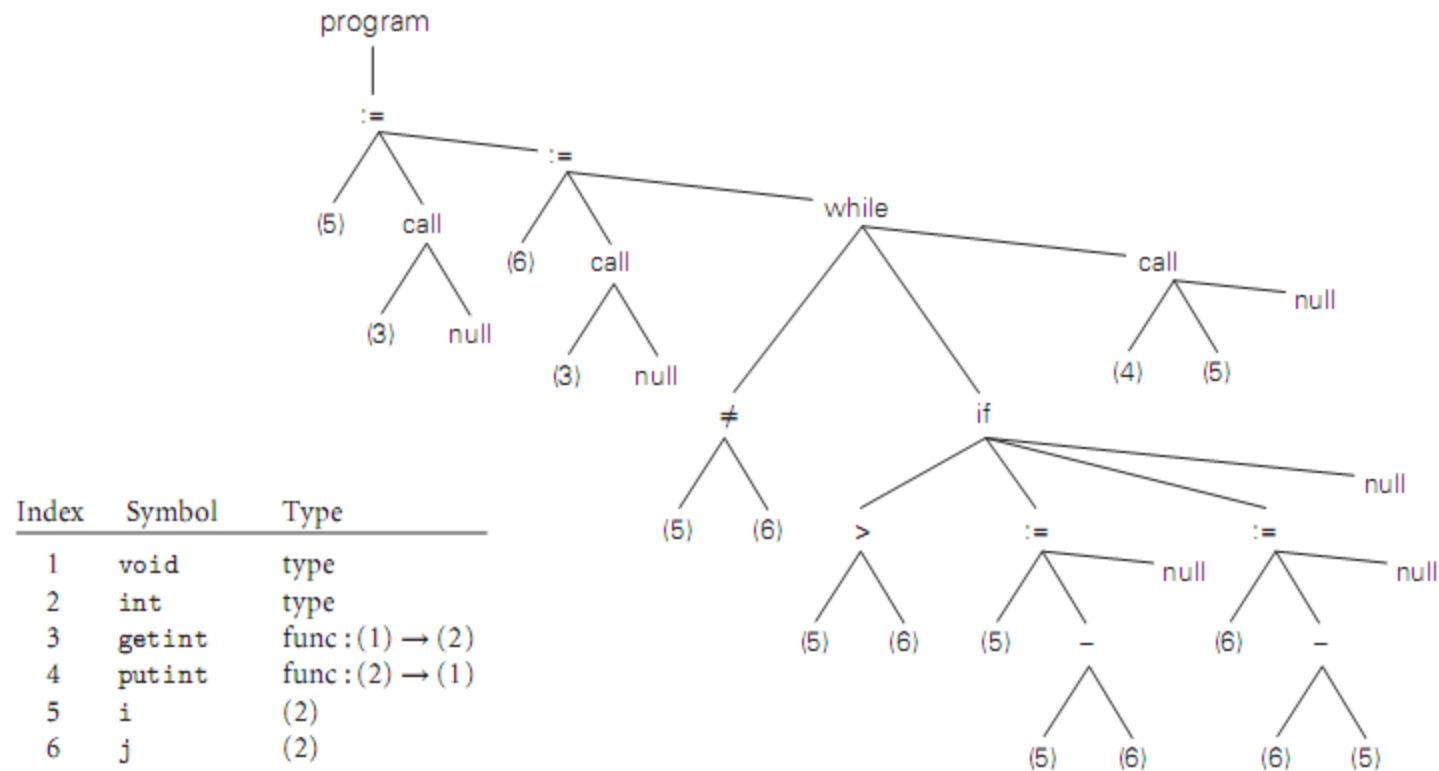
Back-End Compiler Structure

Figure 14.2 Syntax tree and symbol table for the GCD program. The only difference from Figure 1.5 is the addition of explicit null nodes to indicate empty argument lists and to terminate statement lists.

Back-End Compiler Structure

- Following machine-independent code improvement, the next phase of compilation is *target code generation*
    - This phase strings the basic blocks together into a linear program, translating each block into the instruction set of the target machine and generating branch instructions (or "fall-throughs") that correspond to the arcs of the control flow graph.

- The output of this phase differs from real assembly language primarily in its continued reliance on virtual registers
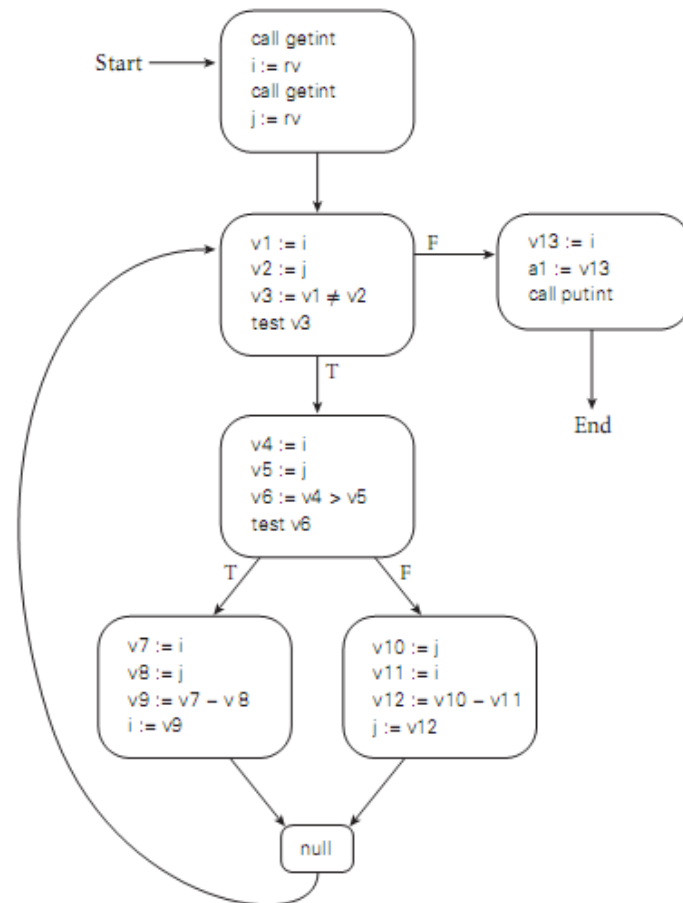
Back-End Compiler Structure

Figure 14.3 Control flow graph for the GCD program. Code within basic blocks is shown in the pseudo-assembly notation introduced on page 212, with a different virtual register (here named v1...v13) for every computed value. Registers a1 and rv are used to pass values to and from subroutines.

Back-End Compiler Structure

- The final phase of our example compiler structure consists of register allocation and instruction scheduling - machine-specific code improvement

- Register allocation requires that we map the unlimited virtual registers onto the bounded set of registers available in the target machine
  - If there aren't enough architectural registers to go around, we may need to generate additional loads and stores to multiplex a given architectural register among two or more virtual registers
  - As described in Section 5.5, instruction scheduling consists of reordering the instructions of each basic block to fill the pipeline(s) of the target machine

Back-End Compiler Structure

- ## Phases and Passes
  - A *pass* of compilation is a phase or sequence of phases that is serialized with respect to the rest of compilation
    - it does not start until previous phases have completed
    - it finishes before any subsequent phases start.
    - if desired, a pass may be written as a separate program, reading its input from a file and writing its output to a file.
  - Two-pass compilers are particularly common
    - they may be divided between the front end and the back end (between semantic analysis and intermediate code generation)

      or

    - they may be divided between intermediate code generation and global code improvement
    - In the latter case, the first pass is still commonly referred to as the front end and the second pass as the back end

ELSEVIER

Intermediate Forms

- An *intermediate form* (IF) provides the connection between the front end and the back end of the compiler, and continues to represent the program during the various back-end phases.

- IFs can be classified in terms of their *level*, or degree of machine dependence.

- High level IFs
  - IFs are often based on trees or directed acyclic graphs (DAGs) that directly capture the structure of modern programming languages
  - facilitates certain kinds of machine-independent code improvement, incremental program updates, direct interpretation, and other operations based strongly on the structure of the source
  - Because the permissible structure of a tree can be described formally by a set of productions (cf., Section 4.6), manipulations of tree-based forms can be written as attribute grammars
  - *Stack-based* languages are another common type of high level IF

Intermediate Forms

- The most common medium level IFs consist of three-address instructions for a simple idealized machine, typically one with an unlimited number of registers
  - Since the typical instruction specifies two operands, an operator, and a destination, three-address instructions are called *quadruples*
  - In older compilers, one may sometimes find an intermediate form consisting of *triples* or *indirect triples* in which destinations are specified implicitly
    - the index of a triple in the instruction stream is the name of the result
    - an operand is generally named by specifying the index of the triple that produced it.

Intermediate Forms

- Different compilers use different Ifs
  - Many compilers use more than one IF internally, though in the common two-pass organization one of these is distinguished as "the" intermediate form
    - connection between the front end and the back end.
  - the syntax trees passed from semantic analysis to intermediate code generation constitute a high level IF
  - control flow graphs containing pseudo-assembly language (passed in and out of machine-independent code improvement) are a medium level IF
  - the assembly language of the target machine serves as a low level IF

- Compilers that have back ends for different target architectures do as much work as possible on a high or medium level IF
  - the machine-independent parts of the code improver can be shared by different back ends
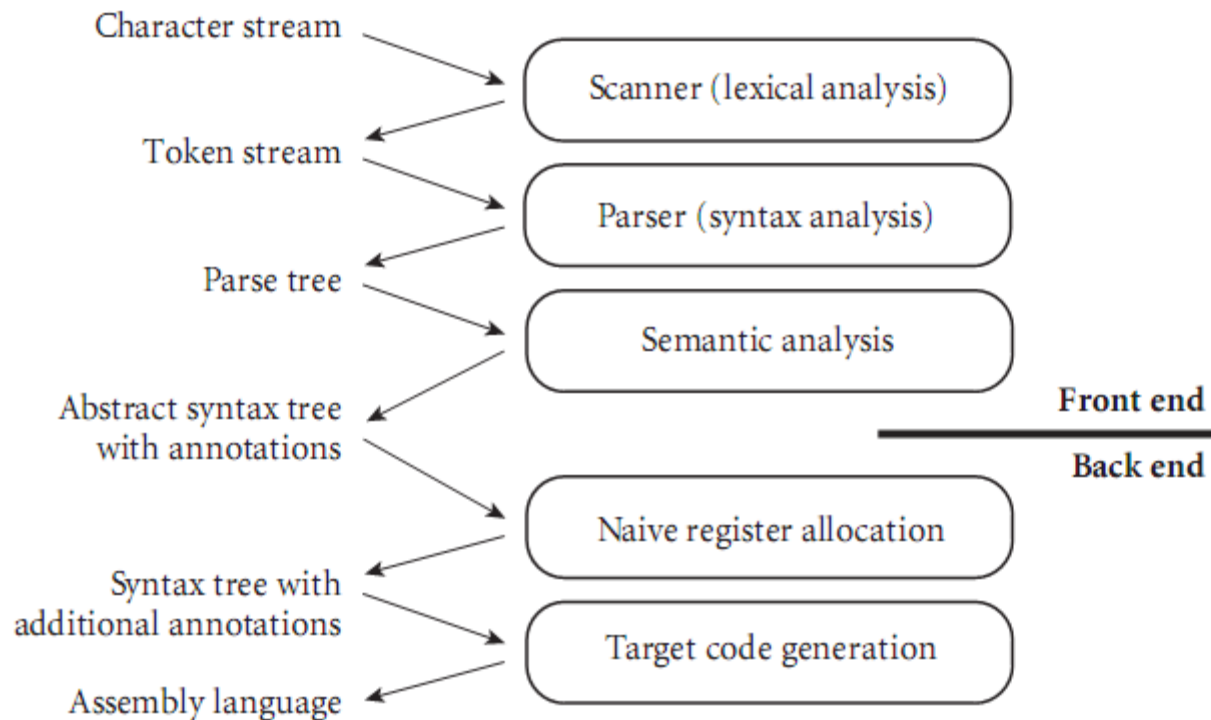
Intermediate Forms

Figure 14.5 A simpler, nonoptimizing compiler structure, assumed in Section 14.3. The target code generation phase closely resembles the intermediate code generation phase of Figure 14.1.

Code Generation

- The back end of Figure 14.1 is too complex to present in any detail in a single chapter
  - To limit the scope of our discussion, we will content ourselves in this chapter with producing correct but naive code
  - This choice will allow us to consider a significantly simpler back end.
  - Starting with Figure 14.1, we drop the machine-independent code improver and then merge intermediate and target code generation into a single phase
    - generates linear assembly language - no code improvements for control flow, therefore, there is no need to represent that flow explicitly in a control flow graph

Code Generation

- We also adopt a much simpler register allocation algorithm
  - operates directly on the syntax tree prior to code generation - eliminates need for virtual registers and the subsequent mapping onto architectural registers
- Finally, we drop instruction scheduling. The resulting compiler structure appears in Figure 14.5.
  - Its code generation phase closely resembles the intermediate code generation of Figure 14.1.
- An Attribute Grammar for GCD Example is presented in Section 14.3.1

ELSEVIER

Code Generation

- Register Allocation
  - Evaluation of the rules of the attribute grammar itself consists of two main tasks
  - In each subtree we first determine the registers that will be used to hold various quantities at run time; then we generate code.
  - Our naive register allocation strategy uses the next_free_reg inherited attribute to manage registers $r_1 \ldots r_k$ as an expression evaluation stack

- To calculate the value of (a + b) × (c - (d / e)) for example, we would generate the following:

Code Generation

$$r1 := a \qquad \text{-- push } a$$
$$r2 := b \qquad \text{-- push } b$$
$$r1 := r1 + r2 \qquad \text{-- add}$$
$$r2 := c \qquad \text{-- push } c$$
$$r3 := d \qquad \text{-- push } d$$
$$r4 := e \qquad \text{-- push } e$$
$$r3 := r3 \, / \, r4 \qquad \text{-- divide}$$
$$r2 := r2 - r3 \qquad \text{-- subtract}$$
$$r1 := r1 \times r2 \qquad \text{-- multiply}$$

Code Generation

- In a particularly complicated fragment of code it is possible to run out of architectural registers.
  - In this case we must *spill* one or more registers to memory
- Our naive register allocator pushes a register onto the program's subroutine call stack
  - In effect, architectural registers hold the top k elements of an expression evaluatio stack of effectively unlimited size
- It should be emphasized that our register allocation algorithm, makes very poor use of machine resources
- If we were generating medium level intermediate code, we would employ virtual registers, rather than architectural ones
  - Mapping of virtual registers to architectural registers would occur much later in the compilation process.
- Target code for the GCD program appears in Figure 14.7.

Address Space Organization

- Assemblers, linkers, and loaders typically operate on a pair of related file formats
  - *relocatable* object code
  - *executable* object code
- Relocatable object code is acceptable as input to a linker
  - multiple files in this format can be combined to create an executable program
- Executable object code is acceptable as input to a loader:
  - it can be brought into memory and run

ELSEVIER

Address Space Organization

- A relocatable object file includes the following descriptive information:
  - *import table:* Identifies instructions that refer to named locations whose addresses are unknown, but are presumed to lie in other files yet to be linked to this one
  - *relocation table:* Identifies instructions that refer to locations within the current file, but that must be modified at link time to reflect the offset of the current file within the final, executable program
  - *export table:* Lists the names and addresses of locations in the current file that may be referred to in other files
- Imported and exported names are known as *external symbols*

ELSEVIER

Address Space Organization

- Running program segments
  - *code*
  - *constants*
  - *initialized data*
  - *uninitialized data:* may be allocated at load time or on demand in response to page faults
    - Usually zero filled, both to provide repeatable symptoms for programs that erroneously read data they have not yet written
  - *stack:* may be allocated in some fixed amount at load time
    - more commonly, is given a small initial size, and then
    - extended automatically by the operating system in response to (faulting) accesses beyond the current segment end.
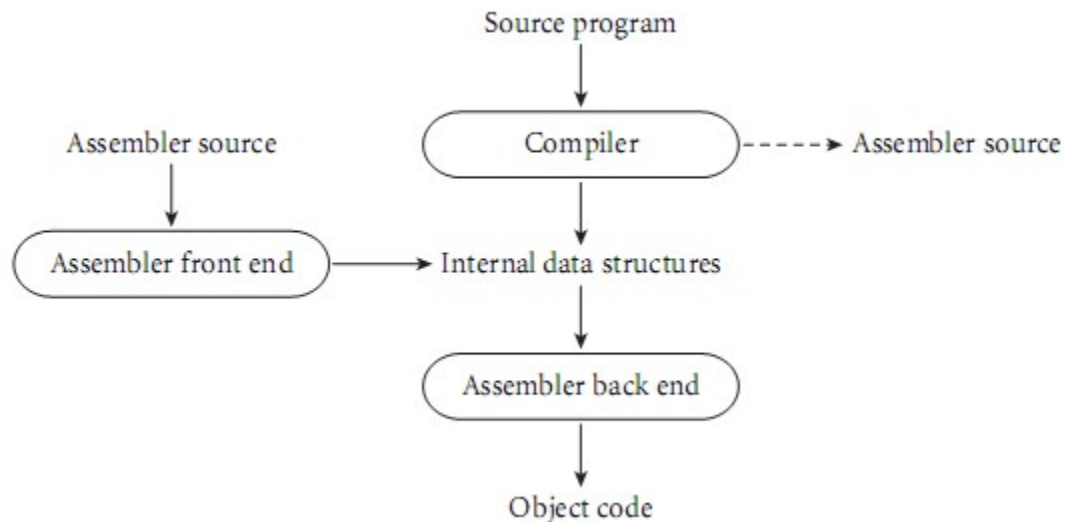
Address Space Organization

- Running program segments (2):
  - *heap:* may also be allocated in some fixed amount at load time.
    - more commonly, is given a small initial size, and is then
    - extended in response to explicit requests from heap-management library routines
  - *files:* In many systems, library routines allow a program to *map* a file into memory
    - The map routine interacts with the operating system to create a new segment for the file, and returns the address of the beginning of the segment
    - the contents of the segment are usually fetched from disk on demand, in response to page faults
  - *dynamic libraries:* Modern operating systems typically arrange for most programs to share a single copy of the code for popular libraries

Assembly

- Some compilers translate source files directly into object files acceptable to the linker
- More commonly, they generate assembly language that must subsequently be processed by an assembler to create an object file
  - symbolic (textual) notation for code.
  - within a compiler it would still be symbolic, most likely consisting of records and linked lists
- To translate this symbolic representation into executable code, we must
  - replace opcodes and operands with their machine language encodings
  - replace uses of symbolic names with actual addresses

ELSEVIER

Assembly

- When passing assembly language from the compiler to the assembler, it makes sense to use some internal (records and linked lists) representation
- At the same time, we must provide a textual front end to accommodate the occasional need for human input:

Assembly

- An alternative organization has the compiler generate object code directly
  - This organization gives the compiler a bit more flexibility: operations normally performed by an assembler (e.g., assignment of addresses to variables) can be performed earlier if desired.
  - Because there is no separate assembly pass, the overall translation to object code may be slightly faster
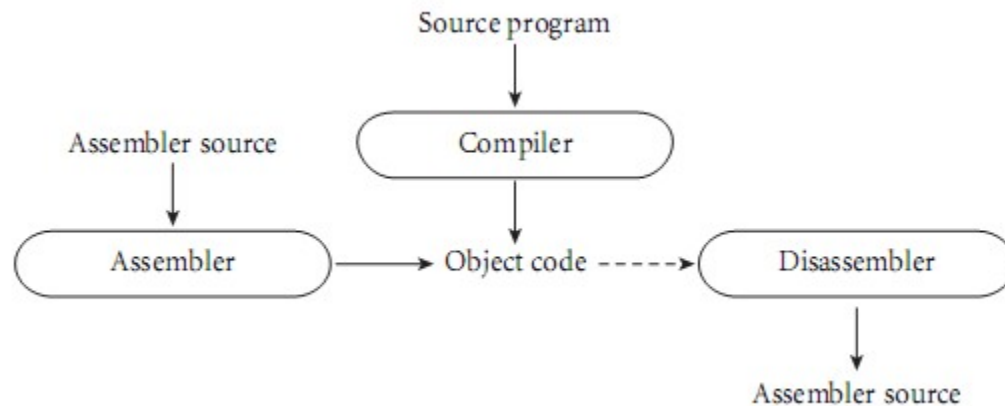
Assembly

- Emitting Instructions
  - The most basic task of the assembler is to translate symbolic representations of instructions into binary form
  - In some assemblers this is easy
    - there is a one-one correspondence between mnemonic operations and instruction op-codes
  - Many assemblers extend the instruction set in minor ways to make the assembly language easier for human beings to read
  - Most MIPS assemblers, for example, provide a large number of pseudo-instructions that translate into different real instructions depending on their arguments, or that correspond to multi-instruction sequences

Assembly

- Assemblers respond to a variety of *directives* (MIPS):
  - *segment switching*
    - .text directive indicates that subsequent instructions and data should be placed in the code (text) segment.
    - .data directive indicates that subsequent instructions and data should be placed in the initialized data segment.
    - .space n directive indicates that n bytes of space should be reserved in the uninitialized data segment
    - .byte, .half, .word, .float, and .double directives each take a sequence of arguments
    - related .ascii directive takes a single character string as argument, which it places in consecutive bytes
  - *symbol identification*
    - .global *name* directive indicates that *name* should be entered into the table of exported symbols.
  - *alignment*
    - .align n directive causes the subsequent output to be aligned at an address evenly divisible by $2n$

Assembly

- RISC assemblers implement a virtual machine - instruction set is "nicer" than that of the real hardware
  - In addition to pseudo-instructions, the virtual machine may have non-delayed branches
  - If desired, the compiler or assembly language programmer can ignore the existence of branch delays
  - The assembler will move nearby instructions to fill delay slots if possible, or generate nops if necessary.
  - To support systems programmers, the assembler must also make it possible to specify that delay slots have already been filled

Assembly

- Assemblers commonly work in several phases
  - if the input is textual, an initial phase scans and parses the input, and builds an internal representation
  - there are two additional phases.
    - first phase identifies all internal and external (imported) symbols, assigning locations to the internal ones
      - complicated by the length of some instructions (on a CISC machine)

        or
      - complicated by number of real instructions produced by a pseudo-instruction (on a RISC machine)
    - final phase produces object code

Assembly

- CISC assemblers distinguish between *absolute* and *relocatable* words in an object file
- Absolute words are known at assembly time; they need not be changed by the linker
  - constants and register-register instructions
- A relocatable word must be modified by adding to it the address within the final program of the code or data segment of the current object file
  - A CISC jump instruction might consist of a one-byte jmp opcode followed by a four-byte target address
  - For a local target, the address bytes in the object file would contain the symbol's offset within the file
  - The linker finalizes the address by adding the offset of the file's code segment within the final program

Linking

- Most language implementations - certainly all that are intended for the construction of large programs - support separate compilation
  - fragments of the program can be compiled and assembled more-or-less independently
- After compilation, these fragments (known as *compilation units*) are "glued together" by a *linker*
  - programmer explicitly divides the program into modules or files separately compiled
  - integrated environments may abandon the notion of a file in favor of a database of subroutines separately compiled
- Linker joins together compilation units

Linking

- A *static linker* does its work prior to program execution, producing an executable object file

- A *dynamic linker* does its work after the program has been brought into memory for execution

- Each of the compilation units of a program to be linked must be a relocatable object file
  - some files will have been produced by compiling fragments of the application being constructed
  - others will be general purpose library packages needed by the application

- Since most programs make use of libraries, even a "one-file" application typically needs to be linked

Linking

- Linking involves two subtasks: relocation and the resolution of external references

- Some authors refer to relocation as *loading*, and call the entire "joining together" process "link-loading."

- In this book we use "loading" to refer to the process of bringing an executable object file into memory for execution
  - on very simple machines loading entails relocation
  - the operating system uses virtual memory to giving the impression that it starts at some standard address (zero)
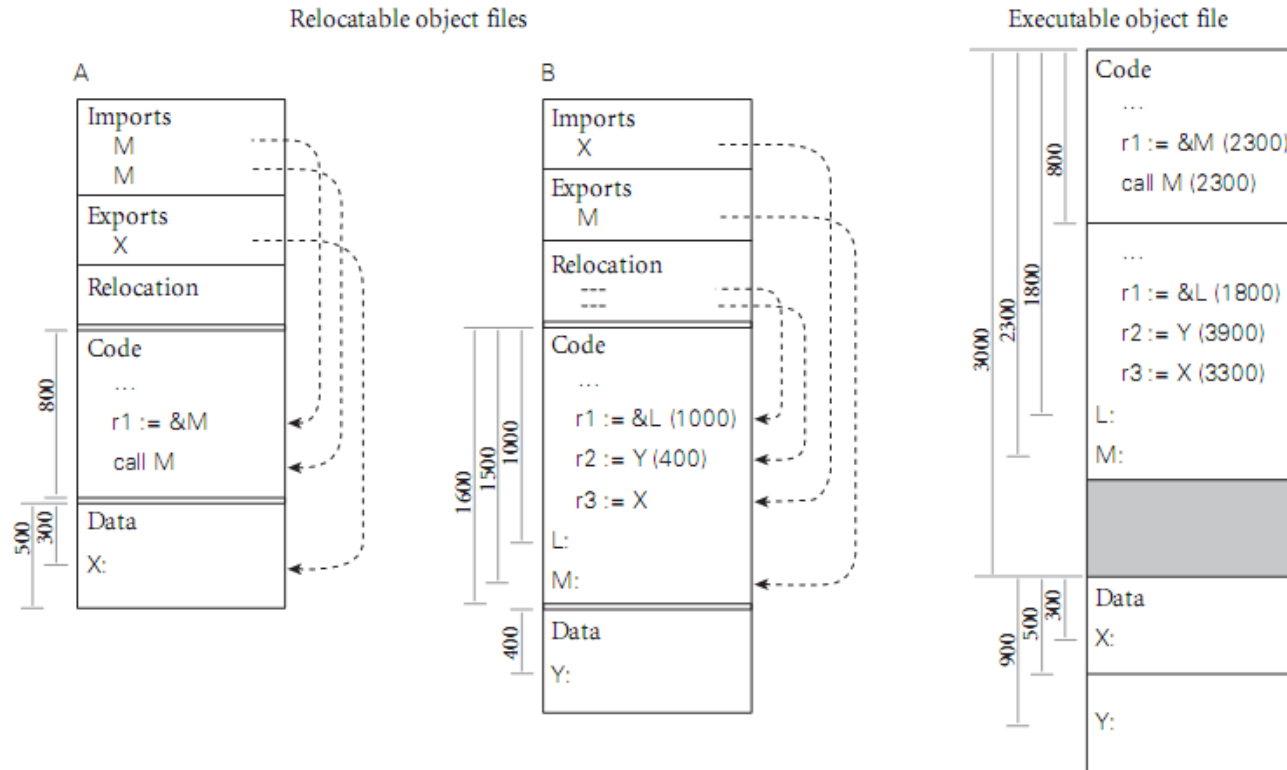 - often loading also entails a certain amount of linking

ELSEVIER

Linking

Figure 14.9 Linking relocatable object files A and B to make an executable object file. A's code section has been placed at offset 0, with B's code section immediately after, at offset 800 (addresses increase down the page). To allow the operating system to establish different protections for the code and data segments, A's data section has been placed at the next page boundary (offset 3000), with B's data section immediately after (offset 3500). External references to M and X have been set to use the appropriate addresses. Internal references to L and Y have been updated by adding in the starting addresses of B's code and data sections, respectively.

Dynamic Linking

- On a multi-user system, it is common for several instances of a program (an editor or web browser, for example) to be executing simultaneously

  - It would be highly wasteful to allocate space in memory for a separate, identical copy of the code of such a program for every running instance

- Many operating systems therefore keep track of the programs that are running, and set up memory mapping tables so that all instances of the same program share the same read-only copy of the program's code segment

  - Each instance receives its own writable copy of the data segment
  - Code segment sharing can save enormous amounts of space
  - It does not work, however, for instances of programs that are similar but not identical