Compilers: Principles, Techniques, & Tools

By Alfred V. Aho et all

First Edition

CONTENTS

CHAPTER 1

INTRODUCTION

The world as we know it depends on programming languages, because all the software running on all the computers was written in some programming language. But, before a program can be run, it first must be translated into a form in which it can be executed by a computer.

The software systems that do this translation are called *compilers*.

1 Language Processors

Simply stated, a compiler is a program that can read a program in one language–the *source* language–and translate it into an equivalent program in another language–the *target* language.

An *interpreter* is another common kind of language processor.

The task of collecting the source program is sometimes entrusted to a separate program, called a *preprocessor*.

The compiler may produce an assembly-language program as its output, because assembly language is easier to produce as output and is easier to debug. The assembly language is then processed by a program called an *assembler* that produces relocatable machine code as its output.

The *linker* resolves external memory addresses, where the code in one file may refer to a location in another file. The *loader* then puts together all of the executable object files into memory for execution.

2 The Structure of a Compiler

Up to this point we have treated a compiler as a single box that maps a source program into a semantically equivalent target program. If we open up this box a little, we see that there are two parts to this mapping: analysis and synthesis.

The *analysis* part breaks up the source program into constituent pieces and imposes a grammatical structure on them. The analysis part also collects information about the source program and stores it in a data structure called a *symbol table*, which is passed along with the intermediate representation to the synthesis part.

The *synthesis* part constructs the desired target program from the intermediate representation and the information in the symbol table. The analysis part is often called the *front end* of the compiler; the synthesis part is the *back end*.

If we examine the compilation process in more detail, we see that it operates as a sequence of *phases*, each of which transforms one representation of the source program to another.

2.1 Lexical Analysis

The first phase of a compiler is called *lexical analysis* or *scanning*. The lexical analyzer reads the stream of characters making up the source program and groups the characters into meaningful sequences called *lexemes*. For each lexeme, the lexical analyzer produces as output a *token* of the form

$$\langle \textit{token-name, attribute-value} \rangle$$

that is passes on to the subsequent phase, syntax analysis. In the token, the first component *token-name* is an abstract symbol that is used during syntax analysis, and the second component *attribute-value* points to an entry in the symbol table for this token.

2.2 Syntax Analysis

The second phase of the compiler is *syntax analysis* or *parsing*. A typical representation is a *syntax tree* in which each interior node represents an operation and the children of the node represent the arguments of the operation.

2.3 Semantic Analysis

The *semantic analyzer* uses the syntax tree and the information in the symbol table to check the source program for semantic consistency with the language definition.

An important part of semantic analysis is *type checking*, where the compiler checks that each operator has matching operands.

The language specification may permit some type conversions called *coercions*.

2.4 Intermediate Code Generation

We consider an intermediate form called *three-address code*, which consists of a sequence of assembly-like instructions with three operands per instruction.

2.5 The Grouping of Phases into Passes

In an implementation, activities from several phases may be grouped together into a *pass* that reads an input file and writes an output file.

2.6 Compiler-Construction Tools

Some commonly used compiler-construction tools include

1. *Parser generators* that automatically produce syntax analyzers from a grammatical description of a programming language.
2. *Scanner generators* that produce lexical analyzers from a regular-expression description of the tokens of a language.
3. *Syntax-directed translation engines* that produce collections of routines for walking a parse tree and generating intermediate code.
4. *Code-generator generators* that produce a code generator from a collection of rules for translating each operation of the intermediate language into the machine language for a target machine.
5. *Data-flow analysis engines* that facilitate the gathering of information about how values are transmitted from one part of a program to each other part.
6. *Compiler-construction toolkits* that provide an integrated set of routines for construction various phases of a compiler.

3 The Evolution of Programming Language

3.1 The Move to Higher-Level Languages

One classification is by generation. *First-generation languages* are the machine languages, *second-generation* the assembly languages, and *third-generation* the higher-level languages. *Fourth-generation languages* are languages designed for specific applications. The term *fifth-generation language* has been applied to logic- and constraint-based languages.

Another classification of languages uses the term *imperative* for languages in which a program specifies *how* a computation is to be done and *declarative* for languages in which a program specifies *what* computation is to be done.

The term *von Neumann language* is applied to programming languages whose computational model is based on the von Neumann computer architecture.

An *object-oriented language* is one that supports object-oriented programming, a programming style in which a program consists of a collection of objects that interact with one another.

Scripting languages are interpreted languages with high-level operators designed for "gluing together" computations.