Table of Contents

[1. Introduction 12](#_Toc59126052)

[1.1. Overview 12](#_Toc59126053)

[1.2. The Compiler Phases 13](#_Toc59126054)

[1.2.1. The Preprocessor 13](#_Toc59126055)

[1.2.2. Scanning 13](#_Toc59126056)

[1.2.3. Parsing and Middle Code Generation 13](#_Toc59126057)

[1.2.4. Middle Code Optimization 14](#_Toc59126058)

[1.2.5. Initialization 14](#_Toc59126059)

[1.2.6. Static Address 14](#_Toc59126060)

[1.2.7. Declarators and Declaration Specifiers 14](#_Toc59126061)

[1.2.8. The Symbol Table 15](#_Toc59126062)

[1.2.9. The Type System 15](#_Toc59126063)

[1.2.10. Assembly Code Generation 15](#_Toc59126064)

[1.2.1. Register Allocation 15](#_Toc59126065)

[1.2.2. The Object Code Generator and Linker 16](#_Toc59126066)

[1.2.3. The Standard Library 16](#_Toc59126067)

[1.3. The Main Class 16](#_Toc59126068)

[1.3.1. Generating the Assembly File 17](#_Toc59126069)

[1.3.1. Generate the Make File 20](#_Toc59126070)

[1.3.2. Is the Object File Fresh? 22](#_Toc59126071)

[2. Scanning 25](#_Toc59126072)

[2.1. Slash Sequences 25](#_Toc59126073)

[2.2. The typedef-name Problem 27](#_Toc59126074)

[2.3. The Scanner 28](#_Toc59126075)

[3. Parsing 36](#_Toc59126076)

[3.1. The MiddleOperator Enumeration 36](#_Toc59126077)

[3.2. Declarations 39](#_Toc59126078)

[3.2.1. Function Definition 40](#_Toc59126079)

[3.2.2. Specifier List 41](#_Toc59126080)

[3.2.3. Structs and Unions 42](#_Toc59126081)

[3.2.4. Enumeration 43](#_Toc59126082)

[3.2.5. Declarators 45](#_Toc59126083)

[3.2.6. Pointer Declarators 45](#_Toc59126084)

[3.2.7. Direct Declarator 46](#_Toc59126085)

[3.2.8. Initialization 48](#_Toc59126086)

[3.2.9. Abstract Declarator 49](#_Toc59126087)

[3.3. Statements 49](#_Toc59126088)

[3.3.1. The if-else Problem 50](#_Toc59126089)

[3.3.1. The Forward-Jump Problem (Backpatching) 51](#_Toc59126090)

[3.3.2. The Statement Class 53](#_Toc59126091)

[3.3.3. Statements 54](#_Toc59126092)

[3.3.4. The If Statement 54](#_Toc59126093)

[3.3.1. The If-Else Statement 54](#_Toc59126094)

[3.3.2. The Switch Statement 54](#_Toc59126095)

[3.3.3. The Case Statement 54](#_Toc59126096)

[3.3.1. The Default Statement 55](#_Toc59126097)

[3.3.2. The While Statement 55](#_Toc59126098)

[3.3.3. The Do Statement 55](#_Toc59126099)

[3.3.4. The For Statement 55](#_Toc59126100)

[3.3.5. Label and Jump Statement 55](#_Toc59126101)

[3.3.6. The Return Statement 56](#_Toc59126102)

[3.3.7. Optional Expression Statements 56](#_Toc59126103)

[3.3.8. Block Statements 56](#_Toc59126104)

[3.3.9. Jump Register Statements 57](#_Toc59126105)

[3.3.10. Interrupt Statements 57](#_Toc59126106)

[3.3.11. System Call Statements 57](#_Toc59126107)

[3.4. Expressions 57](#_Toc59126108)

[3.4.1. The Expression Class (Short and Long List) 57](#_Toc59126109)

[3.4.2. Optional Expressions 58](#_Toc59126110)

[3.4.3. The Comma Expression 59](#_Toc59126111)

[3.4.4. The Assignment Expression 59](#_Toc59126112)

[3.4.1. The Condition Expression 59](#_Toc59126113)

[3.4.1. Constant Expression 60](#_Toc59126114)

[3.4.1. Logical Expressions 60](#_Toc59126115)

[3.4.2. Bitwise Expressions 60](#_Toc59126116)

[3.4.1. Shift Expression 61](#_Toc59126117)

[3.4.2. Equality and Relation Expressions 61](#_Toc59126118)

[3.4.3. Addition and Subtraction Expression 62](#_Toc59126119)

[3.4.4. Multiplication Expressions 62](#_Toc59126120)

[3.4.5. Cast Expressions 63](#_Toc59126121)

[3.4.6. Prefix Expression 63](#_Toc59126122)

[3.4.7. Postfix Expression 64](#_Toc59126123)

[3.4.8. Primary Expressions 65](#_Toc59126124)

[4. Middle Code Generation 67](#_Toc59126125)

[4.1. The MiddleCode Class 67](#_Toc59126126)

[4.1.1. Backpatching 70](#_Toc59126127)

[4.2. Declarations 71](#_Toc59126128)

[4.2.1. Function Definition 71](#_Toc59126129)

[4.2.2. Specifier List 75](#_Toc59126130)

[4.2.3. Structs and Unions 75](#_Toc59126131)

[4.2.4. Enumeration 76](#_Toc59126132)

[4.2.5. Declarator 77](#_Toc59126133)

[4.2.6. Pointer declarator 80](#_Toc59126134)

[4.2.7. Direct Declarator 80](#_Toc59126135)

[4.2.8. Parameters 82](#_Toc59126136)

[4.2.9. Abstract Declarator 83](#_Toc59126137)

[4.3. Statements 83](#_Toc59126138)

[4.3.1. The If Statement 83](#_Toc59126139)

[4.3.2. The If-Else Statement 84](#_Toc59126140)

[4.3.3. The Switch Statement 84](#_Toc59126141)

[4.3.4. The Case Statement 86](#_Toc59126142)

[4.3.5. The Default Statement 87](#_Toc59126143)

[4.3.6. The While Statement 87](#_Toc59126144)

[4.3.7. The Do Statement 88](#_Toc59126145)

[4.3.8. The For Statement 88](#_Toc59126146)

[4.3.9. The Label and Goto Statements 89](#_Toc59126147)

[4.3.10. Return Statement 90](#_Toc59126148)

[4.3.11. Optional Expression Statement 91](#_Toc59126149)

[4.3.12. Jump Register Statements 92](#_Toc59126150)

[4.3.1. Interrupt Statements 92](#_Toc59126151)

[4.3.1. System Call Statements 92](#_Toc59126152)

[4.4. Expressions 92](#_Toc59126153)

[4.4.1. The Comma Expression 92](#_Toc59126154)

[4.4.2. The Assignment Expression 93](#_Toc59126155)

[4.4.3. The Condition Expression 95](#_Toc59126156)

[4.4.4. Constant Expression 98](#_Toc59126157)

[4.4.5. The Logical Or Expression 98](#_Toc59126158)

[4.4.1. The Logical And Expression 99](#_Toc59126159)

[1.1.1. Bitwise Expressions 100](#_Toc59126160)

[4.4.2. Shift Expression 101](#_Toc59126161)

[4.4.3. Equality and Relation Expressions 101](#_Toc59126162)

[4.4.4. Addition and Subtraction Expression 103](#_Toc59126163)

[4.4.5. Multiplication Expressions // XXX 106](#_Toc59126164)

[4.4.6. Cast Expressions 107](#_Toc59126165)

[4.4.7. Unary Addition Expressions 107](#_Toc59126166)

[4.4.8. Logical Not Expression 107](#_Toc59126167)

[4.4.9. Bitwise Not Expression 108](#_Toc59126168)

[4.4.10. The sizeof Expression 108](#_Toc59126169)

[4.4.11. Address Expression // XXX 109](#_Toc59126170)

[4.4.12. Dereference Expression 109](#_Toc59126171)

[4.4.13. Prefix Increment and Decrement Expression 109](#_Toc59126172)

[4.4.14. Postfix Increment and Decrement Expression 111](#_Toc59126173)

[4.4.15. Arrow Expression 112](#_Toc59126174)

[4.4.16. Index Expression 112](#_Toc59126175)

[4.4.1. Dot Expression 114](#_Toc59126176)

[4.4.2. Function Call Expression 114](#_Toc59126177)

[4.4.3. Argument Expression List 117](#_Toc59126178)

[4.4.4. Primary Expressions 118](#_Toc59126179)

[5. Declaration Specifiers and Declarators 120](#_Toc59126180)

[5.1. Declarators 126](#_Toc59126181)

[6. The Symbol Table 128](#_Toc59126182)

[6.1. The Symbol 134](#_Toc59126183)

[6.2. The Static Symbol 140](#_Toc59126184)

[7. The Type System 143](#_Toc59126185)

[7.1. The Sort Enumeration 143](#_Toc59126186)

[7.2. The Type Class 143](#_Toc59126187)

[7.2.1. Integral, Floating, and String 143](#_Toc59126188)

[7.2.2. Logical Types 145](#_Toc59126189)

[7.2.3. Bitfields 145](#_Toc59126190)

[7.2.4. Pointers 146](#_Toc59126191)

[7.2.5. Arrays 146](#_Toc59126192)

[7.2.6. Functions 147](#_Toc59126193)

[7.2.7. Structs and Unions 149](#_Toc59126194)

[7.2.8. Enumerations 150](#_Toc59126195)

[7.2.9. Type Size 151](#_Toc59126196)

[7.2.10. Complete Types 152](#_Toc59126197)

[7.2.11. Constant and Volatile 152](#_Toc59126198)

[7.2.12. Hash Code and Equals 153](#_Toc59126199)

[7.2.13. Predefined Types 154](#_Toc59126200)

[7.2.14. ToString 154](#_Toc59126201)

[7.3. Type Size 154](#_Toc59126202)

[7.4. Type Casting 157](#_Toc59126203)

[7.4.1. Implicit Cast 158](#_Toc59126204)

[7.4.2. Explicit Cast 158](#_Toc59126205)

[7.4.3. Type Promotion 162](#_Toc59126206)

[8. Constant Expression 164](#_Toc59126207)

[8.1. Unary and Binary Expressions 164](#_Toc59126208)

[8.1.1. Relation Expressions 164](#_Toc59126209)

[8.1.2. Arithmetic Expressions 168](#_Toc59126210)

[8.2. Constant Type Cast 174](#_Toc59126211)

[8.2.1. Constant Value 176](#_Toc59126212)

[9. Static Address Expression 178](#_Toc59126213)

[9.1.1. Static Value and Address 178](#_Toc59126214)

[9.1.2. Static Expression 179](#_Toc59126215)

[10. Initialization 183](#_Toc59126216)

[10.1.1. Auto Initialization 183](#_Toc59126217)

[10.1.1. Static Initialization 186](#_Toc59126218)

[10.1.2. Modify Initializer 189](#_Toc59126219)

[11. Middle Code Optimization 192](#_Toc59126220)

[11.1.1. Object to Integer Addresses 193](#_Toc59126221)

[11.1.2. Jump Next Instructions 193](#_Toc59126222)

[11.1.3. Next-Double Jump Statements 194](#_Toc59126223)

[11.1.4. Jump Chains 195](#_Toc59126224)

[11.1.5. Remove Unreachable Code 196](#_Toc59126225)

[11.1.1. Remove Push-Pop Chains 197](#_Toc59126226)

[11.1.1. Merge Pop-Push Chains 197](#_Toc59126227)

[11.1.2. Change Top-Pop to Pop 198](#_Toc59126228)

[11.1.3. Merge Binary 198](#_Toc59126229)

[11.1.4. Semantic Optimization 199](#_Toc59126230)

[11.1.5. Optimize Relation Expression 202](#_Toc59126231)

[11.1.1. Optimize Communicative Expression 203](#_Toc59126232)

[11.1.2. Remove Trivial Assignment 203](#_Toc59126233)

[11.1.3. Remove Cleared Code 204](#_Toc59126234)

[12. Assembly Code Generation 205](#_Toc59126235)

[12.1. Runtime Management 205](#_Toc59126236)

[12.2. Assembly Operator 207](#_Toc59126237)

[12.3. Assembly Code 208](#_Toc59126238)

[12.3.1. Assembly Code Optimization 209](#_Toc59126239)

[12.3.2. Operator Test Methods 211](#_Toc59126240)

[12.3.3. Register Overlapping 212](#_Toc59126241)

[12.3.4. Register Size 213](#_Toc59126242)

[12.3.5. ToString 216](#_Toc59126243)

[12.4. Tracks 220](#_Toc59126244)

[12.5. Register Allocation 222](#_Toc59126245)

[12.6. Assembly Code Generation 227](#_Toc59126246)

[12.6.1. The Long Switch Statement 228](#_Toc59126247)

[12.6.2. Track Set Generation 233](#_Toc59126248)

[12.6.3. Function Calls 234](#_Toc59126249)

[12.6.4. Loading Values into Registers 239](#_Toc59126250)

[12.6.5. Return, Exit, and Jump 242](#_Toc59126251)

[12.6.6. Load and Inspect Registers 245](#_Toc59126252)

[12.6.7. Initialization 246](#_Toc59126253)

[12.6.8. Integral Multiplication, Division, and Modulo 246](#_Toc59126254)

[12.6.1. Integral Assignment and Parameters 248](#_Toc59126255)

[12.6.2. Unary Integral Operations 251](#_Toc59126256)

[12.6.3. Integral Binary 253](#_Toc59126257)

[12.6.1. Base and Offset 257](#_Toc59126258)

[12.6.2. Case 258](#_Toc59126259)

[12.6.3. Address 258](#_Toc59126260)

[12.6.1. Floating Binary 259](#_Toc59126261)

[12.6.1. Floating Relation 260](#_Toc59126262)

[12.6.2. Floating Push and Pop 260](#_Toc59126263)

[12.6.3. Type Conversion 263](#_Toc59126264)

[12.6.4. Struct and Union 265](#_Toc59126265)

[12.6.5. Initialization Code 266](#_Toc59126266)

[12.6.6. Command Line Arguments 267](#_Toc59126267)

[12.6.1. Text List 270](#_Toc59126268)

[13. Executable Code Generation 272](#_Toc59126269)

[13.1. The Windows environment 272](#_Toc59126270)

[13.1.1. Main 272](#_Toc59126271)

[13.1.2. Type Size 275](#_Toc59126272)

[13.1.1. Static Symbol 276](#_Toc59126273)

[13.1.2. Static Value 279](#_Toc59126274)

[13.1.3. Function End 279](#_Toc59126275)

[13.1.4. Target Code Generation 280](#_Toc59126276)

[13.1.5. Exit 280](#_Toc59126277)

[13.1.6. Initialization Code 280](#_Toc59126278)

[13.1.7. Command Line Arguments 281](#_Toc59126279)

[13.1.1. Windows Jump Info 286](#_Toc59126280)

[13.1.2. Windows Byte List 289](#_Toc59126281)

[13.1.3. Byte List 290](#_Toc59126282)

[13.2. Linking 297](#_Toc59126283)

[13.2.1. The Linker Class 298](#_Toc59126284)

[A. The Standard Library 304](#_Toc59126285)

[13.3. Integral and Floating Limits 304](#_Toc59126286)

[13.4. The assert Macro 305](#_Toc59126287)

[13.5. Locale Data 306](#_Toc59126288)

[13.6. Character Types 308](#_Toc59126289)

[13.7. Strings 311](#_Toc59126290)

[13.8. Long Jumps 317](#_Toc59126291)

[13.9. Mathematical Functions 319](#_Toc59126292)

[13.9.1. Exponent and Logarithm Functions 320](#_Toc59126293)

[13.9.2. Power Functions 322](#_Toc59126294)

[13.9.1. Square Root 323](#_Toc59126295)

[13.9.1. Modulo Functions 324](#_Toc59126296)

[13.9.2. Trigonometric Functions 324](#_Toc59126297)

[13.9.3. Inverted Trigonometric Functions 325](#_Toc59126298)

[13.9.4. Hyperbolic Trigonometric Functions 328](#_Toc59126299)

[13.9.5. Floor, Ceiling, Absolute, and Rounding Functions 328](#_Toc59126300)

[13.10. Standard Input/Output 329](#_Toc59126301)

[13.10.1. Printing 329](#_Toc59126302)

[13.10.2. Scanning 341](#_Toc59126303)

[13.10.3. File Handling 352](#_Toc59126304)

[13.11. The Standard Library 362](#_Toc59126305)

[13.12. Time 368](#_Toc59126306)

[B. The Preprocessor 375](#_Toc59126307)

[13.13. The Expression Scanner and Parser 375](#_Toc59126308)

[13.13.1. The Grammar 375](#_Toc59126309)

[13.13.2. The Parser 376](#_Toc59126310)

[13.13.3. The Scanner 380](#_Toc59126311)

[13.14. The Preprocessor Scanner and Parser 381](#_Toc59126312)

[13.14.1. If-Else-Chain 383](#_Toc59126313)

[13.15. The Preprocessor 384](#_Toc59126314)

[13.15.1. Tri Graphs 386](#_Toc59126315)

[13.15.1. Comments, Strings, and Characters 387](#_Toc59126316)

[13.15.2. Slash Codes 389](#_Toc59126317)

[13.15.3. The Line List 389](#_Toc59126318)

[13.15.4. Lines 393](#_Toc59126319)

[13.15.5. Includes 393](#_Toc59126320)

[13.15.6. Macros 394](#_Toc59126321)

[13.15.7. Conditional Programming 399](#_Toc59126322)

[13.15.8. Macro Expansion 401](#_Toc59126323)

[13.15.9. String Merging 404](#_Toc59126324)

[13.16. Register Allocation 405](#_Toc59126325)

[C. The Register Set 406](#_Toc59126326)

[D. The C Grammar 407](#_Toc59126327)

[13.17. The Preprocessor Grammar 407](#_Toc59126328)

[13.18. The Language Grammar 407](#_Toc59126329)

[E. The Gardens Point Tools 414](#_Toc59126330)

[13.18.1. The Language 414](#_Toc59126331)

[13.18.2. The Grammar 415](#_Toc59126332)

[13.18.3. GPPG 416](#_Toc59126333)

[13.18.4. JPlex 420](#_Toc59126334)

[13.18.5. Main 422](#_Toc59126335)

[F. Auxiliary Classes 424](#_Toc59126336)

[13.19. Error Handling 424](#_Toc59126337)

[13.20. Container Classes 428](#_Toc59126338)

[13.20.1. Ordered Pair 428](#_Toc59126339)

[13.21. Graph 429](#_Toc59126340)

[13.21.1. Addition and Removal of Vertices and Edges 430](#_Toc59126341)

[13.21.2. Graph Partition 430](#_Toc59126342)

[G. The ASCII Table 432](#_Toc59126343)

[H. A Crash Course in C 433](#_Toc59126344)

[13.22. The Compiler and the Linker 433](#_Toc59126345)

[13.23. The Hello-World Program 434](#_Toc59126346)

[13.24. Comments 435](#_Toc59126347)

[13.25. Types and Variables 435](#_Toc59126348)

[1.1.1 Simple Types 435](#_Toc59126349)

[2.1.1 Variables 436](#_Toc59126350)

[3.1.1 Constants 437](#_Toc59126351)

[4.1.1 Output 437](#_Toc59126352)

[5.1.1 Input 438](#_Toc59126353)

[6.1.1 Enumerations 439](#_Toc59126354)

[7.1.1 Arrays 440](#_Toc59126355)

[8.1.1 Characters and Strings 440](#_Toc59126356)

[9.1.1 The ASCII Table 441](#_Toc59126357)

[10.1.1 Pointers 441](#_Toc59126358)

[11.1.1 Pointers and Dynamic Memory 442](#_Toc59126359)

[12.1.1 Defining our own types 444](#_Toc59126360)

[13.1.1 The Type Size 444](#_Toc59126361)

[13.26. Expressions and Operators 445](#_Toc59126362)

[14.1.1 Arithmetic Operators 445](#_Toc59126363)

[15.1.1 Pointer Arithmetic 445](#_Toc59126364)

[16.1.1 Increment and decrement 447](#_Toc59126365)

[17.1.1 Relational Operators 447](#_Toc59126366)

[18.1.1 Logical Operators 447](#_Toc59126367)

[19.1.1 Bitwise Operators 448](#_Toc59126368)

[20.1.1 Assignment 448](#_Toc59126369)

[21.1.1 The Condition Operator 449](#_Toc59126370)

[22.1.1 Precedence and Associatively 449](#_Toc59126371)

[13.27. Statements 450](#_Toc59126372)

[23.1.1 Selection statements 450](#_Toc59126373)

[24.1.1 Iteration statements 452](#_Toc59126374)

[25.1.1 Jump statements 453](#_Toc59126375)

[26.1.1 Expression statements 454](#_Toc59126376)

[27.1.1 Volatile 454](#_Toc59126377)

[13.28. Functions 454](#_Toc59126378)

[13.28.1. Void Functions 455](#_Toc59126379)

[13.28.2. Local and global variables 456](#_Toc59126380)

[13.28.3. Inner Blocks with Local Variables 456](#_Toc59126381)

[13.28.4. Call-by-Value and Call-by-Reference 458](#_Toc59126382)

[13.28.5. Static, Extern, and Register Variables 459](#_Toc59126383)

[13.28.6. Recursion 460](#_Toc59126384)

[13.28.7. Definition and Declaration 460](#_Toc59126385)

[13.28.8. Higher-Order Functions 461](#_Toc59126386)

[13.29. Structures and Linked Lists 463](#_Toc59126387)

[13.29.1. Stacks and Linked Lists 463](#_Toc59126388)

[13.29.2. Unions 466](#_Toc59126389)

[13.29.3. Bitfields 467](#_Toc59126390)

[13.30. Structures with function pointers 467](#_Toc59126391)

[13.31. The Preprocessor 468](#_Toc59126392)

[13.32. The Standard Library 469](#_Toc59126393)

[13.32.1. File Management 469](#_Toc59126394)

[13.32.2. Program Parameters 470](#_Toc59126395)

[13.32.3. Environment Functions 471](#_Toc59126396)

[13.32.4. Searching and Sorting 472](#_Toc59126397)

[13.32.5. Functions with Variable Number of Parameters 472](#_Toc59126398)

[13.32.6. String Management 473](#_Toc59126399)

[13.32.7. Memory Management 474](#_Toc59126400)

[13.32.8. Mathematical Functions 475](#_Toc59126401)

[13.32.9. Long Jumps 475](#_Toc59126402)

[13.32.10. Time 476](#_Toc59126403)

[13.32.11. Random Numbers 476](#_Toc59126404)

[13.32.12. Limits of Integral and Floating Types 477](#_Toc59126405)

[13.32.13. Character Functions 477](#_Toc59126406)

[13.33. Further Reading 478](#_Toc59126407)

[Foreword 479](#_Toc59126408)

# Introduction

This is a book about compiler construction. More specifically, about the construction of an optimized compiler for the ANSI C. It can be set to generate two kinds of target code:

Assembly code for the Intel 64-bit Linux system together with a makefile with instructions for further assemblering and linking.

A file in the .com file format holding assembled and linked code the 16-bit Windows system, ready to be executed directly.

## Overview

The compiler is divided into several phases. The input is ANSI C source code, the output is Linux assembly code or Windows executable code.

Preprocessor

ANSI C Source Code

Preprocessed Source Code

Parser

Scanner

Symbol Table

Type System

Middle Code

Middle Code Optimizer

Optimized Middle Code

Assembly Code Generator

64-bit Linux Assembly Code

Object Code Generator

Object Code

Linker

16-bit Windows Executable Code

The compiler is made up be a sequence of phases.

Preprocessor

ANSI C Source Code

Preprocessed Source Code

Parser

Scanner

Symbol Table

Type System

Middle Code

Middle Code Optimizer

Optimized Middle Code

Assembly Code Generator

64-bit Linux Assembly Code

Object Code Generator

Object Code

Linker

16-bit Windows Executable Code

## The Compiler Phases

The compiler is made up be a sequence of phases, which each take code in some form and generates more refined code.

### The Preprocessor

As the name implies, the preprocessor processes the source code before the actual compilation starts. Its task is to remove comments, include files, replace macros, and perform conditional programming. We look into the preprocessor in Appendix 0.

### Scanning

The scanner is responsible for interpreting sequences of characters into tokens: the least significant parts of the source code. For instance, the characters ‘i’ and ‘f’ are interpreted as the keyword if and the characters ‘3’, ‘.’, ‘1’, and ‘4’ are interpreted as the floating-point value 3.14.

### Parsing and Middle Code Generation

Every programming language has a syntax, that describes the form of the code. In this book, the syntax is defined a grammar, described in Appendix 13.16 The parser checks whether the source code is correct with regard to the grammar by requesting tokens from the scanner when needed. When the declarations are parsed the symbol table is generated, which holds information of variables, types, and functions. The type system is also used to perform type checking and type castings. The output of the parser is a sequence of middle code. As the name implies, the middle code is a simple notation holding the code between the parsing and the final target code generation. More specifically, each instruction can refer to three values at most; therefore, the middle code notation is called three-address-code.

Below is an example of how a sample of C code is translated into the middle code.

|  |  |
| --- | --- |
| if (a < b)  a = -b;  else  b = -a; | 1. if a < b goto 3  2. goto 5  3. a = -b;  4. goto 6  5. b = -a  6. ... |
| (a) C code | (b) Middle code |

As each middle code instruction can refer to three values at the most, we introduce temporary values to hold sub expressions.

|  |  |
| --- | --- |
| x = (a + b) \* (c - d); | 1. temp1 = a + b  2. temp2 = c - d  3. temp3 = temp1 \* temp2  4. x = temp3 |
| (a) C code | (b) Middle code |

### Middle Code Optimization

The purpose of the middle code optimization phase is to make the middle code more effective, to provide for more effective assembly code generation. Reduction of condition jumps is one example of such optimization. For instance, by swapping the condition in if statement of the example below, we can remove a goto instruction.

|  |  |
| --- | --- |
| 1. if a < b goto 3  2. goto 5  3. a = -b;  4. goto 6  5. b = -a  6. ... | 1. if a >= b goto 4  2. a = -b;  3. goto 5  4. b = -a;  5. ... |
| (a) Before | (b) After |

### Initialization

A constant or variable may be initialized. The initializer may be a single value, a list of values, or a list of other lists. In the following example, and two-dimensional array is initialized by lists.

int a[][3] = {{1, 2, 3}, {4, 5, 6}};

However, the same array may also be initialized by one list.

int a[][3] = {1, 2, 3, 4, 5, 6};

### Static Address

A static address is an address which value is determined in compile-time. For instance, the &a[3] expression is calculated to a specific constant address by the compiler and linker.

### Declarators and Declaration Specifiers

In the following code, struct {int i;} is a specifier while s and \*p are declarators.

struct {int i;} s, \*p;

A declarator can also by initialized, and in a struct or union be marked as a bitfield.

struct {int i;} s = {1}, \*p = NULL;

struct {int i : 3;};

### The Symbol Table

The symbol table holds information of the symbol of the code, such as variables, constants, and functions as well as struct and union tags of extern, static, auto, register, and typedef storage. The symbol table is actually a hierarchy of tables, where the root table represent global space and the other tables represents functions or blocks in functions, or members of structs or unions.

### The Type System

ANSI C holds a rather large set of types. The basic types are the integral types and floating-point types. Then there are pointers, arrays, function, structs and unions. It is possible to type cast between several of these types. Moreover, in some situation is an array interpreted as a pointer its type, and a function is interpreted as a pointer to the function.

### Assembly Code Generation

When the middle code has been generated and optimized, we generate the assembly code. In the first step, we generate the assembly and use tracks to mark where in the assembly code the yet unknown registers shall be placed. In the second step we perform register allocation to find the optimal us of the register set, and finally replace the tracks with the actual register.

### Register Allocation

The register allocator creates a graph where the vertices hold tracks, and two vertices are connected by an edge if their tracks overlaps in the code. Then we find a match where two overlapping tracks are assigned non-overlapping registers. Let us look at the following example, we assume that x, a, b, c, and d are static 16-bit variables stored at specific addresses.

|  |  |
| --- | --- |
| x = a + b & c - d; | 1. temporary0 = a + b  2. temporary1 = c - d  3. temporary2 = t0 & t1  4. x = temporary2 |
| (a) C Code | (b) Middle Code |

In the assembly code generation, the registers are represented by tracks, that are replaced by proper registers by the register allocator.

|  |  |
| --- | --- |
| mov track0, [aaddress]  add track0, [baddress]  mov track1, [caddress]  sub track1, [daddress]  and track0, track1  mov [xaddress], track0 | mov ax, [aaddress]  add ax, [baddress]  mov bx, [caddress]  sub bx, [daddress]  and ax, bx  mov [xaddress], ax |
| (c) Assembly Code with Tracks | (d) Final Assembly Code |

In the final assembly code, we cannot assign track0 and track1 the same register, since they overlap.

### The Object Code Generator and Linker

At mentioned at the beginning of this chapter, there are two target systems: 64-bit Linux and 16-bit Windows. In the latter case, we generate files with object code and link the object code together in a file executable in the .com file format. However, we will ignore this alternative until Chapter 13, and focus on the 64-bit Linux alternative up until then.

### The Standard Library

The C language comes with a set of functions, macros, and instructions for many services. It is mostly implemented in C. However, there are some additional non-standard elements for accessing registers and perform system calls.

## The Main Class

The Main class handles the overall compiling and linking. It compiles the source code and generates the assembly files together with a makefile, that is used to assembly and link the code to the final executable file.

Main.cs

using System;

using System.IO;

using System.Text;

using System.Globalization;

using System.Collections.Generic;

namespace CCompiler {

public class Start {

First of all, there is the Linux and Windows fields, that indicates whether the output of the compiler is a set of assembly files and a makefile for the Linux system, or an executable file for the Windows system. Throughout this book, there will be several statements that test whether the Windows or Linux field is true. We will look into the Linux cases through the first part of this book and wait with the Windows case until Chapter 13.

public static bool Linux, Windows;

public static string SourcePath =

@"C:\Users\Stefan\Documents\vagrant\homestead\code\code\",

TargetPath = @"C:\D\";

As mentioned above, we simply ignore the Start.Windows it this point, an revisit them in Chapter 13.

public static void Main(string[] args){

if (Start.Windows) {

// ...

}

We set the current culture for decimal numbers to be properly interpreted.

System.Threading.Thread.CurrentThread.CurrentCulture =

CultureInfo.InvariantCulture;

If there are no command line arguments, we report an error.

if (args.Length == 0) {

Assert.Error("usage: compiler <filename>");

}

The rebuild flag indicates that each source file shall be compiled, regardless of whether the generated assembly file is fresh; that is, was last modified after the source file or any of its included files was last modified. The print flag indicates that information about the compiling process shall be printed to stdout.

List<string> argList = new List<string>(args);

bool rebuild = argList.Remove("-rebuild"),

print = argList.Remove("-print");

Then we start by iterating through the files of the command line arguments. The IsGeneratedFileFresh call decides if the assembly file is fresh; that is, if neither its source file nor any of the included files has been modified after generated assembly file. If the rebuild flag is set or if the source file is not fresh, it becomes compiled. The rebuild flag has just that effect, that the file shall be recompiled regardless of whether it is necessary or not.

try {

if (Start.Linux) {

foreach (string arg in argList) {

FileInfo file = new FileInfo(SourcePath + arg);

if (rebuild || !IsGeneratedFileFresh(file, ".asm")) {

if (print) {

Console.Out.WriteLine("Compiling \"" +

file.FullName + ".c\".");

}

CompileSourceFile(file);

}

}

Finally, we call the GenerateMakeFile method to create the makefile.

GenerateMakeFile(argList);

}

if (Start.Windows) {

// ...

}

}

In case of a runtime error (most likely: file error), we catch the exception and report an error.

catch (Exception exception) {

Console.Out.WriteLine(exception.StackTrace);

Assert.Error(exception.Message, Message.Parse\_error);

}

}

### Generating the Assembly File

The CompileSourceFile method compiles the source file into an assembly file. We begin by preprocessing the source code by creating an object of the Preprocessor class. The preprocessing has two outputs: the preprocessed code and the include set. We use the include set when calling GenerateDependencyFile to establish the dependency file of the source file; that is, the name of the source file itself and the files included (directly or indirectly) by the source file.

public static void CompileSourceFile(FileInfo file) {

FileInfo sourceFile = new FileInfo(file.FullName + ".c");

Preprocessor preprocessor = new Preprocessor(sourceFile);

GenerateDependencyFile(file, preprocessor.IncludeSet);

We convert the preprocessed source code to a byte array, that we use as input in a memory stream, that finally become input to the scanner.

byte[] byteArray =

Encoding.ASCII.GetBytes(preprocessor.PreprocessedCode);

MemoryStream memoryStream = new MemoryStream(byteArray);

CCompiler\_Main.Scanner scanner =

new CCompiler\_Main.Scanner(memoryStream);

We set the root symbol table to represent the global space of the source code and parse the source code. The Path and Line fields of the scanner is initialized in order to report the correct file and number in cases of error messages. If the parser encounters an error, it returns false and we report a syntax error.

try {

SymbolTable.CurrentTable = new SymbolTable(null, Scope.Global);

CCompiler\_Main.Parser parser = new CCompiler\_Main.Parser(scanner);

Assert.Error(parser.Parse(), Message.Syntax\_error);

}

catch (IOException ioException) {

Assert.Error(false, ioException.StackTrace, Message.Syntax\_error);

}

The result of the parsing is the SymbolTable.GlobalStaticSet set, holding a set of static symbols, such as functions and static variables. The resulting assembly file shall is made up by three parts:

Extern. The names of objects defined in other files and accessible in this file.

Global. The names of objects defined in this file and accessible in other files.

Code. The actual assembly code of the symbols of the global static set.

We start by defining the total extern set; that is, the union of the extern sets of the symbol in the global static set of the file.

if (Start.Linux) {

ISet<string> totalExternSet = new HashSet<string>();

foreach (StaticSymbol staticSymbol in SymbolTable.GlobalStaticSet) {

StaticSymbolLinux staticSymbolLinux =

(StaticSymbolLinux) staticSymbol;

totalExternSet.UnionWith(staticSymbolLinux.ExternSet);

}

We remove the names of the symbols of the global static set, since objects defined in this file are accessible in this file and shall not be stated as extern.

foreach (StaticSymbol staticSymbol in SymbolTable.GlobalStaticSet) {

totalExternSet.Remove(staticSymbol.UniqueName);

}

Now we are ready to write the actual assembly file, we start by creating a file and connect a stream to in. Is there is already a file with the same name, it will be overwritten.

FileInfo assemblyFile = new FileInfo(file.FullName + ".asm");

StreamWriter streamWriter = new StreamWriter(assemblyFile.FullName);

We iterate through the symbols and add the names of all symbol, with the exceptions of the symbols without external linkage (which have a separator identifier in their names) and symbols that represents constant values (which have a numeric identifier in their names). Neither symbols without external linkage nor value symbol shall be visible for other files.

foreach (StaticSymbol staticSymbol in SymbolTable.GlobalStaticSet) {

if (!staticSymbol.UniqueName.Contains(Symbol.SeparatorId) &&

!staticSymbol.UniqueName.Contains(Symbol.NumberId)) {

streamWriter.WriteLine("\tglobal " + staticSymbol.UniqueName);

}

}

streamWriter.WriteLine();

Then we add the names of all extern references.

foreach (string externName in totalExternSet) {

streamWriter.WriteLine("\textern " + externName);

}

If this file holds the initialization code, we state that the \_start marker and CallStackStart shall marker shall be global; that is, present in this file and accessible in other files. The \_start marker marks the start point of the execution of the code, while CallStackStart marks the beginning of the call stack.

if (SymbolTable.InitSymbol != null) {

streamWriter.WriteLine("\tglobal \_start");

streamWriter.WriteLine("\tglobal " + Linker.CallStackStart);

}

If this file does not hold the initialization code, we state that the \_start marker is extern; that is, present in another file and accessible in this file.

else {

streamWriter.WriteLine("\textern " + Linker.CallStackStart);

}

streamWriter.WriteLine();

Then we iterate through the global static set and add the text lists (that holds the assembly code) of the symbols.

foreach (StaticSymbol staticSymbol in SymbolTable.GlobalStaticSet) {

StaticSymbolLinux staticSymbolLinux =

(StaticSymbolLinux) staticSymbol;

streamWriter.WriteLine();

foreach (string line in staticSymbolLinux.TextList) {

streamWriter.WriteLine(line);

}

}

If this file holds the initialization code, we add the call stack for the activation records at the end of the file. We allocate one megabyte (1 048 576 bytes) for the call stack.

if (SymbolTable.InitSymbol != null) {

streamWriter.WriteLine();

streamWriter.WriteLine("section .data");

streamWriter.WriteLine(Linker.CallStackStart +

":\ttimes 1048576 db 0");

}

streamWriter.Close();

}

if (Start.Windows) {

// ...

}

}

### Generate the Make File

The GenerateMakeFile method generate a makefile for assembling and linking the target files into an executable file.

private static void GenerateMakeFile(List<string> argList) {

StreamWriter makeStream = new StreamWriter(SourcePath + "makefile");

We call the resulting executable file “main”, and we state that it is dependent of the object files; that is, the file names with the “.o” suffix.

makeStream.Write("main:");

foreach (string arg in argList) {

makeStream.Write(" " + arg.ToLower() + ".o");

}

makeStream.WriteLine();

We link the object files together with the GNU linker (ld) system command, which perform the linking of the object files.

makeStream.Write("\tld -o main");

foreach (string arg in argList) {

makeStream.Write(" " + arg.ToLower() + ".o");

}

makeStream.WriteLine();

makeStream.WriteLine();

We assembly each file with the Netwide Assembler (nasm), to assembly the assembly file into an object file.

foreach (string arg in argList) {

makeStream.WriteLine(arg.ToLower() + ".o: " + arg.ToLower() + ".asm");

makeStream.WriteLine("\tnasm -f elf64 -o " + arg.ToLower() + ".o "

+ arg.ToLower() + ".asm");

makeStream.WriteLine();

}

Finally, we generate the clear list by listing all the object files and the final main file.

makeStream.WriteLine("clear:");

foreach (string arg in argList) {

makeStream.WriteLine("\trm " + arg.ToLower() + ".o");

}

makeStream.WriteLine("\trm main");

makeStream.Close();

}

Below is an example of how the makefile may look:

main: main.o malloc.o ctype.o errno.o locale.o math.o setjmp.o signal.o file.o

temp.o scanf.o printf.o stdlib.o time.o string.o

ld -o main main.o malloc.o ctype.o errno.o locale.o math.o setjmp.o

signal.o file.o temp.o scanf.o printf.o stdlib.o time.o string.o

main.o: main.asm

nasm -f elf64 -o main.o main.asm

malloc.o: malloc.asm

nasm -f elf64 -o malloc.o malloc.asm

ctype.o: ctype.asm

nasm -f elf64 -o ctype.o ctype.asm

errno.o: errno.asm

nasm -f elf64 -o errno.o errno.asm

locale.o: locale.asm

nasm -f elf64 -o locale.o locale.asm

math.o: math.asm

nasm -f elf64 -o math.o math.asm

setjmp.o: setjmp.asm

nasm -f elf64 -o setjmp.o setjmp.asm

signal.o: signal.asm

nasm -f elf64 -o signal.o signal.asm

file.o: file.asm

nasm -f elf64 -o file.o file.asm

temp.o: temp.asm

nasm -f elf64 -o temp.o temp.asm

scanf.o: scanf.asm

nasm -f elf64 -o scanf.o scanf.asm

printf.o: printf.asm

nasm -f elf64 -o printf.o printf.asm

stdlib.o: stdlib.asm

nasm -f elf64 -o stdlib.o stdlib.asm

time.o: time.asm

nasm -f elf64 -o time.o time.asm

string.o: string.asm

nasm -f elf64 -o string.o string.asm

clear:

rm main.o

rm malloc.o

rm ctype.o

rm errno.o

rm locale.o

rm math.o

rm setjmp.o

rm signal.o

rm file.o

rm temp.o

rm scanf.o

rm printf.o

rm stdlib.o

rm time.o

rm string.o

rm main

### Is the Object File Fresh?

The GenerateDependencyFile method defines for each source code file its dependency set; that is, the name of the source file and the names of the files (directly or indirectly) included with the #include preprocessor directive. We use the set of include files information extracted by the preprocessor.

private static void GenerateDependencyFile(FileInfo file,

ISet<FileInfo> includeSet) {

FileInfo dependencySetFile = new FileInfo(file.FullName + ".dependency");

StreamWriter dependencyWriter =

new StreamWriter(File.Open(SourcePath + dependencySetFile.Name,

FileMode.Create));

dependencyWriter.Write(file.Name + ".c");

foreach (FileInfo includeFile in includeSet) {

dependencyWriter.Write(" " + includeFile.Name);

}

dependencyWriter.Close();

}

The IsGeneratedFileFresh method decides whether a generated file is fresh; that is, if none of its dependency files has been modified since the generated file was modified.

public static bool IsGeneratedFileFresh(FileInfo file, string suffix) {

FileInfo generatedFile = new FileInfo(file.FullName + suffix),

dependencySetFile = new FileInfo(file.FullName + ".dependency");

If the generated file does not exist, it is obviously not fresh, so we return false. If the dependency set file does not exist, we have to assume that it is not fresh due to dependencies, and we return false.

if (!generatedFile.Exists || !dependencySetFile.Exists) {

return false;

}

If none of the cases above apply, we continue to look into the dependency files; that is, the files that the generated file is (directly or indirectly) dependent of.

if (dependencySetFile.Exists) {

try {

StreamReader dependencySetReader =

new StreamReader(File.OpenRead(dependencySetFile.FullName));

string dependencySetText = dependencySetReader.ReadToEnd();

dependencySetReader.Close();

We split the file text into dependency file names and check whether the files have been modified after generated file.

if (dependencySetText.Length > 0) {

string[] dependencyNameArray = dependencySetText.Split(' ');

foreach (string dependencyName in dependencyNameArray) {

FileInfo dependencyFile =

new FileInfo(SourcePath + dependencyName);

If the file has been modified after the generated file, the generated file is not fresh, and we return false.

if (dependencyFile.LastWriteTime > generatedFile.LastWriteTime){

return false;

}

}

}

}

catch (IOException ioException) {

Console.Out.WriteLine(ioException.StackTrace);

return false;

}

}

When we have iterated through the without finding any dependencies, the generated file is fresh and we return true.

return true;

}

}

}

In order for the parsers of this book to work properly, we add three parser classes. We need them in order to use the parsers.

namespace CCompiler\_Main {

public partial class Parser :

QUT.Gppg.ShiftReduceParser<ValueType, QUT.Gppg.LexLocation> {

public Parser(Scanner scanner)

:base(scanner) {

// Empty.

}

}

}

The expression parser is used by the preprocessor, to parse expressions in the #if and #ifelse directives. In the expression grammar there is the possibility to test whether a macro has been defined. Therefore, we let the constructor accepts the macro map as a parameter.

namespace CCompiler\_Exp {

public partial class Parser :

QUT.Gppg.ShiftReduceParser<ValueType, QUT.Gppg.LexLocation> {

public static IDictionary<string,CCompiler.Macro> m\_macroMap;

public Parser(Scanner scanner,

IDictionary<string,CCompiler.Macro> macroMap)

:base(scanner) {

m\_macroMap = macroMap;

}

}

}

namespace CCompiler\_Pre {

public partial class Parser :

QUT.Gppg.ShiftReduceParser<ValueType, QUT.Gppg.LexLocation> {

public Parser(Scanner scanner)

:base(scanner) {

// Empty.

}

}

}

# Scanning

In this project we use Garden Point Lex, which is a scanner-generation tool based on the classic Lex tool. See Appendix E for a crash course.

## Slash Sequences

The first problem for the scanner to solve is to transform the escape sequences of characters and string into their proper characters. For instance, the character sequence ’\n’ shall be transformed into the character newline, with the ASCII value of 10. See Appendix G for the ASCII table.

Slash.cs

using System.Text;

using System.Collections.Generic;

namespace CCompiler {

class Slash {

The m\_escapeMap map holds the escape sequences of ANSI C

private static IDictionary<char,char> m\_slashMap =

new Dictionary<char,char>() {

// Key, ASCII value

{'0', '\0'}, // Terminator, 0

{'a', '\a'}, // Alert (Beep, Bell), 7

{'b', '\b'}, // Backspace, 8

{'f', '\f'}, // Form Feed (Page Break), 12

{'n', '\n'}, // New Line (Line Feed), 10

{'r', '\r'}, // Carrige Return, 13

{'t', '\t'}, // Horizontal Tabulator, 9

{'v', '\v'}, // Vertical Tabulator, 11

{'\'', '\''}, // Single Quotation Mark, 39

{'\"', '\"'}, // Double Quotation Mark, 34

{'?', '?'}, // Question Mark, 63

{'\\', '\\'} // Backslash, 92

};

The SlashToChar method takes a string with slash sequences and returns the same string with the slash sequences replaced by the proper characters.

public static string SlashToChar(string text) {

StringBuilder buffer = new StringBuilder(text);

We add three zero-characters to the string to make sure that the char1, char2, and char3 values are valid (that we do not index outside the string).

buffer.Append("\0\0\0");

for (int index = 0; buffer[index] != '\0'; ++index) {

if (buffer[index] == '\\') {

char char1 = buffer[index + 1],

char2 = buffer[index + 2],

char3 = buffer[index + 3];

If the slash is followed by a character stored in the slash map, we replace the slash and the character with the character corresponding to the slash sequence by using the m\_escapeMap map above.

if (m\_escapeMap.ContainsKey(char1)) {

buffer.Remove(index, 2);

buffer.Insert(index, m\_escapeMap[char1]);

}

If the three characters following the slash are octal digits, we calculate the ASCII value for the character.

else if (IsOctal(char1) && IsOctal(char2) && IsOctal(char3)) {

int octValue = 64 \* CharToOctal(char1) +

8 \* CharToOctal(char2) +

CharToOctal(char3);

An ASCII value is not allowed to exceed 255; if it does, we report an error.

Assert.Error(octValue <= 255, Message.Invalid\_octal\_sequence);

We remove four characters; that is, the slash and its following three octal digits, and we insert the resulting character in its place.

buffer.Remove(index, 4);

buffer.Insert(index, (char) octValue);

}

If the slash is followed by two octal digits only, we use them to calculate the ASCII value. In this case, we do not need to check whether the ASCII exceeds 255, since the highest possible value with two octal digits is 63 (778 = 8 \* 7 + 7 = 6310).

else if (IsOctal(char1) && IsOctal(char2)) {

int octValue = 8 \* CharToOctal(char1) +

CharToOctal(char2);

buffer.Remove(index, 3);

buffer.Insert(index, (char)octValue);

}

If the slash is followed by one octal digit only, we use it to calculate the ASCII value.

else if (IsOctal(char1)) {

int octValue = CharToOctal(char1);

buffer.Remove(index, 2);

buffer.Insert(index, (char) octValue);

}

If the slash is followed by a lowercase ‘x’ or an uppercase ‘X’ and two hexadecimal digits, we use them to calculate the ASCII value. Also in this case we need not check whether the ASCII value exceeds 255, since the highest possible value with two hexadecimal digits is 255 (FF16 = 15 \* 16 + 15 = 25510).

else if (char.ToLower(char1) == 'x') {

if (IsHex(char1) && IsHex(char2)) {

int hexValue = 16 \* CharToHex(char1) + CharToHex(char2);

buffer.Remove(index, 3);

buffer.Insert(index, (char) hexValue);

}

If the slash is followed by a lowercase ‘x’ or an uppercase ‘X’ and hexadecimal digit only, we use it to calculate the ASCII value.

else if (IsHex(char1)) {

int hexValue = CharToHex(char1);

buffer.Remove(index, 2);

buffer.Insert(index, (char) hexValue);

}

If the slash is followed by lowercase ‘x’ or an uppercase ‘X’ without at least one hexadecimal digit, we report an error.

else {

Assert.Error(char1.ToString(),

Message.Invalid\_hexadecimal\_code);

}

}

Finally, if the slash is not followed by a character in the m\_escapeMap map, an octal digit, a lowercase ‘x’ or an uppercase ‘X’, we report an error.

else {

Assert.Error(buffer[index + 1].ToString(),

Message.Invalid\_slash\_sequence);

}

}

}

When we have traversed through the string, we remove the three zero-character we added at the beginning and return the string.

buffer.Remove(buffer.Length - 3, 3);

return buffer.ToString();

}

The IsOctal and IsHex methods return true if the given character is and octal or hexadecimal digit, respectively.

private static bool IsOctal(char c) {

return "01234567".Contains(c.ToString());

}

private static bool IsHex(char c) {

return "0123456789abcdef".Contains(c.ToString().ToLower());

}

The CharToOctal and CharToHex methods return the numerical value corresponding to the given character.

private static int CharToOctal(char c) {

return "01234567".IndexOf(c);

}

private static int CharToHex(char c) {

return "0123456789abcdef".IndexOf(c.ToString().ToLower());

}

}

}

## The typedef-name Problem

Let us take look at the two source code lines below. Intuitively, the first line looks like an expression statement where x and y are variables of integral or floating types. Admittedly, the expression lacks side effects and is therefore meaningless, but it is still a valid expression statement. On the other hand, the second line looks like a pointer declaration where T is a type defined by and earlier typedef definition and p is a pointer to that type.

x \* y;

T\* p;

However, syntactically it is the same thing: a name, an asterisk, another name, and a semicolon. The parser cannot distinguish between the two cases. The solution to that problem is to give the scanner access to the symbol table, so it can look up the whether the name is an earlier typedef definition. In that way, the parser can distinguish between the two cases.

## The Scanner

The scanner identifies keywords, operators, strings, characters, and numerical values.

Scanner.gplex

%namespace CCompiler\_Main

%using CCompiler;

%using System.Numerics;

%{

public static FileInfo Path = null;

public static int Line = 1;

We have now reached the part of the scanner where we define the scanner rules. We need rules for names as well as octal, decimal, and hexadecimal, floating, string and character values. However, the regular expressions for values are a bit more complicated. There are several kinds of values:

Octal. An octal value starts with a zero that is followed by zero or more digits between zero and eight, inclusive.

OCTAL\_VALUE [\+\-]?0[0-7]\*

Decimal. A decimal value is made up by one or more digits.

DECIMAL\_VALUE [\+\-]?[1-9][0-9]\*

Hexadecimal. A hexadecimal value starts with zero, followed by a lowercase ‘x’ or an uppercase ‘X’ and one or more hexadecimal digits.

HEXADECIMAL\_VALUE [\+\-]?0[xX][0-9a-fA-F]+

Floating point. A floating value starts with one or more digits followed by a dot and zero or more digits or starts with a dot followed by one or more digits. It is possible to add an exponent part that starts with a lowercase ‘e’ or an uppercase ‘E’, followed by a potential plus or minus sign and one or more digits. Finally, a floating value can also be made up by one or more digits without a dot, followed by a lowercase ‘e’ or an uppercase ‘E’, a possible plus or minus sign, and one or more digits.

DECIMAL\_PART [\+\-]?([0-9]+|[0-9]+\.[0-9]\*|\.[0-9]+)

EXPONENT\_PART (([eE][\+\-]?[0-9]+)?|[0-9]+[eE][\+\-]?[0-9]+)([fF]|[lL])?

FLOATING\_VALUE {DECIMAL\_PART}{EXPONENT\_PART}

The decimal values can be appended by the lowercase ‘u’ or an uppercase ‘U’ to indicate an unsigned value and the lowercase ‘s’ or ‘l’ or an uppercase ‘S’ or ‘L’ to indicate a short or long value. In the same way, a floating value can be appended by a lowercase ‘f’ or ‘l’ or an uppercase ‘F’ or ‘L’ to indicate a float or long double.

POSTFIX ([uU]?[sSlL]?)|([sSlL]?[uU]?)

INTEGRAL\_VALUE ({DECIMAL\_VALUE}|{OCTAL\_VALUE}|{HEXADECIMAL\_VALUE}){POSTFIX}

A character is either two single quotation enclosing another single quotation preceded by a backslash (‘\’’) or two single quotation mark enclosing everything except single quotation marks.

CHAR\_VALUE (\'\\\'\')|(\'[^\']\*\')

A character starts with single quotation mark followed by one or more occurrences of anything except a single quotation mark and is terminated by another single quotation mark. However, there may be a single quotation mark if it is preceded by a backslash. Since the quotation mark and backslash are character with special meaning, we need to precede them with backslashes. Note that we do not need the backslash when the quotation mark is enclosed by brackets.

CHAR\_VALUE \'(\\\'|[^'])\*\'

In the same way, a string starts with a double quotation mark followed by anything except double quotation mark and is terminated by another double quotation mark. However, the string may hold double quotations marks if they are proceeded by backslashes.

STRING\_VALUE \"(\\\"|[^"])\*\"

Names are used to identify variables, constants, struct and unions, enumerations, functions, and macros. They start with a letter or an underscore (’\_’), that is optionally followed by letters, digits, or underscores.

NAME [a-zA-Z\_][a-zA-Z0-9\_]\*

Register names are used internally only, when performing system calls. They start with the text “register\_” followed by the name of the register.

REGISTER\_NAME "register\_"[a-z]+

The path line is used to keeping track of the current line number. It is used by the preprocessor macro \_\_LINE\_\_, and when reporting errors. The path line starts with a dollar sign followed by anything except newline (the dot represents every character except newline) and ends with another dollar sign.

PATH\_LINE \$.\*\$

A white space is a space or any character that can substitute as a space; that is horizontal tabulator, return, newline, and form feed.

WHITE\_SPACE [ \t\r\n\f]

The next section holds the actions of keywords, operators, and the rules defined above.

%%

"auto" { return ((int) Tokens.AUTO); }

"break" { return ((int) Tokens.BREAK); }

"case" { return ((int) Tokens.CASE); }

"carry\_flag" { return ((int) Tokens.CARRY\_FLAG); }

"char" { return ((int) Tokens.CHAR); }

"const" { return ((int) Tokens.CONSTANT); }

"continue" { return ((int) Tokens.CONTINUE); }

"default" { return ((int) Tokens.DEFAULT); }

"do" { return ((int) Tokens.DO); }

"double" { return ((int) Tokens.DOUBLE); }

"else" { return ((int) Tokens.ELSE); }

"enum" { return ((int) Tokens.ENUM); }

"extern" { return ((int) Tokens.EXTERN); }

"float" { return ((int) Tokens.FLOAT); }

"for" { return ((int) Tokens.FOR); }

"goto" { return ((int) Tokens.GOTO); }

"int" { return ((int) Tokens.INT); }

"interrupt" { return ((int) Tokens.INTERUPT); }

"if" { return ((int) Tokens.IF); }

"jump\_register" { return ((int) Tokens.JUMP\_REGISTER); }

"long" { return ((int) Tokens.LONG); }

"register" { return ((int) Tokens.REGISTER); }

"return" { return ((int) Tokens.RETURN); }

"short" { return ((int) Tokens.SHORT); }

"signed" { return ((int) Tokens.SIGNED); }

"sizeof" { return ((int) Tokens.SIZEOF); }

"static" { return ((int) Tokens.STATIC); }

"struct" { return ((int) Tokens.STRUCT); }

"switch" { return ((int) Tokens.SWITCH); }

"syscall" { return ((int) Tokens.SYSCALL); }

"typedef" { return ((int) Tokens.TYPEDEF); }

"union" { return ((int) Tokens.UNION); }

"unsigned" { return ((int) Tokens.UNSIGNED); }

"while" { return ((int) Tokens.WHILE); }

"void" { return ((int) Tokens.VOID); }

"volatile" { return ((int) Tokens.VOLATILE); }

";" { return ((int) Tokens.SEMICOLON); }

":" { return ((int) Tokens.COLON); }

"," { return ((int) Tokens.COMMA); }

"." { return ((int) Tokens.DOT); }

"->" { return ((int) Tokens.ARROW); }

"..." { return ((int) Tokens.ELLIPSE); }

"(" { return ((int) Tokens.LEFT\_PAREN); }

")" { return ((int) Tokens.RIGHT\_PAREN); }

"{" { return ((int) Tokens.LEFT\_BLOCK); }

"}" { return ((int) Tokens.RIGHT\_BLOCK); }

"[" { return ((int) Tokens.LEFT\_SQUARE); }

"]" { return ((int) Tokens.RIGHT\_SQUARE); }

"\*" { return ((int) Tokens.ASTERRISK); }

"?" { return ((int) Tokens.QUESTION\_MARK); }

"||" { return ((int) Tokens.LOGICAL\_OR); }

"&&" { return ((int) Tokens.LOGICAL\_AND); }

"!" { return ((int) Tokens.LOGICAL\_NOT); }

"&" { return ((int) Tokens.AMPERSAND); }

"^" { return ((int) Tokens.BITWISE\_XOR); }

"|" { return ((int) Tokens.BITWISE\_OR); }

"~" { return ((int) Tokens.BITWISE\_NOT); }

"==" { return ((int) Tokens.EQUAL); }

"!=" { return ((int) Tokens.NOT\_EQUAL); }

"<" { return ((int) Tokens.LESS\_THAN); }

"<=" { return ((int) Tokens.LESS\_THAN\_EQUAL); }

">" { return ((int) Tokens.GREATER\_THAN); }

">=" { return ((int) Tokens.GREATER\_THAN\_EQUAL); }

"<<" { return ((int) Tokens.LEFT\_SHIFT); }

">>" { return ((int) Tokens.RIGHT\_SHIFT); }

"+" { return ((int) Tokens.PLUS); }

"-" { return ((int) Tokens.MINUS); }

"/" { return ((int) Tokens.DIVIDE); }

"%" { return ((int) Tokens.MODULO); }

"++" { return ((int) Tokens.INCREMENT); }

"--" { return ((int) Tokens.DECREMENT); }

"=" { return ((int) Tokens.ASSIGN); }

"+=" { return ((int) Tokens.ADD\_ASSIGN); }

"-=" { return ((int) Tokens.SUBTRACT\_ASSIGN); }

"\*=" { return ((int) Tokens.MULTIPLY\_ASSIGN); }

"/=" { return ((int) Tokens.DIVIDE\_ASSIGN); }

"%=" { return ((int) Tokens.MODULO\_ASSIGN); }

"<<=" { return ((int) Tokens.LEFT\_SHIFT\_ASSIGN); }

">>=" { return ((int) Tokens.RIGHT\_SHIFT\_ASSIGN); }

"&=" { return ((int) Tokens.AND\_ASSIGN); }

"^=" { return ((int) Tokens.XOR\_ASSIGN); }

"|=" { return ((int) Tokens.OR\_ASSIGN); }

When we encounter a register name, we look up the register, and report an error there is no register with the name.

{REGISTER\_NAME} {

{ Register register;

string text = yytext.Substring(9);

if (Enum.TryParse<Register>(text, out register)) {

yylval.register = register;

return ((int) Tokens.REGISTER\_NAME);

}

Assert.Error(text, Message.Unknown\_register);

}

}

When we encounter a name, we need to check whether it represent a typedef name or a regular name; that is, the name of a variable, constant, struct or union, enumeration, or a function. We look up the name in the current symbol table in accordance with the typedef-name problem above and return the typedef-name token in case of a typedef name. In case of a regular name, we return the name token.

{NAME} {

{ string name = yytext;

if (Start.Linux && (name.Equals("abs") ||

name.Equals("\_start") || name.Equals("section") ||

name.Equals("extern") || name.Equals("global"))) {

name = Symbol.FileMarker + name;

}

Symbol symbol = SymbolTable.CurrentTable.LookupSymbol(name);

if ((symbol != null) && symbol.IsTypedef()) {

yylval.type = symbol.Type;

return ((int) Tokens.TYPEDEF\_NAME);

}

else {

yylval.name = name;

return ((int) Tokens.NAME);

}

}

}

When it comes to integral values, we need to decide the sign, base, and type of the value.

{INTEGRAL\_VALUE} {

{ string text = yytext.Trim().ToLower();

The ToUInt64 method below does not accept plus or minus signs. Therefore, we need to remove the sign and assign true to the minus variable in case of a negative value.

bool minus = false;

if (text.StartsWith("+")) {

text = text.Substring(1);

}

else if (text.StartsWith("-")) {

minus = true;

text = text.Substring(1);

}

Then we decide the base of the value. If the text start with “0x” or “0X”, the value is a hexadecimal with the base 16. If the text start with “0”, the value is an octal value with the base 8. Otherwise, it is a decimal value with the base 10.

int fromBase;

if (text.StartsWith("0x")) {

fromBase = 16;

text = text.Substring(2);

}

else if (text.StartsWith("0")) {

fromBase = 8;

}

else {

fromBase = 10;

}

Then we decide the type of the value. If the text ends with any combination of lowercase ‘s’ or ‘u’ or an uppercase ‘S’ or ‘U’, the type is unsigned small integer. If the text ends with a lowercase ‘s’ or an uppercase ‘S’, the type is small integer. If the text ends with lowercase ‘l’ or an uppercase ‘l’, the type is large integer. If the text ends with lowercase ‘u’ or an uppercase ‘U’, the type is unsigned integer. Otherwise, the type is signed integer.

CCompiler.Type type;

if (text.EndsWith("us") || text.EndsWith("su")) {

type = CCompiler.Type.UnsignedShortIntegerType;

text = text.Substring(0, text.Length - 2);

}

If the text ends with any combination of lowercase ‘l’ or ‘u’ or an uppercase ‘L’ or ‘U’, the type is large unsigned integer.

else if (text.EndsWith("ul") || text.EndsWith("lu")) {

type = CCompiler.Type.UnsignedLongIntegerType;

text = text.Substring(0, text.Length - 2);

}

If the text ends with a single lowercase ‘s’ or an uppercase ‘S’, the type is signed short integer.

else if (text.EndsWith("s")) {

type = CCompiler.Type.SignedShortIntegerType;

text = text.Substring(0, text.Length - 1);

}

If the text ends with a single lowercase ‘l’ or an uppercase ‘L’, the type is signed long integer.

else if (text.EndsWith("l")) {

type = CCompiler.Type.SignedLongIntegerType;

text = text.Substring(0, text.Length - 1);

}

If the text ends with a single lowercase ‘u’ or an uppercase ‘U’, the type is unsigned integer.

else if (text.EndsWith("u")) {

type = CCompiler.Type.UnsignedIntegerType;

text = text.Substring(0, text.Length - 1);

}

Otherwise, the type is signed integer.

else {

type = CCompiler.Type.SignedIntegerType;

}

Finally, we decide the value by calling the ToUInt64 method in the Convert standard class.

try {

ulong unsignedValue = Convert.ToUInt64(text, fromBase);

We convert the value to an object of the standard BigInteger class, and if minus is true, we make the value negative.

BigInteger bigValue = new BigInteger(unsignedValue);

if (minus) {

bigValue = -bigValue;

}

We then create a symbol of the value that we return. We also store a static symbol in the global static set.

yylval.symbol = new Symbol(type, bigValue);

yylval.symbol.StaticSymbol =

ConstantExpression.Value(yylval.symbol.UniqueName, type, bigValue);

SymbolTable.StaticSet.Add(yylval.symbol.StaticSymbol);

}

catch (OverflowException) {

Assert.Error("X " + type + ": " + text, Message.Value\_overflow);

}

Finally, we return the value token.

return ((int) Tokens.VALUE);

}

}

When it comes to floating values, we do not need to look for its plus or minus sign. However, we do notice if the text ends with lowercase ‘f’ and ‘l’ or an uppercase ‘F’ or ‘L’, in which case the type becomes float or long double, respectively. If the text does not end with ‘f’, ‘F’, ‘l’, or ‘L’, the type is a double.

{FLOATING\_VALUE} {

{ string text = yytext.ToLower();

CCompiler.Type type = CCompiler.Type.DoubleType;

if (text.EndsWith("f")) {

type = CCompiler.Type.FloatType;

text = text.Substring(0, text.Length - 1);

}

else if (text.EndsWith("l")) {

type = CCompiler.Type.LongDoubleType;

text = text.Substring(0, text.Length - 1);

}

We call the Parse method of the decimal standard class to obtain the value. Then we create and return a value symbol. We also add a static value to the global static set.

try {

decimal value = decimal.Parse(text, NumberStyles.Float);

yylval.symbol = new Symbol(type, value);

yylval.symbol.StaticSymbol =

ConstantExpression.Value(yylval.symbol.UniqueName, type, value);

SymbolTable.StaticSet.Add(yylval.symbol.StaticSymbol);

}

catch (OverflowException) {

Assert.Error("Y " + type + ": " + text, Message.Value\_overflow);

}

return ((int) Tokens.VALUE);

}

}

In the case of a character value, we call the SlashToChar method of the Slash class to clear the text from escape characters. We then check that it holds three characters (including the two single quotation marks), and return a symbol holding the value.

{CHAR\_VALUE} {

{ CCompiler.Type type = new CCompiler.Type(Sort.SignedChar);

string text = Slash.SlashToChar(yytext);

Assert.Error(text.Length == 3, yytext, Message.Invalid\_char\_sequence);

yylval.symbol = new Symbol(type, (BigInteger) ((int) text[1]));

return ((int) Tokens.VALUE);

}

}

In the case of a string value, we also call the SlashToChar method of the Slash class to clear the text from escape characters. We then return a symbol holding the value. In this case we do not need to check the length of the string.

{STRING\_VALUE} {

{ CCompiler.Type type = new CCompiler.Type(Sort.String);

string text = Slash.SlashToChar(yytext);

object value = text.Substring(1, text.Length - 2);

yylval.symbol = new Symbol(type, value);

yylval.symbol.StaticSymbol =

ConstantExpression.Value(yylval.symbol.UniqueName, type, value);

SymbolTable.StaticSet.Add(yylval.symbol.StaticSymbol);

return ((int) Tokens.VALUE);

}

}

Path lines are generated by the preprocessor for the \_\_FILE\_\_ and \_\_LINE\_\_ macros to hold the correct file name and line number. It is also used by the compiler when reporting error. Its text is made up by two dollar signs (‘$’) at the beginning and end, and the file name and line number separated by a comma (‘,’). The text also holds plus signs (‘+’) in place of spaces.

{PATH\_LINE} {

{ string text = yytext.Substring(1, yyleng - 2);

int index = text.IndexOf(',');

Path = new FileInfo(text.Substring(0, index).Replace("+", " "));

Line = int.Parse(text.Substring(index + 1));

}

}

In case of a white space, we only update the Line variable if the character is a newline. Note that we do not return anything in this case, which causes the scanner to proceed with the next character in the input stream.

{WHITE\_SPACE} {

if (yytext.Equals("\n")) {

++Line;

}

}

Finally, when the scanner finds a character that has not been handled by the rules above, a error is reported.

. { Assert.Error(yytext, Message.Unknown\_character); }

# Parsing

The parser can be considered to be the heart of the compiler. It requests token from the scanner, checks that the source code complies with the grammar, builds the symbol table, performs type checking, and generates the middle code.

## The MiddleOperator Enumeration

Before we start the parsing, we need to look into the middle code generated by the parser. The MiddleOperator enumeration holds the operator of the middle code.

MiddleOperator.cs

namespace CCompiler {

public enum MiddleOperator

{Initializer, InitializerZero, AssignInitSize, Assign,

The addition, subtraction, increment, and decrement operations perform the same task with signed and unsigned values, while the multiplication, division, and modulo operations need to take into consideration whether the values are signed or unsigned.

UnaryAdd, UnarySubtract, BinaryAdd, BinarySubtract,

Increment, Decrement,

SignedMultiply, SignedDivide, SignedModulo,

UnsignedMultiply, UnsignedDivide, UnsignedModulo,

LogicalOr, LogicalAnd,

BitwiseNot, BitwiseOr, BitwiseXOr, BitwiseAnd,

ShiftLeft, ShiftRight,

Compare, Equal, NotEqual, SignedLessThan,

SignedLessThanEqual, SignedGreaterThan,

SignedGreaterThanEqual, UnsignedLessThan,

UnsignedLessThanEqual, UnsignedGreaterThan,

UnsignedGreaterThanEqual,

Carry, NotCarry, Case, CaseEnd, Jump,

Index, Dot, Address, Dereference,

IntegralToIntegral, IntegralToFloating, FloatingToIntegral,

PushZero, PushOne, PushFloat, PopFloat, TopFloat, PopEmpty,

PreCall, Call, PostCall, DecreaseStack,

ParameterInitSize, Parameter,

GetReturnValue, SetReturnValue, Return, Exit,

AssignRegister, InspectRegister, StackTop,

JumpRegister, SysCall, Interrupt,

FunctionEnd, Empty};

}

MainParser.gppg

%namespace CCompiler\_Main

%partial

%using CCompiler;

%using System.Numerics;

%{

public static Stack<Specifier> SpecifierListStack = new Stack<Specifier>();

public static Stack<BigInteger> EnumValueStack = new Stack<BigInteger>();

public static Stack<Scope> ScopeStack = new Stack<Scope>();

%}

The parser is made up by tokens and rules, where tokens correspond to operators, keywords, and characters. The following tokens are used by the parser, and they are also returned by the scanner in the Chapter 2.

A variable or function has auto, register, static, extern, or typedef storage.

%token AUTO REGISTER STATIC EXTERN TYPEDEF

It can also be qualified as constant or volatile.

CONSTANT VOLATILE

An integral type can be signed or unsigned char, short int, int, or long int.

SIGNED UNSIGNED CHAR SHORT INT LONG

A floating type can be float, double, or long double.

FLOAT DOUBLE

Moreover, there are enumerations, structs, and unions.

ENUM STRUCT UNION

Finally, there is also the void type, that marks the absence of a type. It is used to mark the absence of a parameter list or return type of a function, and as pointer to void.

VOID

The asterisk (’\*’) is used both for multiplication and dereferencing pointers.

PLUS MINUS ASTERRISK DIVIDE MODULO INCREMENT DECREMENT

EQUAL NOT\_EQUAL LESS\_THAN LESS\_THAN\_EQUAL GREATER\_THAN

GREATER\_THAN\_EQUAL

The assignment is both simple and compound.

ASSIGN ADD\_ASSIGN SUBTRACT\_ASSIGN MULTIPLY\_ASSIGN

DIVIDE\_ASSIGN MODULO\_ASSIGN AND\_ASSIGN OR\_ASSIGN XOR\_ASSIGN

LEFT\_SHIFT\_ASSIGN RIGHT\_SHIFT\_ASSIGN

LOGICAL\_OR LOGICAL\_AND LOGICAL\_NOT

The ampersand (’&’) is used both bitwise and as well as address operator.

AMPERSAND BITWISE\_XOR BITWISE\_OR BITWISE\_NOT

The ellipse is made up by threee dots (’...’) and is used when defining functions with a variable number of parameters, such as printf or scanf.

QUESTION\_MARK COLON COMMA SEMICOLON ELLIPSE DOT ARROW SIZEOF

In C, we have the regular parentheses ’(’ and ’)’, block parentheses ’{’ and ’}’, and squares ’[’ and ’]’

LEFT\_PARENTHESIS RIGHT\_PARENTHESIS LEFT\_BLOCK RIGHT\_BLOCK

LEFT\_SQUARE RIGHT\_SQUARE

We need a set of keywords for the statements of C.

IF ELSE SWITCH CASE DEFAULT FOR WHILE DO BREAK CONTINUE RETURN GOTO

The interrupt, jump\_register, syscall, and carry flag tokens are used for internal system calls only.

INTERRUPT JUMP\_REGISTER SYSCALL CARRY\_FLAG

Note that we do not need any tokens for comments, since they have already been taken care of by the preprocessor.

The next part of the scanner is the union section, where we define the attributes of the rules below.

%union {

public string name;

The Register enumeration holds the registers of the target machine.

public Register register;

Each symbol has a type, which is described by the Type class.

public CCompiler.Type type;

public List<CCompiler.Type> type\_list;

public Sort sort;

public Symbol symbol;

public IDictionary<string,Symbol> symbol\_map;

public ISet<Pair<Symbol,bool>> symbol\_bool\_pair\_set;

public Pair<Symbol,bool> symbol\_bool\_pair;

public List<Symbol> symbol\_list;

public List<string> string\_list;

public Declarator declarator;

public List<Declarator> declarator\_list;

public MiddleOperator middle\_operator;

public Expression expression;

public List<Expression> expression\_list;

public Statement statement;

public Pair<List<Symbol>,Boolean> parameter\_pair;

public List<MiddleCode> middle\_code\_list;

public object obj;

public List<object> object\_list;

}

%token <name> NAME

%token <register> REGISTER\_NAME

%token <type> TYPEDEF\_NAME

%token <symbol> VALUE

%type <obj> declaration\_specifier declaration\_specifier\_list\_x

%type <object\_list> declaration\_specifier\_list

%type <name> optional\_name

%type <type> struct\_or\_union\_specifier

%type <sort> struct\_or\_union

%type <type> enum\_specifier

%type <symbol\_bool\_pair\_set> enum\_list

%type <symbol\_bool\_pair> enum

%type <middle\_code\_list> declarator\_list declaration

declaration\_list optional\_declaration\_list

initialization\_bitfield\_simple\_declarator

%type <declarator> optional\_simple\_declarator declarator direct\_declarator

%type <type\_list> optional\_pointer\_list pointer\_list

%type <type> pointer\_marker

%type <parameter\_pair> optional\_parameter\_ellipse\_list

parameter\_ellipse\_list

%type <symbol\_list> parameter\_list

%type <symbol> parameter\_declaration

%type <string\_list> optional\_name\_list identifier\_list

%type <object\_list> initializer\_list

%type <obj> initializer

%type <type> type\_name

%type <declarator> abstract\_declarator direct\_abstract\_declarator

%type <middle\_operator> assignment\_operator equality\_operator

relation\_operator add\_operator shift\_operator

multiply\_operator prefix\_add\_operator

increment\_operator

%type <expression> optional\_constant\_integral\_expression

constant\_integral\_expression optional\_expression

expression assignment\_expression

condition\_expression logical\_or\_expression

logical\_and\_expression BITWISE\_OR\_expression

bitwise\_xor\_expression bitwise\_and\_expression

equality\_expression relation\_expression

shift\_expression add\_expression

multiply\_expression type\_cast\_expression

prefix\_expression postfix\_expression

primary\_expression

%type <statement> optional\_statement\_list statement

closed\_statement opened\_statement

%type <expression\_list> optional\_argument\_expression\_list

argument\_expression\_list

## Declarations

Finally, we have reached the last (and largest) part of the parser, where the rules are defined. The source\_code\_file rule is the start of the grammar.

%start source\_code\_file

%%

A source code file is made up of a sequence of external declarations, which may be a function definition or a regular declaration.

source\_code\_file:

external\_declaration

| source\_code\_file external\_declaration;

external\_declaration:

function\_definition

| declaration;

### Function Definition

A function definition is made up by declarator, possible preceded by a declaration specifier list, followed by an optional declaration list, and a block with an optional statement list. For instance, in the function below unsigned long int is the declaration specifier list, square(int value) is the declarator, and return value \* value ; is the only statement in the statement list.

unsigned long int square(int value) { return value \* value; }

In C, there are two ways to define a function, the old way where the parameter list hold holds the names of the parameters and where there types are defined afterwards, and new way where the parameter list holds the names and types of the parameters. In the following examples f is defined by the old way and g is defined by the new way.

double f(a, b)

int a;

double b; {

return a + b;

}

double f(int a, double b) {

return a + b;

}

The first part of the function definition rule handles the case where there is specifier list of at least specifier present. There are three methods to handle the function definition: FunctionHeader that handles the header of the function, CheckFunctionDefinition that makes sure the function is defined in either the old or new way, and FunctionEnd that generates the assembly code of the function and saves it in the global static set.

MainParser.gppg

function\_definition:

declaration\_specifier\_list\_x declarator {

MiddleCodeGenerator.FunctionHeader

(SpecifierListStack.Pop(), $2);

}

optional\_declaration\_list {

MiddleCodeGenerator.CheckFunctionDefinition();

}

LEFT\_BLOCK optional\_declaration\_list optional\_statement\_list RIGHT\_BLOCK {

$8.CodeList.InsertRange(0, $7);

MiddleCodeGenerator.BackpatchGoto();

MiddleCodeGenerator.FunctionEnd($8);

}

The second rule handles the case where there is no specifier list. In that case, the function return type is assumed to be a signed integer.

| declarator {

MiddleCodeGenerator.FunctionHeader(null, $1);

}

optional\_declaration\_list {

MiddleCodeGenerator.CheckFunctionDefinition();

}

LEFT\_BLOCK optional\_declaration\_list optional\_statement\_list RIGHT\_BLOCK {

$7.CodeList.InsertRange(0, $6);

MiddleCodeGenerator.BackpatchGoto();

MiddleCodeGenerator.FunctionEnd($7);

};

optional\_declaration\_list:

/\* Empty \*/ { $$ = new List<MiddleCode>(); }

| declaration\_list { $$ = $1; };

### Specifier List

A declaration is made up by a declaration specifier list followed by an optional declarator list, and a semicolon. For instance, in the following declaration, static long int is the specifier list while \*p, and a[3] are declarators.

static long int \*p, a[3];

There is possible to define a declaration without a declarator list. This feature is useful when declare a named struct or union:

struct \_s {int i};

However, this feature makes it also possible to declare unnamed specifier lists. The following declaration are syntactically correct but without meaning, they will result in no generated code.

unsigned short int;

union {short s;}

However, the following declaration has meaning since the enumeration values can be used.

enum {ZERO = 0, ONE = 1, TEN = 10};

MainParser.gppg

declaration:

declaration\_specifier\_list SEMICOLON {

SpecifierListStack.Push(Specifier.SpecifierList($1));

$$ = new List<MiddleCode>();

}

| declaration\_specifier\_list\_x declarator\_list SEMICOLON {

SpecifierListStack.Pop();

$$ = $2;

};

optional\_declaration\_specifier\_list\_x:

/\* empty \*/

| declaration\_specifier\_list\_x;

Due to parser-specific limitations we add an extra rule for the specifier list and push the list at specifier list stack. The stack is inspected for each of the following declarator. It is necessary to decide the type completely in case of initialization of the declarator.

declaration\_specifier\_list\_x:

declaration\_specifier\_list {

SpecifierListStack.Push(Specifier.SpecifierList($1));

};

A declaration specifier list is made up by one or more declaration specifiers, which are stored in a list.

declaration\_specifier\_list:

declaration\_specifier {

$$ = new List<object>();

$$.Add($1);

}

| declaration\_specifier declaration\_specifier\_list {

$2.Add($1);

$$ = $2;

};

There is a set of different kind of specifiers. The following specifiers return a value of the Mask enumeration.

Storage specifiers: auto, register, static, extern, or typedef

Type qualifiers: constant or volatile

Type specifiers: void, char, short, int, long, float, double, signed, or unsigned

The following specifiers return a reference to the Type class. Note that a specifier can be a struct, union, or enumeration declaration, or a typedef name as well as a keyword.

Struct or union specifier

Enum specifier

Typedef name

declaration\_specifier:

AUTO { $$ = Mask.Auto; }

| REGISTER { $$ = Mask.Register; }

| STATIC { $$ = Mask.Static; }

| EXTERN { $$ = Mask.Extern; }

| TYPEDEF { $$ = Mask.Typedef; }

| CONSTANT { $$ = Mask.Constant; }

| VOLATILE { $$ = Mask.Volatile; }

| VOID { $$ = Mask.Void; }

| CHAR { $$ = Mask.Char; }

| SHORT { $$ = Mask.Short; }

| INT { $$ = Mask.Int; }

| LONG { $$ = Mask.Long; }

| FLOAT { $$ = Mask.Float; }

| DOUBLE { $$ = Mask.Double; }

| SIGNED { $$ = Mask.Signed; }

| UNSIGNED { $$ = Mask.Unsigned; }

| struct\_or\_union\_specifier { $$ = $1; }

| enum\_specifier { $$ = $1; }

| TYPEDEF\_NAME { $$ = $1; };

### Structs and Unions

A struct or union specifier holds an optional name and a declaration list within brackets. It can also refer to a declared struct or union by its name. When the declaration list is parsed, the struct or union is given a symbol table of its own and each member becomes added to symbol table.

MainParser.gppg

The call to StructOrUnionHeader adds the struct or union to the symbol table if the optional name is not null. If the optional name is not null, we add a tag to the current symbol table. We need to do this in order for recursive pointers to work. For instance:

struct Cell {

int value;

struct Cell\* next;

}

If we do not add the tag, there is a risk that the next field in the code above will point to another struct with the same name, defined in a surrounding block or in global space.

struct\_or\_union\_specifier:

struct\_or\_union optional\_name {

if ($2 != null) {

SymbolTable.CurrentTable.AddTag($2, new CCompiler.Type($1));

}

Before the parsing of the members we assign them a symbol table of their own.

SymbolTable.CurrentTable =

new SymbolTable(SymbolTable.CurrentTable, (Scope) $1);

}

LEFT\_BLOCK declaration\_list RIGHT\_BLOCK {

$$ = MiddleCodeGenerator.StructUnionSpecifier($2, $1);

After the parsing of the member list of the struct, we restore the original symbol table.

SymbolTable.CurrentTable = SymbolTable.CurrentTable.ParentTable;

}

In case of a struct of union without a declaration list, but with un obligatory name, we look up the name.

| struct\_or\_union NAME {

$$ = MiddleCodeGenerator.LookupStructUnionSpecifier($2, $1);

};

struct\_or\_union:

STRUCT { $$ = Sort.Struct; }

| UNION { $$ = Sort.Union; };

optional\_name:

/\* Empty \*/ { $$ = null; }

| NAME { $$ = $1; };

declaration\_list:

declaration {

$$ = $1;

}

| declaration\_list declaration {

$1.AddRange($2);

$$ = $1;

};

### Enumeration

An enumeration declaration may hold a list of enumeration items, which are assigned values. Each item is declared as a constant signed integer, with a value implicitly or explicitly assigned.

MainParser.gppg

enum\_specifier:

ENUM optional\_name {

EnumValueStack.Push(BigInteger.Zero);

}

LEFT\_BLOCK enum\_list RIGHT\_BLOCK {

EnumValueStack.Pop();

$$ = MiddleCodeGenerator.EnumSpecifier($2, $5);

}

An enumeration declaration may also hold a name without an enumeration item list. In that case, we look up the name and returns its type, which is a constant signed integer. If the name of the enumeration does not exist, we report an error.

| ENUM NAME {

$$ = MiddleCodeGenerator.LookupEnum($2);

};

The result of enum\_list is a set holding the symbols of the enumeration items and a Boolean value indicating whether the item was explicitly assigned. We need the Boolean value in case the enumeration symbol is given external storage, in which case assignment of explicit values are not allowed.

enum\_list:

enum {

ISet<Pair<Symbol,bool>> memberSet = new HashSet<Pair<Symbol,bool>>();

memberSet.Add($1);

$$ = memberSet;

}

| enum\_list COMMA enum {

ISet<Pair<Symbol,bool>> memberSet = $1;

memberSet.Add($3);

$$ = memberSet;

};

When it comes to enumerations, we have a potential problem. Each enumeration item is stored in the symbol table as a signed integer with a value. As stated above, the value may be explicitly assigned or implicitly given a value. If the storage of the enumeration is extern, it is not allowed to assign values to the enumeration items. However, it is in C allowed to state the declaration specifiers in arbitrary order. For instance, the following declaration is valid:

enum {a, b} extern;

The following declaration is invalid, since it has extern storage and a is assigned a value:

enum {a = 1, b} extern;

The problem is that when we parse enumeration declaration we might not yet know if its storage is extern. Therefore, we must also store, for each item, whether it has been explicitly assigned. The handling of the storage is managed by the Specifier class (see Chapter 4) after all declaration specifiers have been parsed. The result is that for each item a pair is returned, holding the symbol of the item and a Boolean value indicating whether the item has been explicitly assigned a value.

enum:

NAME {

Symbol symbol = MiddleCodeGenerator.EnumItem($1, null);

$$ = new Pair<Symbol,bool>(symbol, false);

}

| NAME ASSIGN constant\_integral\_expression {

Symbol symbol = MiddleCodeGenerator.EnumItem($1, $3.Symbol);

$$ = new Pair<Symbol,bool>(symbol, true);

};

### Declarators

As mentioned in the previous section, a declaration is made up by a declaration specifier list, an optional declarator list, and a semicolon. In the following declaration, static long int is the specifier list while \*p, and a[3] are declarators.

static long int \*p, a[3];

A declarator list is made up declarators separated by commas. There are three kinds of declarators:

* Simple, such as int i;
* With initialization, such as int i = 3;
* With bitfield, only allowed in structs and unions, such as int i : 3;. It also possible to omit the declarator; in the int : 3; declaration, it is only stated that three bits shall be unused.

Note that it is not possible to combine initializations with bitfield markers, since bitfields are only allowed as struct or union members that cannot be initialized.

MainParser.gppg

declarator\_list:

initialization\_bitfield\_declarator {

$$ = $1;

}

| declarator\_list COMMA initialization\_bitfield\_declarator {

$1.AddRange($3);

$$ = $1;

};

When calling the declarator methods in the MiddleCodeDeclarator class, we submit the specifier list at the top of the specifier list stack, in order to build the complete type. In some cases, the specifier list constitutes the type; for instance, in int i;, the integer is the complete type. In other cases, such as int \*p, f(int);, the pointer to integer and the function return integer are the complete types.

initialization\_bitfield\_declarator:

declarator {

MiddleCodeGenerator.Declarator(SpecifierListStack.Peek(), $1);

$$ = new List<MiddleCode>();

}

| declarator ASSIGN initializer {

$$ = MiddleCodeGenerator.InitializedDeclarator

(SpecifierListStack.Peek(), $1, $3);

}

| optional\_declarator COLON constant\_integral\_expression {

MiddleCodeGenerator.BitfieldDeclarator

(SpecifierListStack.Peek(), $1, $3.Symbol);

$$ = new List<MiddleCode>();

};

### Pointer Declarators

A pointer declarator is made up by an optional list of pointer markers and a direct declarator. A pointer marker is made up by an optional qualifier (constant or volatile) list and an asterisk (‘\*’). In other words, each asterisk can be qualified by the word constant and/or volatile, and it is possible to add more than one pointer marker to establish a pointer-to-pointer effect.

MainParser.gppg

declarator:

optional\_pointer\_list direct\_declarator {

$$ = MiddleCodeGenerator.PointerDeclarator($1, $2);

};

optional\_pointer\_list:

/\* Empty \*/ { $$ = new List<CCompiler.Type>(); }

| pointer\_list { $$ = $1; };

pointer\_list:

pointer\_marker {

$$ = new List<CCompiler.Type>();

$$.Add($1);

}

| pointer\_list pointer\_marker {

$1.Add($2);

$$ = $1;

};

pointer\_marker:

ASTERRISK optional\_qualifier\_list {

$$ = Specifier.QualifierList($2);

};

A pointer can be quilified with constant and volatile. Note the difference between a constant pointer and a pointer at a constant value. In the former case the pointer itself is constant, it cannot be assigned another address but the value it points at is not constant. In the latter case the pointer can be assigned to another address, but the value it points at is always constant. Naturally, a constant pointer can also point at a constant value.

|  |  |  |
| --- | --- | --- |
| int \* const p; | const int \* p; | const int const \* p; |
| (a) Constant pointer | (b) Pointer at constant value | (c) Constant pointer at  constant value |

optional\_qualifier\_list:

/\* Empty \*/ {

$$ = new List<Mask>();

}

| optional\_qualifier\_list qualifier {

$$ = $1;

$$.Add($2);

};

qualifier:

CONSTANT { $$ = Mask.Constant; }

VOLATILE { $$ = Mask.Volatile; };

### Direct Declarator

The next step after we have handled pointer markers is the direct declarator, which may hold an array or function declaration, or just a single name.

direct\_declarator:

NAME {

$$ = new Declarator($1);

}

A declarator may also another declarator enclosed by parenthesis, in which case we simply return the declarator. The parentheses are present only to change the precedence of the declarator.

| LEFT\_PARENTHESIS declarator RIGHT\_PARENTHESIS {

$$ = $2;

}

A direct declarator can be an array declarator, with a constant expression enclosed by brackets.

| direct\_declarator LEFT\_SQUARE

optional\_constant\_integral\_expression RIGHT\_SQUARE {

$$ = MiddleCodeGenerator.ArrayType($1, $3);

}

Similar to the pointer case above, it is possible to omit the declarator of an array declaration. It is also possible to omit the expression within the brackets. However, the brackets must always be present in an array declaration.

MainParser.gppg

| direct\_declarator

LEFT\_PARENTHESIS parameter\_ellipse\_list RIGHT\_PARENTHESIS {

$$ = MiddleCodeGenerator.

NewFunctionDeclaration($1, $3.First, $3.Second);

}

The next declarator is the old-style function declarator, where the parameter list is a list of strings.

MainParser.gppg

| direct\_declarator LEFT\_PARENTHESIS

optional\_name\_list RIGHT\_PARENTHESIS {

$$ = MiddleCodeGenerator.OldFunctionDeclaration($1, $3);

};

optional\_parameter\_ellipse\_list:

/\* Empty \*/ { $$ = null; }

| parameter\_ellipse\_list { $$ = $1; };

The parameter\_ellipse\_list rule returns a pair of the parameter list and a Boolean value representing the presence of the ellipse marker.

parameter\_ellipse\_list:

parameter\_list {

$$ = new Pair<List<Symbol>,Boolean>($1, false);

}

| parameter\_list COMMA ELLIPSE {

$$ = new Pair<List<Symbol>,Boolean>($1, true);

};

For each parameter declaration, we push the scope of the current symbol table on a stack that we pop after the parsing of the declaration. This is because there are special rules for the storage specifiers of parameters (only auto or register allowed) for the Specifier class to handle (see Chapter 5).

parameter\_list:

{ ScopeStack.Push(SymbolTable.CurrentTable.Scope);

SymbolTable.CurrentTable.Scope = Scope.Parameter;

}

parameter\_declaration {

SymbolTable.CurrentTable.Scope = ScopeStack.Pop();

$$ = new List<Symbol>();

$$.Add($2);

}

| parameter\_list COMMA {

ScopeStack.Push(SymbolTable.CurrentTable.Scope);

SymbolTable.CurrentTable.Scope = Scope.Parameter;

}

parameter\_declaration {

SymbolTable.CurrentTable.Scope = ScopeStack.Pop();

$1.Add($4);

$$ = $1;

};

A parameter declaration can be made up by a declaration specifier list together with a declarator or an abstract declarator, or only a declaration specifier list, since function parameters do not need to be named.

parameter\_declaration:

declaration\_specifier\_list {

$$ = MiddleCodeGenerator.Parameter(Specifier.SpecifierList($1), null);

}

| declaration\_specifier\_list\_x declarator {

$$ = MiddleCodeGenerator.Parameter(SpecifierListStack.Pop(), $2);

}

| declaration\_specifier\_list\_x abstract\_declarator {

$$ = MiddleCodeGenerator.Parameter(SpecifierListStack.Pop(), $2);

};

MainParser.gppg

optional\_name\_list:

/\* Empty \*/ { $$ = new List<string>(); }

| name\_list { $$ = $1; };

name\_list:

NAME {

$$ = new List<string>();

$$.Add($1);

}

| name\_list COMMA NAME {

$1.Add($3);

$$ = $1;

};

### Initialization

The initializer can be either an assignment expression or a block, which is made up by assignment expressions or other blocks. The assignment expression makes sure that the expression cannot hold assignment or comma.

initializer:

assignment\_expression {

$$ = $1;

}

The block holds a list of initializers, which means that initializers can be nested. However, at the deepest level, there are always expressions in the initialization lists.

| LEFT\_BLOCK initializer\_list optional\_comma RIGHT\_BLOCK {

$$ = $2;

};

optional\_comma:

/\* Empty \*/

| COMMA;

initializer\_list:

initializer {

$$ = new List<object>();

$$.Add($1);

}

| initializer\_list COMMA initializer {

$1.Add($3);

$$ = $1;

};

### Abstract Declarator

An abstract declarator is a direct declarator without a name, which is allowed in function parameters and as operand of the sizeof operator. It works in the same way as the regular declarator. The only difference is that an abstract declarator cannot be made up by a name.

abstract\_declarator:

pointer\_list {

$$ = MiddleCodeGenerator.PointerDeclarator ($1, null);

}

| optional\_pointer\_list direct\_abstract\_declarator {

$$ = MiddleCodeGenerator.PointerDeclarator($1, $2);

};

direct\_abstract\_declarator:

LEFT\_PARENTHESIS abstract\_declarator RIGHT\_PARENTHESIS {

$$ = $2;

}

| LEFT\_SQUARE optional\_constant\_integral\_expression RIGHT\_SQUARE {

$$ = MiddleCodeGenerator.ArrayType(null, $2);

}

| direct\_abstract\_declarator

LEFT\_SQUARE optional\_constant\_integral\_expression RIGHT\_SQUARE {

$$ = MiddleCodeGenerator.ArrayType($1, $3);

}

| LEFT\_PARENTHESIS optional\_parameter\_ellipse\_list RIGHT\_PARENTHESIS {

$$ = MiddleCodeGenerator.

NewFunctionDeclaration(null, $2.First, $2.Second);

}

| direct\_abstract\_declarator

LEFT\_PARENTHESIS optional\_parameter\_ellipse\_list RIGHT\_PARENTHESIS {

$$ = MiddleCodeGenerator.

NewFunctionDeclaration($1, $3.First, $3.Second);

};

## Statements

The GenerateFunctionCode is called after the middle code of the function has been generated. Its task is to generate the assembly and object code of the function. If the function returns void and is the main function, the Exit middle code is added to middle code list. If the it returns void and is not the main, the Return middle code instruction is added.

Then the middle code is being optimized and a list of base blocks is generated.

The object code is actually generated twice, since we want to know which symbols are not used and can be removed from the symbol table.

If the function is the main function we have to modify the entry point of the function.

Finally, the function is added to the linker map and the symbol table is popped.

The generateAssemblyText method generates the assembly code for the function.

When a function containing goto statements has been parsed, the addresses need to be set with the help of the label map.

### The if-else Problem

The if-else problem is the problem of syntactically interpret the leftmost source code below. Semantically, the middle interpretation of the left statement is the correct one, each else shall be connected to the latest preceding if.

|  |  |  |
| --- | --- | --- |
| if (a < b)  if (c < d)  e = 1;  else  f = 2;  (a) Ambiguous C code | if (a < b) {  if (c < d)  e = 1;  else  f = 2;  }  (b) Correct Interpretation | if (a < b)  if (c < d) {  e = 1;  }  else  f = 2;  (c) Incorrect interpretation |

Below is a simple set of statement rules. Unfortunately, they are ambiguous in that way that the an else does not have to be connected to the latest preceding if, resulting in both the middle and rightmost semantically interpretation above, depending in which order the rules are applied.

statement ::=

IF LEFT\_PAREN logical\_expression RIGHT\_PAREN statement

| IF LEFT\_PAREN logical\_expression RIGHT\_PAREN statement ELSE statement

| ...

To solve the problem, we need a more complicated set of rules that works with open and closed statements. The following set is unambiguous in that way that it always connects each *else* with the latest preceding *if*. A theoretical explanation of the open-closed statement solution is beyond the scope of this book, but I recommend the Dragon Book by Aho et al. for a closer look.

statement:

opened\_statement

| closed\_statement

opened\_statement:

if ( expression ) statement

| if ( expression ) closed\_statement else opened\_statement

| switch ( expression ) opened\_statement

| case constant\_integral\_expression : opened\_statement

| while ( expression ) opened\_statement

| for ( optional\_expression ; optional\_expression ; optional\_expression )

opened\_statement

| name : opened\_statement

closed\_statement:

if ( expression ) closed\_statement else closed\_statement

| switch ( expression ) closed\_statement

| while ( expression ) closed\_statement

| for ( optional\_expression ; optional\_expression ; optional\_expression )

closed\_statement

| do statement while ( expression ) ;

| case constant\_integral\_expression : closed\_statement

| default : closed\_statement

| continue ;

| break ;

| { optional\_statement\_list }

| goto name ;

| return optional\_expression ;

| optional\_expression ;

| declaration ;

| jump\_register ( register\_name ) ;

| interrupt ( constant\_integral\_expression ) ;

| syscall ( ) ;

### The Forward-Jump Problem (Backpatching)

When generating middle code instructions for expressions or statements, a common situation is that we generate forward jump instructions without yet knowing the target address. One way to solve the problem is to work with backpatching, which means that we use sets to keep track of jump instructions with yet unknown target addresses that eventually becomes filled in when the target addresses have become known.

For instance, let us look at the while statement to the left. The parsing of expression a < b will generate the middle code in the middle and the sets to the right. The middle code line will a have undefined targets since we not yet know where to jump when the expression becomes evaluated to true or false.

|  |  |  |
| --- | --- | --- |
| while (a < b)  ++a;  (a) C code | 1. if a < b goto ?  2. goto ?  (a) Middle code | a < b: true\_set = {1}  false\_set = {2}  (a) True and false-sets |

When parsing the while statement, the sets of the expression becomes backpatched: If the a < b expression is true, line 1 shall jump to line 3. In case of false, line 2 shall jump to line 5, which is the middle code instruction following the middle code for the while statement.

backpatch(true\_set, 3); => backpatch({1}, 3);

backpatch(false\_set, 5); => backpatch({2}, 5);

1. if a < b goto 3

2. goto 5

3. ++a

4. goto 1

5. ...

|  |  |  |
| --- | --- | --- |
|  |  |  |

If we let the while statement above be surrounded by an if-else statement, we have both the true and false-sets of the if expression c < d and the sets of the while expression a < b. Note that line 6 and 7 do not need to be backpatched, since they jump backwards to known line numbers.

|  |  |  |
| --- | --- | --- |
| if (c < d)  while (a < b)  ++a;  else  ++b; | 1. if c < d goto ?  2. goto ?  3. if a < b goto ?  4. goto ?  5. ++a  6. goto 3  7. goto 9  8. ++b  9. ... | c < d: true\_set1: {1}  false\_set1: {2}  a < b: true\_set2: {3}  false\_set2: {4} |

When the if and while statements become parsed the following call will occur:

backpatch(true\_set\_1, 3); => backpatch({1}, 3);

backpatch(false\_set\_1, 8); => backpatch({2}, 8);

backpatch(true\_set\_2, 5); => backpatch({3}, 5);

backpatch(false\_set\_2, 7); => backpatch({4}, 7);

|  |  |  |
| --- | --- | --- |
| 1. if c < d goto 3  2. goto 8  3. if a < b goto 5  4. goto 7  5. ++a  6. goto 3  7. goto 9  8. ++b  9. ... |  |  |

The generated middle code above is ineffective, but the middle code optimizer of Chapter 11 will change it into the following middle code.

|  |  |  |
| --- | --- | --- |
| 1. if c >= d goto 5  2. if a >= b goto 6  3. ++a  4. goto 2  5. ++b  6. ... |  |  |

Another example is the evaluation of a logical and expression. Since C takes advantage of lazy evaluation, we have to insert jump instructions in the middle code.

|  |  |  |
| --- | --- | --- |
| if ((a < b) && (c < d))  ++a;  else  ++b; | 1. if a < b goto ?  2. goto ?  3. if c < d goto ?  4. goto ?  5. ++a  6. goto 7  7. ++b  8. ... |  |
| (a) C code | (b) Middle code before  backpatching | |
|  |  |  |
| 1. if a < b goto ?  2. goto ?  3. if c < d goto ?  4. goto ?  5. ++a  6. goto 7  7. ++b  8. ... | 1. if a < b goto ?  2. goto ?  3. if c < d goto ?  4. goto ?  5. ++a  6. goto 7  7. ++b  8. ... |  |
| (c) Middle code after  backpatching | (c) Middle code after  optimization |  |

When the first expression (a < b) is evaluated, its true-set is {1} and its false-set is {2}. When the second expression is evaluated the true-set of the first expression is backpatched to the beginning of the second expression. This means that if the first expression is true we continue to evaluate the second expression. However, if the first expression is false the second expression will never be evaluated since it not necessary, the total expression will be false regardless of the value of the second expression. The result will be that the true-set of the total expression is the one of the second expression and the false-set is the union of the falses sets of both expressions.

|  |  |  |
| --- | --- | --- |
| 1. if a < b goto 3  2. goto ?  3. if c < d goto ?  4. goto ?  5. ++a  6. goto 7  7. ++b | true-set: {3}  false-set: {2, 4} |  |

In a similar way, the true-set of an expression with logical-or will be the union of both true-sets while the false-set will the false-set of the second expression.

|  |  |  |
| --- | --- | --- |
| if ((a < b) || (c < d))  ++a;  else  ++b; | 1. if a < b goto ?  2. goto 3  3. if c < d goto ?  4. goto ?  5. ++a  6. goto 7  7. ++b  8. ... | true-set: {1, 3}  false-set: {4} |

### The Statement Class

Let us look at the Statement class. A statement holds a list of middle code instructions, and the next-set, which is a set of jump instructions that shall be backpatched to jump to the next instruction after the current statement.

Statement.cs

using System.Collections.Generic;

namespace CCompiler {

public class Statement {

private List<MiddleCode> m\_codeList;

private ISet<MiddleCode> m\_nextSet;

public Statement(List<MiddleCode> codeList,

ISet<MiddleCode> nextSet = null) {

Assert.ErrorXXX(codeList != null);

m\_codeList = codeList;

m\_nextSet = (nextSet != null) ? nextSet : (new HashSet<MiddleCode>());

}

public List<MiddleCode> CodeList {

get { return m\_codeList; }

}

public ISet<MiddleCode> NextSet {

get { return m\_nextSet; }

}

}

}

### Statements

As stated above, a statement can be an open or a closed statement.

MainParser.cs

statement:

opened\_statement { $$ = $1; }

| closed\_statement { $$ = $1; };

### The If Statement

opened\_statement:

IF LEFT\_PARENTHESIS expression RIGHT\_PARENTHESIS statement {

$$ = MiddleCodeGenerator.IfStatement($3, $5);

};

### The If-Else Statement

opened\_statement:

IF LEFT\_PARENTHESIS expression RIGHT\_PARENTHESIS closed\_statement

ELSE opened\_statement {

$$ = MiddleCodeGenerator.IfElseStatement($3, $5, $7);

};

closed\_statement:

IF LEFT\_PARENTHESIS expression RIGHT\_PARENTHESIS closed\_statement

ELSE closed\_statement {

$$ = MiddleCodeGenerator.IfElseStatement($3, $5, $7);

};

### The Switch Statement

switch\_header:

/\* Empty. \*/ { MiddleCodeGenerator.SwitchHeader(); };

opened\_statement:

SWITCH switch\_header LEFT\_PARENTHESIS expression RIGHT\_PARENTHESIS

opened\_statement {

$$ = MiddleCodeGenerator.SwitchStatement($4, $6);

};

closed\_statement:

SWITCH switch\_header LEFT\_PARENTHESIS expression RIGHT\_PARENTHESIS

closed\_statement {

$$ = MiddleCodeGenerator.SwitchStatement($4, $6);

};

### The Case Statement

opened\_statement:

CASE constant\_integral\_expression COLON opened\_statement {

$$ = MiddleCodeGenerator.CaseStatement($2, $4);

};

closed\_statement:

CASE constant\_integral\_expression COLON closed\_statement {

$$ = MiddleCodeGenerator.CaseStatement($2, $4);

};

### The Default Statement

closed\_statement:

DEFAULT COLON closed\_statement {

$$ = MiddleCodeGenerator.DefaultStatement($3);

};

### The While Statement

loop\_header:

/\* Empty. \*/ { MiddleCodeGenerator.LoopHeader(); };

opened\_statement:

WHILE loop\_header LEFT\_PARENTHESIS expression RIGHT\_PARENTHESIS

opened\_statement {

$$ = MiddleCodeGenerator.WhileStatement($4, $6);

};

closed\_statement:

WHILE loop\_header LEFT\_PARENTHESIS expression RIGHT\_PARENTHESIS

closed\_statement {

$$ = MiddleCodeGenerator.WhileStatement($4, $6);

};

### The Do Statement

closed\_statement:

DO loop\_header statement WHILE

LEFT\_PARENTHESIS expression RIGHT\_PARENTHESIS SEMICOLON {

$$ = MiddleCodeGenerator.DoStatement($3, $6);

};

### The For Statement

opened\_statement:

FOR loop\_header LEFT\_PARENTHESIS optional\_expression SEMICOLON

optional\_expression SEMICOLON optional\_expression RIGHT\_PARENTHESIS

opened\_statement {

$$ = MiddleCodeGenerator.ForStatement($4, $6, $8, $10);

};

closed\_statement:

FOR loop\_header LEFT\_PARENTHESIS optional\_expression SEMICOLON

optional\_expression SEMICOLON optional\_expression RIGHT\_PARENTHESIS

closed\_statement {

$$ = MiddleCodeGenerator.ForStatement($4, $6, $8, $10);

};

### Label and Jump Statement

opened\_statement:

NAME COLON opened\_statement {

$$ = MiddleCodeGenerator.LabelStatement($1, $3);

};

closed\_statement:

GOTO NAME SEMICOLON {

$$ = MiddleCodeGenerator.GotoStatement($2);

};

To keep track of the label statements we have the label map m\_labelMap with the label name as key and the first middle code instruction of the following statement as value.

### The Return Statement

The return statement may have an optional expression.

closed\_statement:

RETURN optional\_expression SEMICOLON {

$$ = MiddleCodeGenerator.ReturnStatement($2);

};

### Optional Expression Statements

A statement can also be made up by an optional expression; that is, the statement is an expression followed by a semicolon, or simply a semicolon.

An expression statement is an optional expression followed by a semicolon. The expression is evaluated and its true and false-sets are backpatched to the beginning of the next statement. Note that we do not use the result of the expression, the only interested part is its potential side effects. If it has no side effects, the statements will be removed by the middle code optimizer in Chapter 0.

MainParser.gppg

closed\_statement:

optional\_expression SEMICOLON {

$$ = MiddleCodeGenerator.ExpressionStatement($1);

};

### Block Statements

A statement can be an optional sequence of statements enclosed in brackets. The sequence is parsed with a new symbol table, since two symbols with the same name can be defined in different blocks.

closed\_statement:

LEFT\_BLOCK {

SymbolTable.CurrentTable =

new SymbolTable(SymbolTable.CurrentTable, Scope.Block);

}

optional\_declaration\_list optional\_statement\_list RIGHT\_BLOCK {

SymbolTable.CurrentTable =

SymbolTable.CurrentTable.ParentTable;

$4.CodeList.InsertRange(0, $3);

$$ = $4;

}

In case of an empty statement list, we return a statement with an empty code list and an empty next-set.

optional\_statement\_list:

/\* Empty \*/ {

$$ = new Statement(new List<MiddleCode>(),

new HashSet<MiddleCode>());

}

In case of a non-empty statement list, we add the code list of the statements of the statement list. For each statement in the list, except the last statement, we backpatch the next-set to the beginning of the code list of the next-set. The result is a statement with the total middle code list and the next-set of the last statement.

| optional\_statement\_list statement {

MiddleCodeGenerator.Backpatch($1.NextSet, $2.CodeList);

List<MiddleCode> codeList = new List<MiddleCode>();

codeList.AddRange($1.CodeList);

codeList.AddRange($2.CodeList);

$$ = new Statement(codeList, $2.NextSet);

};

### Jump Register Statements

When performing a call to a function which address is stored in a pointer variable we need to store the address in a register and jump to that register.

closed\_statement:

JUMP\_REGISTER LEFT\_PARENTHESIS REGISTER\_NAME RIGHT\_PARENTHESIS SEMICOLON {

$$ = MiddleCodeGenerator.JumpRegisterStatement($3);

};

### Interrupt Statements

When making system calls, an interrupt occurs. The operand is an integral value of short size (1 byte).

closed\_statement:

INTERRUPT LEFT\_PARENTHESIS constant\_integral\_expression RIGHT\_PARENTHESIS

SEMICOLON {

$$ = MiddleCodeGenerator.InterruptStatement($3);

};

### System Call Statements

System calls for Linux.

closed\_statement:

SYSCALL LEFT\_PARENTHESIS RIGHT\_PARENTHESIS SEMICOLON {

$$ = MiddleCodeGenerator.SystemCallStatement();

};

## Expressions

The third part of the parser is the expressions. We start with the expression of lowest precedence and add a new rule for each new level of precedence.

### The Expression Class (Short and Long List)

Similar to the statement case above, we have the Expression class to handle expressions. It handles a symbol, a register, and the short list and the long list. The short list and long list both hold the middle code of the expression. The difference is that the short list holds only the side effects of the expression, while the long holds the whole expression. For instance, in the following expression the long list holds the function call, decrement, and addition. The short list holds the only the side effects: the function call and decrement, but not the addition.

f(x) + a--;

The following statement is correct, but since in holds no side-effects its short list will be empty. In the end, no code will be generated.

a + (b \* c);

The register parameter is used in system calls when we need to access or assign a specific register.

Expression.cs

using System.Numerics;

using System.Collections.Generic;

namespace CCompiler {

public class Expression {

private Symbol m\_symbol;

private List<MiddleCode> m\_shortList;

private List<MiddleCode> m\_longList;

private Register? m\_register;

public Expression(Symbol symbol, List<MiddleCode> shortList,

List<MiddleCode> longList, Register? register = null) {

m\_symbol = symbol;

m\_shortList = (shortList != null) ? shortList : (new List<MiddleCode>());

m\_longList = (longList != null) ? longList : (new List<MiddleCode>());

m\_register = register;

}

public Symbol Symbol {

get { return m\_symbol; }

}

public List<MiddleCode> ShortList {

get { return m\_shortList; }

}

public List<MiddleCode> LongList {

get { return m\_longList; }

}

public Register? Register {

get { return m\_register; }

}

public override string ToString() {

return m\_symbol.ToString();

}

}

}

### Optional Expressions

On several occasions ,such as the for statement, the presence of expressions are optional. Therefore, we have a rule for an optional expression. In the absence of an expression, the result is null.

MainParser.cs

optional\_expression:

/\* Empty \*/ { $$ = null; }

| expression { $$ = $1; };

### The Comma Expression

We begin the parsing of the expressions with the operator of lowest precedence, which is the comma operator. A comma expression may be an assignment expression, or two expressions separated by a comma.

expression:

assignment\_expression {

$$ = $1;

}

| expression COMMA assignment\_expression {

$$ = MiddleCodeGenerator.CommaExpression($1, $3);

};

### The Assignment Expression

There are two kinds of assignment: simple and compound. In the case of compound assignment, we first evaluate the corresponding binary expression (addition in the example below), assign its result to a temporary variable that we then assign to original left operand.

|  |  |
| --- | --- |
| x += y; | t = x + y  x = t  (b) Equivalent simple assignment |
| (a) Compound assignment |

An assignment expression can be a conditional expression, a simple assignment, or a compound assignment.

MainParser.cs

assignment\_expression:

condition\_expression {

$$ = $1;

}

| prefix\_expression assignment\_operator assignment\_expression {

$$ = MiddleCodeGenerator.AssignmentExpression($2, $1, $3);

};

assignment\_operator:

ASSIGN { $$ = MiddleOperator.Assign; }

| ADD\_ASSIGN { $$ = MiddleOperator.BinaryAdd; }

| SUBTRACT\_ASSIGN { $$ = MiddleOperator.BinarySubtract; }

| MULTIPLY\_ASSIGN { $$ = MiddleOperator.SignedMultiply; }

| DIVIDE\_ASSIGN { $$ = MiddleOperator.SignedDivide; }

| MODULO\_ASSIGN { $$ = MiddleOperator.SignedModulo; }

| AND\_ASSIGN { $$ = MiddleOperator.BitwiseAnd; }

| OR\_ASSIGN { $$ = MiddleOperator.BitwiseOr; }

| XOR\_ASSIGN { $$ = MiddleOperator.BitwiseXOr; }

| LEFT\_SHIFT\_ASSIGN { $$ = MiddleOperator.ShiftLeft; }

| RIGHT\_SHIFT\_ASSIGN { $$ = MiddleOperator.ShiftRight; };

### The Condition Expression

The conditional operator is the only operator in C that takes three operands. It applies lazy evaluation, which means that only one of the second or third expression will be evaluated, depending on the value of the first expression.

MainParser.gppg

condition\_expression:

logical\_or\_expression {

$$ = $1;

}

| logical\_or\_expression QUESTION\_MARK expression COLON condition\_expression{

$$ = MiddleCodeGenerator.ConditionalExpression($1, $3, $5);

};

### Constant Expression

On several occasions, such array limits and enumeration values, we need to parse constant integral expressions. The expression is a conditional expression since it cannot hold commas or assignments.

MainParser.cs

optional\_constant\_integral\_expression:

/\* Empty \*/ { $$ = null; }

| constant\_integral\_expression { $$ = $1; };

constant\_integral\_expression:

condition\_expression {

$$ = MiddleCodeGenerator.ConstantIntegralExpression($1);

};

### Logical Expressions

The logical and and or operators have one rule each, since and have higher precedence than or.

MainParser.cs

logical\_or\_expression:

logical\_and\_expression {

$$ = $1;

}

| logical\_or\_expression LOGICAL\_OR logical\_and\_expression {

$$ = MiddleCodeGenerator.LogicalOrExpression($1, $3);

};

logical\_and\_expression:

bitwise\_or\_expression {

$$ = $1;

}

| logical\_and\_expression LOGICAL\_AND bitwise\_or\_expression {

$$ = MiddleCodeGenerator.LogicalAndExpression($1, $3);

};

### Bitwise Expressions

Unlike the logical expressions in the previous section, the bitwise operators do not apply lazy evaluation, which means that the right expressions is always evaluated regardless of the value of the left expression. There are three bitwise operators: or (inclusive or), xor (exclusive or), and and. The expressions have one rule each since they have different precedence. They take two integer values and perform operations on each bit:

* or: the resulting bit is one if at least one bit is one, zero otherwise
* xor: the resulting bit is one if exact one bit (but not both bits) is one, zero otherwise
* and: the resulting bit is one if at both values are one, zero otherwise

MainParser.gppg

bitwise\_or\_expression:

bitwise\_xor\_expression {

$$ = $1;

}

| bitwise\_or\_expression BITWISE\_OR bitwise\_xor\_expression {

$$ = MiddleCodeGenerator.BitwiseExpression

(MiddleOperator.BitwiseOr, $1, $3);

};

bitwise\_xor\_expression:

bitwise\_and\_expression {

$$ = $1;

}

| bitwise\_xor\_expression BITWISE\_XOR bitwise\_and\_expression {

$$ = MiddleCodeGenerator.BitwiseExpression

(MiddleOperator.BitwiseXOr, $1, $3);

};

bitwise\_and\_expression:

equality\_expression {

$$ = $1;

}

| bitwise\_and\_expression AMPERSAND equality\_expression {

$$ = MiddleCodeGenerator.BitwiseExpression

(MiddleOperator.BitwiseAnd, $1, $3);

};

### Shift Expression

The left and right shift expression must hold integral types. The right expression is type cast to a one-byte integer value.

MainParser.gppg

shift\_expression:

add\_expression {

$$ = $1;

}

| shift\_expression shift\_operator add\_expression {

$$ = MiddleCodeGenerator.ShiftExpression($2, $1, $3);

};

shift\_operator:

LEFT\_SHIFT { $$ = MiddleOperator.ShiftLeft; }

| RIGHT\_SHIFT { $$ = MiddleOperator.ShiftRight; };

### Equality and Relation Expressions

Values of all types except structs or unions can be compared with equality and inequality operator. However, arrays and functions are converted to pointers before the comparation. The equality operators have higher precedence than the relation operators, but they both call RelationalExpression in MiddleCodeGenerator.

MainParser.gppg

equality\_expression:

relation\_expression {

$$ = $1;

}

| equality\_expression equality\_operator relation\_expression {

$$ = MiddleCodeGenerator.RelationalExpression($2, $1, $3);

};

equality\_operator:

EQUAL { $$ = MiddleOperator.Equal; }

| NOT\_EQUAL { $$ = MiddleOperator.NotEqual; };

relation\_expression:

shift\_expression {

$$ = $1;

}

| relation\_expression relation\_operator shift\_expression {

$$ = MiddleCodeGenerator.RelationalExpression ($2, $1, $3);

};

relation\_operator:

LESS\_THAN { $$ = MiddleOperator.SignedLessThan; }

| LESS\_THAN\_EQUAL { $$ = MiddleOperator.SignedLessThanEqual; }

| GREATER\_THAN { $$ = MiddleOperator.SignedGreaterThan; }

| GREATER\_THAN\_EQUAL { $$ = MiddleOperator.SignedGreaterThanEqual; };

### Addition and Subtraction Expression

The addition and subtraction expression are a bit more complicated, since we need to take pointer arithmetic into consideration.

MainParser.gppg

add\_expression:

multiply\_expression {

$$ = $1;

}

| add\_expression addition\_operator multiply\_expression {

$$ = MiddleCodeGenerator.AdditionExpression($2, $1, $3);

};

add\_operator:

PLUS { $$ = MiddleOperator.BinaryAdd; }

| MINUS { $$ = MiddleOperator.BinarySubtract; };

### Multiplication Expressions

The multiplication expression takes the multiply, divide, and module operators.

MainParser.gppg

multiply\_expression:

type\_cast\_expression {

$$ = $1;

}

| multiply\_expression multiply\_operator type\_cast\_expression {

$$ = MiddleCodeGenerator.MultiplyExpression($2, $1, $3);

};

The multiplication operators are multiply, division, and modulo.

multiply\_operator:

ASTERRISK { $$ = MiddleOperator.SignedMultiply; }

| DIVIDE { $$ = MiddleOperator.SignedDivide; }

| MODULO { $$ = MiddleOperator.SignedModulo; };

### Cast Expressions

A type cast expression is a type name within parentheses followed by any kind of expression.

MainParser.gppg

type\_cast\_expression:

prefix\_expression {

$$ = $1;

}

| LEFT\_PARENTHESIS type\_name RIGHT\_PARENTHESIS type\_cast\_expression {

$$ = MiddleCodeGenerator.CastExpression($2, $4);

};

The type name is a declaration specifier list with or without an abstract declarator.

type\_name:

declaration\_specifier\_list {

$$ = MiddleCodeGenerator.TypeName(Specifier.SpecifierList($1), null);

}

| declaration\_specifier\_list {

SpecifierListStack.Push(Specifier.SpecifierList($1));

}

abstract\_declarator {

$$ = MiddleCodeGenerator.TypeName(SpecifierListStack.Pop(), $3);

};

### Prefix Expression

In C, there are several prefix expressions:

* Unary add and minus
* Logical not
* Bitwise not
* The sizeof operator
* The address operator
* The dereference operator
* Prefix increment and decrement

prefix\_expression:

postfix\_expression {

$$ = $1;

}

| prefix\_add\_operator type\_cast\_expression {

$$ = MiddleCodeGenerator.UnaryExpression($1, $2);

}

| LOGICAL\_NOT type\_cast\_expression {

$$ = MiddleCodeGenerator.LogicalNotExpression($2);

}

| BITWISE\_NOT type\_cast\_expression {

$$ = MiddleCodeGenerator.BitwiseNotExpression($2);

}

The sizeof operator can be applied to an expression or a type name within parentheses. It does not apply to functions or bitfields. When applied to an array, it gives the total size of the array, not the size of its address.

| SIZEOF prefix\_expression {

$$ = MiddleCodeGenerator.SizeOfExpression($2);

}

| SIZEOF LEFT\_PARENTHESIS type\_name RIGHT\_PARENTHESIS {

$$ = MiddleCodeGenerator.SizeOfType($3);

}

The address operator takes the address of its operand and the return type is a pointer to the operand type, unless it is an array, in which case the result is a pointer to the array type. The operand cannot have register storage.

The address operator does not apply to bitfields symbols or symbols with register storage. A symbol is addressable if it does not have register storage and is not a bitfield in a struct or a union.

AMPERSAND type\_cast\_expression {

$$ = MiddleCodeGenerator.AddressExpression($2);

};

The dereference operator takes a pointer or array as operand.

| ASTERRISK type\_cast\_expression {

$$ = MiddleCodeGenerator.DereferenceExpression($2);

}

The postfix increment and decrement change the operand and return the original value. The operand has to be a left-value of arithmetic or pointer type.

| increment\_operator prefix\_expression {

$$ = MiddleCodeGenerator.PrefixIncrementExpression($1, $2);

};

prefix\_add\_operator:

PLUS { $$ = MiddleOperator.UnaryAdd; }

| MINUS { $$ = MiddleOperator.UnarySubtract; };

increment\_operator:

INCREMENT { $$ = MiddleOperator.Increment; }

| DECREMENT { $$ = MiddleOperator.Decrement; };

### Postfix Expression

In C, there are several postfix expressions

* Prefix increment and decrement
* Dot and Arrow
* Array indexing
* Function calls

postfix\_expression:

primary\_expression {

$$ = $1;

}

| postfix\_expression increment\_operator {

$$ = MiddleCodeGenerator.PostfixIncrementExpression($2, $1);

}

The dot and arrow operator takes a pointer to a struct or union, and the name of one of their members as operands.

| postfix\_expression DOT NAME {

$$ = MiddleCodeGenerator.DotExpression($1, $3);

}

| postfix\_expression ARROW NAME {

$$ = MiddleCodeGenerator.ArrowExpression($1, $3);

}

| postfix\_expression LEFT\_SQUARE expression RIGHT\_SQUARE {

$$ = MiddleCodeGenerator.IndexExpression($1, $3);

}

| postfix\_expression {

MiddleCodeGenerator.FunctionPreCall($1);

}

LEFT\_PARENTHESIS optional\_argument\_expression\_list RIGHT\_PARENTHESIS {

$$ = MiddleCodeGenerator.CallExpression($1, $4);

};

The arguments of the function call are gathered in a list.

argument\_expression\_list:

assignment\_expression {

$$ = new List<Expression>();

$$.Add(MiddleCodeGenerator.ArgumentExpression(0, $1));

}

| argument\_expression\_list COMMA assignment\_expression {

$1.Add(MiddleCodeGenerator.ArgumentExpression($1.Count, $3));

$$ = $1;

};

### Primary Expressions

We have finally reached the last kind of expression. A primary expression may be another expression within parentheses, a value, a symbol, a register, the carry flag, or the stack top position.

In case of an expression within parentheses, the result is simply the expression. The parentheses are only present to change the precedence of the expression.

primary\_expression:

LEFT\_PARENTHESIS expression RIGHT\_PARENTHESIS {

$$ = $2;

}

| VALUE {

$$ = MiddleCodeGenerator.ValueExpression($1);

}

| NAME {

$$ = MiddleCodeGenerator.NameExpression($1);

}

| REGISTER\_NAME {

$$ = MiddleCodeGenerator.RegisterExpression($1);

}

The carry flag is also only used internally in conjunction with system calls. On some occasion the interrupt call returns information stored in the carry flag.

| CARRY\_FLAG {

$$ = MiddleCodeGenerator.CarryFlagExpression();

}

The address of the top position of the call stack is needed by the malloc function in the C standard library, see Appendix A.

| STACK\_TOP {

$$ = MiddleCodeGenerator.StackTopExpression();

};

# Middle Code Generation

The rules of the parser call corresponding methods of the MiddleCodeGenerator class to perform type checking, build the symbol table, and generate the middle code.

## The MiddleCode Class

The middle code of the compiler in this book is three-address code. As the name implies, each instruction is made up by one operator and at most three addresses, which are references to objects of the Object class, or its sub classes, such as Symbol, Register, String, Integer, or Double.

MiddleCode.cs

using System;

using System.Text;

using System.Collections.Generic;

The MiddleCode class holds the operator and the array of the three operands. It also holds a set of test methods.

namespace CCompiler {

public class MiddleCode {

private MiddleOperator m\_middleOperator;

private object[] m\_operandArray = new object[3];

public MiddleCode(MiddleOperator middleOp, object operand0 = null,

object operand1 = null, object operand2 = null) {

m\_middleOperator = middleOp;

m\_operandArray[0] = operand0;

m\_operandArray[1] = operand1;

m\_operandArray[2] = operand2;

}

public MiddleOperator Operator {

get { return m\_middleOperator; }

set { m\_middleOperator = value; }

}

public object this[int index] {

get { return m\_operandArray[index]; }

set { m\_operandArray[index] = value; }

}

public void Clear() {

m\_middleOperator = MiddleOperator.Empty;

m\_operandArray[0] = null;

m\_operandArray[1] = null;

m\_operandArray[2] = null;

}

The IsGoto, IsCarry, and IsRelation methods test whether the middle code is a jump instruction.

public bool IsGoto() {

return (m\_middleOperator == MiddleOperator.Jump);

}

public bool IsCarry() {

return (m\_middleOperator == MiddleOperator.Carry) ||

(m\_middleOperator == MiddleOperator.NotCarry);

}

public bool IsRelation() {

switch (m\_middleOperator) {

case MiddleOperator.Case:

case MiddleOperator.Equal:

case MiddleOperator.NotEqual:

case MiddleOperator.SignedLessThan:

case MiddleOperator.SignedLessThanEqual:

case MiddleOperator.SignedGreaterThan:

case MiddleOperator.SignedGreaterThanEqual:

case MiddleOperator.UnsignedLessThan:

case MiddleOperator.UnsignedLessThanEqual:

case MiddleOperator.UnsignedGreaterThan:

case MiddleOperator.UnsignedGreaterThanEqual:

return true;

default:

return false;

}

}

public bool IsRelationCarryOrGoto() {

return IsRelation() || IsCarry() || IsGoto();

}

The IsBinary, IsCommutative, and IsShift methods test the operator of the middle code.

public bool IsBinary() {

switch (m\_middleOperator) {

case MiddleOperator.BinaryAdd:

case MiddleOperator.BinarySubtract:

case MiddleOperator.SignedMultiply:

case MiddleOperator.SignedDivide:

case MiddleOperator.SignedModulo:

case MiddleOperator.UnsignedMultiply:

case MiddleOperator.UnsignedDivide:

case MiddleOperator.UnsignedModulo:

case MiddleOperator.LogicalOr:

case MiddleOperator.LogicalAnd:

case MiddleOperator.BitwiseOr:

case MiddleOperator.BitwiseXOr:

case MiddleOperator.BitwiseAnd:

case MiddleOperator.ShiftLeft:

case MiddleOperator.ShiftRight:

return true;

default:

return false;

}

}

public bool IsCommutative() {

switch (m\_middleOperator) {

case MiddleOperator.BinaryAdd:

case MiddleOperator.SignedMultiply:

case MiddleOperator.UnsignedMultiply:

case MiddleOperator.BitwiseOr:

case MiddleOperator.BitwiseXOr:

case MiddleOperator.BitwiseAnd:

return true;

default:

return false;

}

}

public static bool IsShift(MiddleOperator middleOp) {

switch (middleOp) {

case MiddleOperator.ShiftLeft:

case MiddleOperator.ShiftRight:

return true;

default:

return false;

}

}

The ToString method simply returns a string holding the operator and the operand array. It is only used when adding comments to the final assembly code.

public override string ToString() {

return m\_middleOperator + ToString(m\_operandArray[0]) +

ToString(m\_operandArray[1]) + ToString(m\_operandArray[2]);

}

private static string ToString(object value) {

if (value != null) {

return (" " + value.ToString().Replace("\n", "\\n"));

}

else {

return "";

}

}

}

}

The MiddleCodeGenerator class is a large class that generates the middle code representing the source code.

MiddleCodeGenerator.cs

using System;

using System.IO;

using System.Linq;

using System.Numerics;

using System.Collections.Generic;

To begin with, we have the AddMiddleCode methods that add a MiddleCode object to the given code list.

namespace CCompiler {

public class MiddleCodeGenerator {

public static MiddleCode AddMiddleCode(List<MiddleCode> codeList,

MiddleOperator op, object operand0 = null,

object operand1 = null, object operand2 = null) {

MiddleCode middleCode = new MiddleCode(op, operand0, operand1, operand2);

codeList.Add(middleCode);

return middleCode;

}

### Backpatching

When generating middle code, there are numerous occasions that we generate forward jump instructions without knowing which instruction to jump to. In those cases, we store the instructions in sets and later go back and fill in the jump address. The process is called backpatching.

The static Backpatch methods backpatch a single address or a set of addresses. The first two methods call the last two methods with the current size of the middle code list as target address. In this way, the target address becomes the next middle code instruction to become generated. The backpatching is performed by setting the third address of the middle code, since the third address is always the jump address in conditional and unconditional jump instructions.

The first Backpatch method takes set of jump instructions and a list of instructions, where its first instruction is the target of the jump instructions. If the code list is empty, we add an empty instruction in order to make sure there is always a target to jump to. Then we call the second Backpatch method with the set of jump instruction and the target instruction.

MiddleCodeGenerator.cs

public static void Backpatch(ISet<MiddleCode> sourceSet,

List<MiddleCode> list) {

if (list.Count == 0) {

AddMiddleCode(list, MiddleOperator.Empty);

}

Backpatch(sourceSet, list[0]);

}

The second Backpatch method iterates through the set of jump instructions and set its first reference (index 0) to the target. The first reference of each jump instruction is its target address.

public static void Backpatch(ISet<MiddleCode> sourceSet,

MiddleCode target) {

foreach (MiddleCode source in sourceSet) {

source[0] = target;

}

}

The GetFirst method returns the first instruction in the middle code instruction list. It also adds an empty instruction if the list is empty. This is done for the backpatching to always have an instruction to set as the target.

private static MiddleCode GetFirst(List<MiddleCode> list) {

if (list.Count == 0) {

AddMiddleCode(list, MiddleOperator.Empty);

}

return list[0];

}

## Declarations

This section holds methods, called by the parser, that generates the symbol table and the middle code list.

### Function Definition

The FunctionHeader method is called before the middle code of the function has been generated. Its task is to store the symbol of the function in the symbol table, and to create a new symbol table for the symbols of the function.

MiddleCodeGenerator.cs

public static void FunctionHeader(Specifier specifier,

Declarator declarator) {

Storage? storage;

Type returnType;

If specifier is not null, we extract the storage and return type of the function from the specifier.

if (specifier != null) {

storage = specifier.Storage;

returnType = specifier.Type;

}

If specifier is null, there was no storage specifier or return type defined, in which case the storage is to be set to extern and the return type to signed integer.

else {

storage = Storage.Extern;

returnType = Type.SignedIntegerType;

}

We add the return type to the pointer declarator. In this way the type become complete: a function with a return type.

declarator.Add(returnType);

Then we perform some error checking. A function may lack a name in a function declaration. However, in a function definition it must have a name.

Assert.Error(declarator.Name != null,

Message.Unnamed\_function\_definition);

Technically, the declaration may have any type, and we must check that it really is a function declaration.

Assert.Error(declarator.Type.IsFunction(), declarator.Name,

Message.Not\_a\_function);

The public and static member variable CurrentFunction is a reference to the symbol holding the current function. This reference is used on every occasion during the parsing of the function body when we add or look up a symbol.

SymbolTable.CurrentFunction =

new Symbol(declarator.Name, storage, declarator.Type);

A function definition must be static or extern, it cannot be auto, register, or typedef.

Assert.Error(SymbolTable.CurrentFunction.IsStaticOrExtern(),

declarator.Name,

Message.A\_function\_must\_be\_static\_or\_extern);

Every function definition, as well as static variables, are added to the global static set. They will in the end be translated into assembly code.

SymbolTable.CurrentFunction.FunctionDefinition = true;

SymbolTable.CurrentTable.AddSymbol(SymbolTable.CurrentFunction);

If this function is the main function of the C code, the return type must be void or (signed or unsigned) integer.

if (SymbolTable.CurrentFunction.UniqueName.Equals("main")) {

Assert.Error(returnType.IsVoid() || returnType.IsInteger(), "main",

Message.Function\_main\_must\_return\_void\_or\_integer);

}

Finally, we create a new symbol for the function. Every symbol defined in this function will be stored in the new symbol table, or in one of its sub tables.

SymbolTable.CurrentTable =

new SymbolTable(SymbolTable.CurrentTable, Scope.Function);

}

The CheckFunctionDefinition method checks that the definition is correct, especially in case of old-style definition. In the old-style function definition, the parameter list must match the declaration list.

public static void CheckFunctionDefinition() {

Type funcType = SymbolTable.CurrentFunction.Type;

if (funcType.Style == Type.FunctionStyle.Old) {

List<string> nameList = funcType.NameList;

IDictionary<string,Symbol> entryMap =

SymbolTable.CurrentTable.EntryMap;

In the old-style case, we must make sure the number of parameters equals the number of declarations.

Assert.Error(nameList.Count == entryMap.Count,

SymbolTable.CurrentFunction.Name, Message.

Unmatched\_number\_of\_parameters\_in\_old\_\_style\_function\_definition);

Then we iterate through the parameter list to make sure they are all declared, and to assigned them the correct offset. When the parameters were declared they were given offsets. However, since the parameters and declaration may come in different order, we need to reassign offsets to match the parameter order.

int offset = SymbolTable.FunctionHeaderSize;

foreach (string name in nameList) {

Symbol symbol;

If one of the parameters has not been declared, we report an error.

if (!entryMap.TryGetValue(name, out symbol)) {

Assert.Error(name, Message.

Undefined\_parameter\_in\_old\_\_style\_function\_definition);

}

symbol.Offset = offset;

offset += symbol.Type.Size();

}

}

In case of a new-style function definition, we need to make sure the entry map of the current function is empty. If it is not empty, we have a mix of old-style and new-style function definition, which is not allowed.

else {

Assert.Error(SymbolTable.CurrentTable.EntryMap.Count == 0,

Message.New\_and\_old\_style\_mixed\_function\_definition);

Then we iterate through the parameters list and add each parameter to the current symbol table of the function. In this way, each parameter is located at a memory offset so they can be assigned in function calls and be treated like regular variables inside the function. The AddSymbol method will also report an error if two parameters have the same name.

foreach (Symbol symbol in funcType.ParameterList) {

SymbolTable.CurrentTable.AddSymbol(symbol);

}

}

When the parameter list has been settled, we check if the function is the main function of the C code. In that case, some special rules apply.

if (SymbolTable.CurrentFunction.UniqueName.Equals("main")) {

First, we call the InitializationCodeList method that adds a few lines of code that initializes the system and will be placed before the code of the main function.

AssemblyCodeGenerator.InitializationCodeList();

We extract the parameter type list from the current function. In case of old-style definition, this list will be null. In case of new-style definition, there is only two allowed parameter list:

* No parameters, marked with void.

int main(void) { /\* ... \*/ }

* Command line arguments. The two cases are actually the same case, since arrays are changed to pointers in parameter types. Note that we do not check the names of the parameters, the parameters may have other names. However, the names cannot be omitted.

int main(int argc, char \*argv[]) { /\* ... \*/ }

or

int main(int argc, char \*\*argv) { /\* ... \*/ }

MiddleCodeGenerator.cs

List<Type> typeList =

SymbolTable.CurrentFunction.Type.TypeList;

If typelist is not null, and there are two parameters, we check if the command line argument case applies. The first parameter shall be a signed or unsigned integer while the second parameter shall be a pointer to pointer to signed or unsigned character.

if ((typeList != null) && (typeList.Count == 2)) {

Assert.Error(typeList[0].IsInteger() &&

typeList[1].IsPointer() &&

typeList[1].PointerType.IsPointer() &&

typeList[1].PointerType.PointerType.IsChar(),

"main", Message.Invalid\_parameter\_list);

AssemblyCodeGenerator.ArgumentCodeList();

}

If typelist is null, we have the old-style function definition, in which case we do not perform any type checking at all. If typelist is empty, we have the void-marked empty parameter list.

else {

Assert.Error((typeList == null) || (typeList.Count == 0),

"main", Message.Invalid\_parameter\_list);

}

}

}

The FunctionEnd method is called when the statements of the function have been parsed. Its task is to check the return statement at the end of the function and generate a static symbol holding the assembly code of the function.

The first step is to backpatch the next-set of the statement. Each statement has a next-set holding jump instructions. For instance, for or while loops have next-sets holding jump instructions jumping out of the loop. We add a new empty statement and backpatch the next-set to it.

public static void FunctionEnd(Statement statement) {

MiddleCode nextCode =

AddMiddleCode(statement.CodeList, MiddleOperator.Empty);

Backpatch(statement.NextSet, nextCode);

If the return type of the function is void, we need to add a Return instruction at the end of the function.

if (SymbolTable.CurrentFunction.Type.ReturnType.IsVoid()) {

AddMiddleCode(statement.CodeList, MiddleOperator.Return);

if (SymbolTable.CurrentFunction.UniqueName.Equals

(AssemblyCodeGenerator.MainName)) {

AddMiddleCode(statement.CodeList, MiddleOperator.Exit);

}

}

We then add a FunctionEnd middle code instruction. Its only purpose is that the middle code optimizer will report an error if it is possible to reach the EndFunction instruction in a function not returning void.

AddMiddleCode(statement.CodeList, MiddleOperator.FunctionEnd,

SymbolTable.CurrentFunction);

When we finally have added all middle code instructions, we optimize the middle code of the function. The optimizer transforms the middle code in several ways, see Chapter 11.

MiddleCodeOptimizer middleCodeOptimizer =

new MiddleCodeOptimizer(statement.CodeList);

middleCodeOptimizer.Optimize();

When the middle code has been optimized, we generate assembly code of the function, see Chapter 12.

List<AssemblyCode> assemblyCodeList = new List<AssemblyCode>();

AssemblyCodeGenerator.GenerateAssembly(statement.CodeList,

assemblyCodeList);

As mention in the introduction chapter of this book, we generate two kinds of target code. In the first part of the book we generate assembly code text for 64-bit Linux on Intel processors, which is than assembled and linked in traditional manner. In the second part of the book we generate an executable target file for 16-bit Windows. In the case where we need to distinguish between the two cases, we test whether the Start.Linux or Start.Windows static variable is true. In this chapter, we perform the actions for the Linux target system, see Chapter 13 for the Windows target system.

In the Linux case, we need to generate the assembly text from the assembly code list. We also need to generate the extern set; that is, the set of accesses of static variables with external linkage and the calls to all functions with external linkage.

if (Start.Linux) {

List<string> textList = new List<string>();

if (SymbolTable.CurrentFunction.UniqueName.

Equals(AssemblyCodeGenerator.MainName)) {

textList.AddRange(SymbolTable.InitSymbol.TextList);

if (SymbolTable.ArgsSymbol != null) {

textList.AddRange(SymbolTable.ArgsSymbol.TextList);

}

}

else {

textList.Add("section .text");

}

ISet<string> externSet = new HashSet<string>();

AssemblyCodeGenerator.LinuxTextList(assemblyCodeList, textList,

externSet);

When the text list and access set has been generated, we create a static symbol that we add to the global static set. In this way, the assembly code text of the function will be added to final assembly file. The symbols of the global static set is then handled by the Main method of Section 1.3

StaticSymbol staticSymbol =

new StaticSymbolLinux(SymbolTable.CurrentFunction.UniqueName,

textList, externSet);

SymbolTable.StaticSet.Add(staticSymbol);

}

if (Start.Windows) {

// ...

}

Finally, we set the former symbol table, which is the parent table of this function’s table, to be the current table. We also set the current function to null, which strictly speaking in not necessary. However, it is logical to set it to null when the parser does not parse the code of a function.

SymbolTable.CurrentTable = SymbolTable.CurrentTable.ParentTable;

SymbolTable.CurrentFunction = null;

}

### Specifier List

### Structs and Unions

MiddleCodeGenerator.cs

public static void StructUnionHeader(string optionalName, Sort sort) {

if (optionalName != null) {

Type type = new Type(sort);

SymbolTable.CurrentTable.AddTag(optionalName, type);

}

}

The StructUnionSpecifier method adds a struct or union. The sort parameter is either Sort.Struct or Sort.Union. If the optional name is not null, we look up the type (it has been added by StructOrUnionHeader above) and sets its member map, which is given by the entry map of the current symbol table.

public static Type StructUnionSpecifier(string optionalName, Sort sort) {

if (optionalName != null) {

Type type = SymbolTable.CurrentTable.LookupTag(optionalName, sort);

type.MemberMap = SymbolTable.CurrentTable.EntryMap;

type.MemberList = SymbolTable.CurrentTable.EntryList;

return type;

}

If the optional name is null, we create and return a type with the entry map of the current symbol table.

else {

return (new Type(sort, SymbolTable.CurrentTable.EntryMap,

SymbolTable.CurrentTable.EntryList));

}

}

The LookupStructUnion method looks a struct or union. The sort parameter is either Sort.Struct or Sort.Union.

public static Type LookupStructUnion(string name, Sort sort) {

Type type = SymbolTable.CurrentTable.LookupTag(name, sort);

If the struct or union exists, we simply return its type.

if (type != null) {

return type;

}

If the struct or union does not exist, we create a new type and add it to the tag map. However, the type lacks a member map, which means that it is not yet possible to define variables of the type. It is only when the struct or union becomes properly defined, with a member map, this it will be possible to define variables.

else {

type = new Type(sort);

SymbolTable.CurrentTable.AddTag(name, type);

return type;

}

}

### Enumeration

The EnumItem method stores the item in the symbol table and checks the optional initialization symbol. The type of the item is constant signed integer.

MiddleCodeGenerator.cs

public static Symbol EnumItem(string itemName,

Symbol optInitializerSymbol) {

Type itemType = new Type(Sort.SignedInt, true);

itemType.Constant = true;

The value of the item is either set by the initialization symbol value, if present, or by the enumeration value stack. The stack holds the value of the current enumeration list. It is popped and pushed with the item value plus one. In that way will each item hold the value of the previous item plus one, the first item holds the value zero.

BigInteger value;

if (optInitializerSymbol != null) {

Assert.Error(optInitializerSymbol.Type.IsIntegral(), itemName,

Message.Non\_\_integral\_enum\_value);

Assert.Error(optInitializerSymbol.IsValue(), itemName,

Message.Non\_\_constant\_enum\_value);

CCompiler\_Main.Parser.EnumValueStack.Pop();

value = (BigInteger) optInitializerSymbol.Value;

}

else {

value = CCompiler\_Main.Parser.EnumValueStack.Pop();

}

Symbol itemSymbol = new Symbol(itemName, null, itemType, value);

SymbolTable.CurrentTable.AddSymbol(itemSymbol);

CCompiler\_Main.Parser.EnumValueStack.Push(value + 1);

return itemSymbol;

}

The EnumSpecifier method add the enumerator to the current symbol table if optional name is not null.

public static Type EnumSpecifier(string optionalName,

ISet<Pair<Symbol,bool>> enumSet) {

Type enumType = new Type(Sort.Enumeration, enumSet);

if (optionalName != null) {

SymbolTable.CurrentTable.AddTag(optionalName, enumType);

}

return enumType;

}

The LookupEnum method looks up the name of an enumeration. Technically, there is no meaning by looking up an enumeration name since an enumeration value has the signed integer type. However, if an unknown name is refereed, an error shall be reported. The following code is value only if an enumeration with the name CarMake has earlier been defined.

Enum CarMake car;

MiddleCodeGenerator.cs

public static Type LookupEnum(string name) {

Type type = SymbolTable.CurrentTable.LookupTag(name, Sort.Enumeration);

Assert.Error(type != null, name, Message.Tag\_not\_found);

return type;

}

### Declarator

The Declarator method handles the cases without assignment or bitfield. The declarator is possible made up by a sequence of pointer, array, or functions declarations, and we add the type of the specifier as the end type of the declarator.

MiddleCodeGenerator.cs

public static void Declarator(Specifier specifier, Declarator declarator){

declarator.Add(specifier.Type);

A function can only hold extern or static storage. This is in fact the only storage check that is performed outside the Specifier class of Chapter 4, since we need the declarator to decide that it is in fact a function declaration.

Storage storage = specifier.Storage;

if (declarator.Type.IsFunction()) {

Assert.Error((storage == Storage.Static) ||

(storage == Storage.Extern), storage, Message.

Only\_extern\_or\_static\_storage\_allowed\_for\_functions);

}

We create the symbol holding the name, external linkage, storage, and type of the declarator, and add it to the current symbol table.

Symbol symbol = new Symbol(declarator.Name, specifier.ExternalLinkage,

storage, declarator.Type);

SymbolTable.CurrentTable.AddSymbol(symbol);

If the symbol has static storage and is not a function, we add it to the global static set. We call the Value method in the ConstantExpression class to generate the global static set. If it is a function definition, it becomes added by the FunctionEnd method of Section 3.2.1. If it is function declaration, is shall not be stored in the global static set at all. Declarations are stored in the symbol table only, not in the global static set.

if ((storage == Storage.Static) && !type.IsFunction()) {

SymbolTable.StaticSet.Add(ConstantExpression.Value(symbol));

}

}

The InitializedDeclarator method handles the cases where the declarator becomes initialized with a value. Similar to the Declarator case above, we begin by adding the type of the specifier to the declarator. Then we obtain the

public static List<MiddleCode> InitializedDeclarator(Specifier specifier,

Declarator declarator, object initializer){

declarator.Add(specifier.Type);

Type type = declarator.Type;

Storage storage = specifier.Storage;

string name = declarator.Name;

A function cannot be initialized, neither can a symbol of extern or typedef storage, or a member of a struct or union.

Assert.Error(!type.IsFunction(), null,

Message.Functions\_cannot\_be\_initialized);

Assert.Error(storage != Storage.Extern, name,

Message.Extern\_cannot\_be\_initialized);

Assert.Error(storage != Storage.Typedef, name,

Message.Typedef\_cannot\_be\_initialized);

Assert.Error((SymbolTable.CurrentTable.Scope != Scope.Struct) &&

(SymbolTable.CurrentTable.Scope != Scope.Union),

name, Message.Struct\_or\_union\_field\_cannot\_be\_initialized);

In case of static storage, we call the GenerateStatic method in the GenerateStaticInitializer class to generate a middle code list holding the initialization code.

if (storage == Storage.Static) {

List<MiddleCode> middleCodeList =

GenerateStaticInitializer.GenerateStatic(type, initializer);

We create a symbol with the name and type of the declarator and the external linkage and storage of the specifier and add it to the current symbol table.

Symbol symbol = new Symbol(name, specifier.ExternalLinkage,

storage, type);

SymbolTable.CurrentTable.AddSymbol(symbol);

Since the storage is static, we also create a static symbol that we add to the global static set. Similar to the Declarator case above, we call the Value method in the ConstantExpression class to generate the value. However, we return an empty code list after we have added the static symbol to the global static set.

StaticSymbol staticSymbol =

ConstantExpression.Value(symbol.UniqueName, type, middleCodeList);

SymbolTable.StaticSet.Add(staticSymbol);

return (new List<MiddleCode>());

}

In the non-static cases, we call the GenerateAuto method in the GenerateAutoInitializer class, which return a list of middle code instructions. The value in this case is not a static value. Instead, it is alist of middle code instructions that assign the value to the symbol in run-time.

else {

Symbol symbol =

new Symbol(name, specifier.ExternalLinkage, storage, type);

symbol.Offset = SymbolTable.CurrentTable.CurrentOffset;

List<MiddleCode> codeList =

GenerateAutoInitializer.GenerateAuto(symbol, initializer);

SymbolTable.CurrentTable.AddSymbol(symbol);

return codeList;

}

}

The BitfieldDeclarator handles the case with bitfields. What is special in this case is that the declarator can actually be omitted.

public static void BitfieldDeclarator(Specifier specifier,

Declarator declarator, Symbol bitsSymbol) {

Storage storage = specifier.Storage;

Bitfields are only allowed in structs or unions and can only have auto or register storage.

Assert.Error((SymbolTable.CurrentTable.Scope == Scope.Struct) ||

(SymbolTable.CurrentTable.Scope == Scope.Union),bitsSymbol,

Message.Bitfields\_only\_allowed\_in\_structs\_or\_unions);

Assert.Error((storage == Storage.Auto) || (storage == Storage.Register),

null, Message.

Only\_auto\_or\_register\_storage\_allowed\_in\_struct\_or\_union);

We calculate the number of bits by parsing the text of the

object bitsValue = bitsSymbol.Value;

int bits = int.Parse(bitsValue.ToString());

If the declarator is not null (it has been present in the code) we add the specifier type to it and extract its final type, which must be integral. The number of bits must be a value between one and the type size in bites (eight times the type size in bytes), inclusive.

if (declarator != null) {

declarator.Add(specifier.Type);

Type type = declarator.Type;

Assert.Error(type.IsIntegral(), type,

Message.Non\_\_integral\_bits\_expression);

Assert.Error((bits >= 1) && (bits <= (8 \* type.Size())),

bitsValue, Message.Bits\_value\_out\_of\_range);

If the number of bits is less then the type size in bits, we need to define the bit mask that is used to delete the bits not covered by the mask.

if (bits < (8 \* type.Size())) {

type.SetBitfieldMask(bits);

}

Symbol symbol = new Symbol(declarator.Name, specifier.ExternalLinkage,

storage, type);

SymbolTable.CurrentTable.AddSymbol(symbol);

Contrary to the previous cases, we do not add the symbol to the global static set since it cannot be static.

}

If the declarator is null (it has been omitted in the code) we just check that the number of bits is in range: at least one and at most the type size in bits.

else {

Assert.Error((bits >= 1) && (bits <= (8 \* 4)), bitsValue,

Message.Bits\_value\_out\_of\_range);

}

}

### Pointer declarator

The PointerDeclarator method takes a type list and a declarator.

MiddleCodeGenerator.cs

public static Declarator PointerDeclarator(List<Type> typeList,

Declarator declarator) {

If the declarator is null, we simply create a new declarator since we always need a declarator, even in the cases where is has been omitted.

if (declarator == null) {

declarator = new Declarator(null);

}

We iterate through the pointer type list and add each pointer type to the declarator. In this way, the declarator is pointer.

foreach (Type pointerType in typeList) {

Type pointerType = new Type((Type) null);

pointerType.Constant = type.Constant;

pointerType.Volatile = type.Volatile;

declarator.Add(pointerType);

}

We will always have at least a declarator with a certain type or a pointer type, since it is syntactically impossible to omit both the declarator and the pointer list.

return declarator;

}

### Direct Declarator

The **ArrayType** method handles the declaration of an array.

MiddleCodeGenerator.cs

public static Declarator ArrayType(Declarator declarator,

Expression optionalSizeExpression) {

if (declarator == null) {

declarator = new Declarator(null);

}

If the array size expression is present, the parser has already made sure it is a constant expression of integral type. But we must also check that the array (the value of the expression) is positive.

Type arrayType;

if (optionalSizeExpression != null) {

Symbol optSizeSymbol = optionalSizeExpression.Symbol;

int arraySize = (int) ((BigInteger) optSizeSymbol.Value);

Assert.Error(arraySize > 0, arraySize,

Message.Non\_\_positive\_array\_size);

arrayType = new Type(arraySize, null);

}

If the array size expression is not present, we create a new array type of zero size.

else {

arrayType = new Type(0, null);

}

We add the array type to the declarator, which we return. Note that the type of the array is yet unknown, which is why it is null.

declarator.Add(arrayType);

return declarator;

}

The next kind of direct declarator is the new-style function declarator, with parentheses enclosing a variable parameter list. A variable parameter list is a parameter list that may be finished with the variable marker (‘…’).

The NewFunctionDeclaration handles a new-style function declaration.

MiddleCodeGenerator.cs

public static Declarator NewFunctionDeclaration(Declarator declarator,

List<Symbol> parameterList,

bool variadic) {

It is not allowed to add a variable marker to an empty parameter list, it must always be at least one parameter.

if (parameterList.Count == 0) {

Assert.Error(!variadic, "...",

Message.An\_variadic\_function\_must\_have\_at\_least\_one\_parameter);

}

It is allowed to have a void parameter if it is the only parameter in the list, and there is no variable marker.

else if ((parameterList.Count == 1) && parameterList[0].Type.IsVoid()) {

Assert.Error(parameterList[0].Name == null,

parameterList[0].Name,

Message.A\_void\_parameter\_cannot\_be\_named);

Assert.Error(!variadic, "...", Message.

An\_variadic\_function\_cannot\_have\_a\_void\_parameter);

parameterList.Clear();

}

Otherwise, a void parameter is not allowed.

else {

foreach (Symbol symbol in parameterList) {

Assert.Error(!symbol.Type.IsVoid(),

Message.Invalid\_void\_parameter);

}

}

Finally, we create a new-style function type with the parameter list and variadic status, add it to the declarator, and return the declarator. Note that we do not add a return type (the first parameter is null), since it will be added to the declarator by the specifier.

declarator.Add(new Type(null, parameterList, variadic));

return declarator;

}

The OldFunctionDeclaration method handles the old-style function declaration. The only thing we need to test is that the same name does not occur twice in the name list.

MiddleCodeGenerator.cs

public static Declarator OldFunctionDeclaration(Declarator declarator,

List<string> nameList) {

We create a set and add the names of the name list to the set. The Add method returns true as long as the name does not already occur in the set.

ISet<string> nameSet = new HashSet<string>();

foreach (string name in nameList) {

Assert.Error(nameSet.Add(name), name, Message.Name\_already\_defined);

}

Finally, we create an old-style function type with the name list, add it to the declarator, and return the declarator. Again, note that we do not add a return type since it will be added to the declarator by the specifier.

declarator.Add(new Type(null, nameList));

return declarator;

}

### Parameters

MiddleCodeGenerator.cs

public static Symbol Parameter(Specifier specifier,

Declarator declarator) {

Storage storage = specifier.Storage;

Type specifierType = specifier.Type;

string name;

Type type;

if (declarator != null) {

name = declarator.Name;

declarator.Add(specifierType);

type = declarator.Type;

}

else {

name = null;

type = specifierType;

}

If the parameter is an array, we set the type to a pointer to the array type, and if it is a function, we set the type to a pointer to the function. In both cases the type becomes constant.

if (type.IsArray()) {

type = new Type(type.ArrayType);

type.Constant = true;

}

else if (type.IsFunction()) {

type = new Type(type);

type.Constant = true;

}

We create and return a symbol, with the name, storage, and type of the parameter. The second value is false since parameters do not have external linkage. The last value is true since it is a parameter, which we need to know that in the assembly code generation phase in Chapter 12.

return (new Symbol(name, false, storage, type, true));

}

### Abstract Declarator

## Statements

### The If Statement

The IfStatement method handles the if statement of the parser. Since the if-statement accepts logical expressions only, we need to start with making sure the expression is or becomes a logical expression.

MiddleCodeGenerator.cs

public static Statement IfStatement(Expression expression,

Statement innerStatement){

expression = TypeCast.ToLogical(expression);

Since we use the value of the expression in the if-statement, we only use the long list.

List<MiddleCode> codeList = expression.LongList;

AddMiddleCode(codeList, MiddleOperator.CheckTrackMapFloatStack);

We backpatch the true value of the expression to the first instruction in the middle code list of the inner statement. This means that when the expression is true, the program will jump to the beginning of the code list of the inner statement.

Backpatch(expression.Symbol.TrueSet, innerStatement.CodeList);

codeList.AddRange(innerStatement.CodeList);

We add a jump instruction at the end of the middle code that shall jump to the instruction following this if-statement.

MiddleCode nextCode = AddMiddleCode(codeList, MiddleOperator.Jump);

We define the next-set of the if-statement as the union of the next-set of the inner statement, the false-set of the expression, and the jump instruction added to the code list of the inner statement.

ISet<MiddleCode> nextSet = new HashSet<MiddleCode>();

nextSet.UnionWith(innerStatement.NextSet);

nextSet.UnionWith(expression.Symbol.FalseSet);

nextSet.Add(nextCode);

Finally, we return on object of the Statement class holding the code list and next-set of the if-statement.

return (new Statement(codeList, nextSet));

}

### The If-Else Statement

The IfElseStatement method is similar to IfStatement above. The difference is that we have two inner statements, that shall be executed in case of a true or false expression.

public static Statement IfElseStatement(Expression expression,

Statement trueStatement,

Statement falseStatement) {

expression = TypeCast.ToLogical(expression);

List<MiddleCode> codeList = expression.LongList;

AddMiddleCode(codeList, MiddleOperator.CheckTrackMapFloatStack);

We backpatch the true-set and false-set to the beginning of the code list of the true and false statement, respectively.

Backpatch(expression.Symbol.TrueSet, trueStatement.CodeList);

Backpatch(expression.Symbol.FalseSet, falseStatement.CodeList);

We add the code list for the true and false statements. We also add a jump instruction after each list, to jump out of the statement. The jump instruction are then added to the next-set of the else-if-statement.

codeList.AddRange(trueStatement.CodeList);

MiddleCode trueNextCode = AddMiddleCode(codeList, MiddleOperator.Jump);

codeList.AddRange(falseStatement.CodeList);

MiddleCode falseGotoCode = AddMiddleCode(codeList, MiddleOperator.Jump);

The next-set of the if-else-statement is the union of the next-sets of the true and false statement, and their extra jump instructions.

ISet<MiddleCode> nextSet = new HashSet<MiddleCode>();

nextSet.UnionWith(trueStatement.NextSet);

nextSet.UnionWith(falseStatement.NextSet);

nextSet.Add(trueNextCode);

nextSet.Add(falseGotoCode);

Finally, we return on object of the Statement class holding the code list and next-set of the if-statement.

return (new Statement(codeList, nextSet));

}

### The Switch Statement

We need a map to keep track of the case statements. Each entry of the map holds the value of the case statement, and the first instruction of its middle code list (if the list is empty, we add an empty instruction). Since the switch statements can be nested, we store the maps for each switch statement in the m\_caseMapStack map.

private static Stack<IDictionary<BigInteger, MiddleCode>> m\_caseMapStack =

new Stack<IDictionary<BigInteger, MiddleCode>>();

In the same way, we need a stack to keep track of the default statement of the switch statement. For each default statement connected to a switch statement, we store the first instruction of the default middle code list in the m\_defaultStack. If there is no default statement, we add null to the stack.

private static Stack<MiddleCode> m\_defaultStack = new Stack<MiddleCode>();

Finally, we also need the m\_breakSetStack stack for the break statements connected to the switch statement. We add the first instruction of the break statement to the sets of the stack. They are later backpatched to jump out of the switch statement.

private static Stack<ISet<MiddleCode>> m\_breakSetStack =

new Stack<ISet<MiddleCode>>();

The SwitchHeader method is called before the parsing of the switch statement. It pushes an empty map to the case map stack, a null reference to the default stack, and an empty set to the break set stack.

public static void SwitchHeader() {

m\_caseMapStack.Push(new Dictionary<BigInteger,MiddleCode>());

DefaultStack.Push(null);

BreakSetStack.Push(new HashSet<MiddleCode>());

}

The SwitchStatement method is called after the parsing of the switch statement. Its task is to generate middle code for the case and default statements, and to backpatch the break statements.

public static Statement SwitchStatement(Expression switchExpression,

Statement innerStatement) {

Since the switch statement requires an integral value rather than a logical value we need to, contrary to the if-statement above, type cast a potential logical value to an integral value.

switchExpression = TypeCast.LogicalToIntegral(switchExpression);

Since we need the value of the switch statement, rather than its side-effects, we are only interested in the long list of the switch expression.

List<MiddleCode> codeList = switchExpression.LongList;

Each case value is type casted to the type of the switch expression.

Type switchType = switchExpression.Symbol.Type;

foreach (KeyValuePair<BigInteger,MiddleCode> entry

in m\_caseMapStack.Pop()) {

BigInteger caseValue = entry.Key;

MiddleCode caseTarget = entry.Value;

Symbol caseSymbol = new Symbol(switchType, caseValue);

We add code where we compare the case value with the switch value and jump to the matching middle code instruction if the values are equal. We use the Case instruction rather than Equal, since we want the value to be kept in a register when we generate the assembly code in Chapter 12.

AddMiddleCode(codeList, MiddleOperator.Case, caseTarget,

switchExpression.Symbol, caseSymbol);

}

After the case values, we add a case end instruction. The reason for this is that the register used for the case comparison in the generated assembly code shall be unallocated.

AddMiddleCode(codeList, MiddleOperator.CaseEnd,

switchExpression.Symbol);

If there is a default statement, we jump to it if the switch statement does not match any of the case statement. In that case, the next-set becomes empty.

MiddleCode defaultCode = DefaultStack.Pop();

if (defaultCode != null) {

AddMiddleCode(codeList, MiddleOperator.Jump, defaultCode);

}

Similar to the if-statement and if-else-statement, the switch-statement has a next-set; that is, a set of middle code instruction that shall jump out of the switch statement.

ISet<MiddleCode> nextSet = new HashSet<MiddleCode>();

If there is no default statement, the execution shall jump out of the switch statement after the last case statement, and we add a jump instruction to the next-set.

if (defaultCode == null) {

nextSet.Add(AddMiddleCode(codeList, MiddleOperator.Jump));

}

We add the code list of the inner statement to the resulting code list.

codeList.AddRange(innerStatement.CodeList);

Finally, we add the next-set of the inner statement and the set of break statements to the next-set of the switch statement.

nextSet.UnionWith(innerStatement.NextSet);

nextSet.UnionWith(BreakSetStack.Pop());

return (new Statement(codeList, nextSet));

}

### The Case Statement

For the case statement the Main.CaseMapStack is not allowed to be empty. If it is empty, the case statement misses an enclosed switch statement, and we report an error.

public static Statement CaseStatement(Expression expression,

Statement statement) {

Assert.Error(m\_caseMapStack.Count > 0, Message.Case\_without\_switch);

expression = TypeCast.LogicalToIntegral(expression);

Moreover, the value of the case expression must be integral and constant, and thereby possible to evaluate in compile-time. If the value of the symbol is null, we have a non-constant case value, and we report an error.

Assert.Error(expression.Symbol.Value != null, expression.Symbol.Name,

Message.Non\_\_constant\_case\_value);

BigInteger caseValue = (BigInteger) expression.Symbol.Value;

We extract the current case map from the top of the case map stack and adds the current case value to the map.

IDictionary<BigInteger, MiddleCode> caseMap = m\_caseMapStack.Peek();

If the map already contains the value, we have two case statements with the same value and we report an error.

Assert.Error(!caseMap.ContainsKey(caseValue), caseValue,

Message.Repeated\_case\_value);

Finally, we add the case value and the first instruction of the statement’s code to the case map of the enclosed switch statements

caseMap.Add(caseValue, GetFirst(statement.CodeList));

return statement;

}

### The Default Statement

If the default stack is empty, we have a default statement without an enclosing switch statement and we report an error.

public static Statement DefaultStatement(Statement statement) {

Assert.Error(DefaultStack.Count > 0, Message.Default\_without\_switch);

There can be only one default statement in a switch statement. Therefore, if the value on the top of the default stack is not null, we have a switch statement with two default statement and we report an error.

Assert.Error(DefaultStack.Pop() == null, Message.Repeted\_default);

Finally, we add the default middle code instruction at the top of the stack.

DefaultStack.Push(GetFirst(statement.CodeList));

return statement;

}

### The While Statement

Similar to the break set stack above, we also need a continue set stack.

private static Stack<ISet<MiddleCode>> m\_continueSetStack =

new Stack<ISet<MiddleCode>>();

The LoopHeader method is called before the parsing of a while, do, or for statement. Its task is to add empty sets at the top of the break set stack and continue set stack.

public static void LoopHeader() {

m\_breakSetStack.Push(new HashSet<MiddleCode>());

m\_continueSetStack.Push(new HashSet<MiddleCode>());

}

The WhileStatement method is called after the while statement has been parsed. Similar to the if statement above, it starts by type casting the expression into logical type.

public static Statement WhileStatement(Expression expression,

Statement innerStatement) {

expression = TypeCast.ToLogical(expression);

List<MiddleCode> codeList = expression.LongList;

AddMiddleCode(codeList, MiddleOperator.CheckTrackMapFloatStack);

We backpatch the true-set of the expression to the code list of the inner statement. If the expression is true, the execution shall jump to the beginning of the inner statement.

Backpatch(expression.Symbol.TrueSet, innerStatement.CodeList);

codeList.AddRange(innerStatement.CodeList);

We define the next-set of the while statement, and add a jump instruction that jumps back to the beginning of the while statement. Since we jump backwards, we do not need to perform backpatching.

ISet<MiddleCode> nextSet = new HashSet<MiddleCode>();

nextSet.Add(AddMiddleCode(codeList, MiddleOperator.Jump,

GetFirst(codeList)));

We also add the false-set of the expression and the break set of the inner statement to the next-set.

nextSet.UnionWith(expression.Symbol.FalseSet);

nextSet.UnionWith(m\_breakSetStack.Pop());

We backpatch the next-set of the inner statement and its continue set to the code list. The jump instruction of the sets shall jump back to the beginning of the while statement.

Backpatch(innerStatement.NextSet, codeList);

Backpatch(m\_continueSetStack.Pop(), codeList);

return (new Statement(codeList, nextSet));

}

### The Do Statement

The do statement starts with a statement followed by an expression. We start by backpatching the next set of the inner statement to the beginning of the middle code of the do statement.

public static Statement DoStatement(Statement innerStatement,

Expression expression) {

List<MiddleCode> codeList = innerStatement.CodeList;

Backpatch(innerStatement.NextSet, codeList);

codeList.AddRange(expression.LongList);

We backpatch the true-set of the expression and the continue set of the inner statement to the beginning of the do statement.

Backpatch(expression.Symbol.TrueSet, codeList);

Backpatch(m\_continueSetStack.Pop(), codeList);

The next-set of the do statement is the union of the false-set of the expression, the break-set of the inner statement, and the jump instruction out of the inner statement.

ISet<MiddleCode> nextSet = new HashSet<MiddleCode>();

nextSet.UnionWith(expression.Symbol.FalseSet);

nextSet.UnionWith(m\_breakSetStack.Pop());

return (new Statement(codeList, nextSet));

}

### The For Statement

In the ForStatement method, we make difference of the short and long list of the expressions. If the test expression is present, we want its value, and therefore we are interested in its long list. On the other hand, if the initializer or next expressions are present, we are only interested in their potential side effects, which are stored in their short lists.

public static Statement ForStatement(Expression initializerExpression,

Expression testExpression, Expression nextExpression,

Statement innerStatement) {

List<MiddleCode> codeList = new List<MiddleCode>();

ISet<MiddleCode> nextSet = new HashSet<MiddleCode>();

We add an empty instruction as the target of the test expression, since we do not know if their in fact is a test expression.

MiddleCode testTarget = AddMiddleCode(codeList, MiddleOperator.Empty);

If the initializer expression is not null, we add its short list. Note that we add the short list instead of the long list since we are only interested of the side effects of the expression, not its value. We also backpatch both its true-set and false-set to the test target; that is, the beginning of the test expression code.

if (initializerExpression != null) {

codeList.AddRange(initializerExpression.ShortList);

Backpatch(initializerExpression.Symbol.TrueSet, testTarget);

Backpatch(initializerExpression.Symbol.FalseSet, testTarget);

}

If the test expression is not null, we start by type casting it into a logical value and add its long list to the code. Note that we add the long list rather than the short list in this case, since we need the value of the expression, rather than its side effects only.

if (testExpression != null) {

testExpression = TypeCast.ToLogical(testExpression);

codeList.AddRange(testExpression.LongList);

AddMiddleCode(codeList, MiddleOperator.CheckTrackMapFloatStack);

We backpatch the true-set of the test expression to the beginning or the code list of the inner statement.

Backpatch(testExpression.Symbol.TrueSet, innerStatement.CodeList);

nextSet.UnionWith(testExpression.Symbol.FalseSet);

}

We add the code list of the inner statement. We then add the next target instruction and backpatch the next-set of the code of the inner statement since we do not know if the next expression is not null.

codeList.AddRange(innerStatement.CodeList);

MiddleCode nextTarget = AddMiddleCode(codeList, MiddleOperator.Empty);

Backpatch(innerStatement.NextSet, nextTarget);

If the next expression is not null, we add its short list and backpatch both its true-set and false-set to the beginning of the test expression.

if (nextExpression != null) {

codeList.AddRange(nextExpression.ShortList);

Backpatch(nextExpression.Symbol.TrueSet, testTarget);

Backpatch(nextExpression.Symbol.FalseSet, testTarget);

}

Finally, like the while case, the continue set is backpatched to the beginning of the test expression and the break set is added to the next-set of the for statement.

AddMiddleCode(codeList, MiddleOperator.Jump, testTarget);

Backpatch(m\_continueSetStack.Pop(), nextTarget);

nextSet.UnionWith(m\_breakSetStack.Pop());

return (new Statement(codeList, nextSet));

}

### The Label and Goto Statements

In C, is is allowed to use goto statements, jumping to labels, even though it is not recommended. For that we need the m\_labelMap and m\_gotoSetMap maps, where m\_labelMap stores the labels of the goto targets and m\_gotoSetMap stores sets of instructions jumping to a target.

public static IDictionary<string, MiddleCode> m\_labelMap =

new Dictionary<string, MiddleCode>();

public static IDictionary<string, ISet<MiddleCode>> m\_gotoSetMap =

new Dictionary<string, ISet<MiddleCode>>();

The LabelStatement add a label to the label map. If the label map already holds a label with the name, we have two labels with the same name, which is not allowed and we report an error.

public static Statement LabelStatement(string labelName,

Statement statement) {

Assert.Error(!m\_labelMap.ContainsKey(labelName),

labelName, Message.Defined\_twice);

m\_labelMap.Add(labelName, GetFirst(statement.CodeList));

return statement;

}

In the GotoStatement method we start by adding middle code holding a goto instruction.

public static Statement GotoStatement(string labelName) {

List<MiddleCode> gotoList = new List<MiddleCode>();

MiddleCode gotoCode = AddMiddleCode(gotoList, MiddleOperator.Jump);

If the label name already exists the goto-set map, we look up the goto-set and add the goto instruction. This situation occurs in when there has been previous jumps to the label.

if (m\_gotoSetMap.ContainsKey(labelName)) {

ISet<MiddleCode> gotoSet = m\_gotoSetMap[labelName];

gotoSet.Add(gotoCode);

}

If the label name is not already a key in the goto-set map, we create a new goto-set, to which we add the goto instruction, and then we add the label name and goto set to the goto set map. This situation occurs in when we encounter the first jump to the label. Note that we do not check whether the label is present in the label map, since it may be a forward jump. The BackpatchGoto method below backpatches all jumps.

else {

ISet<MiddleCode> gotoSet = new HashSet<MiddleCode>();

gotoSet.Add(gotoCode);

m\_gotoSetMap.Add(labelName, gotoSet);

}

return (new Statement(gotoList));

}

The BackpatchGoto method is called at the end of each function definition. It iterates through the goto- set map and, for each label name, looks up the label instruction and backpatch the goto-set to that instruction.

public static void BackpatchGoto() {

foreach (KeyValuePair<string,ISet<MiddleCode>> entry in m\_gotoSetMap) {

string labelName = entry.Key;

ISet<MiddleCode> gotoSet = entry.Value;

If the label name does not exist in the label map, we have a goto statement to an unknown label and we report an error.

MiddleCode labelCode;

Assert.Error(m\_labelMap.TryGetValue(labelName, out labelCode),

labelName, Message.Missing\_goto\_address);

Backpatch(gotoSet, labelCode);

}

}

### Return Statement

The GenerateReturnStatement method generates the code for setting the return value (if present) and return the execution to the calling function. If the function is main function it also adds code for exiting the execution. In that case, the execution shall exit only in case of the original main function. Note thet is allowed to recursively call the main function.

public static Statement ReturnStatement(Expression expression) {

List<MiddleCode> codeList;

If the expression is not null, we need the check that the function does not return void. If it does, we report an error.

if (expression != null) {

Assert.Error(!SymbolTable.CurrentFunction.Type.ReturnType.IsVoid(),

Message.Non\_\_void\_return\_from\_void\_function);

We cast the return expression to the return type of the function.

expression = TypeCast.ImplicitCast(expression,

SymbolTable.CurrentFunction.Type.ReturnType);

codeList = expression.LongList;

AddMiddleCode(codeList, MiddleOperator.SetReturnValue);

AddMiddleCode(codeList, MiddleOperator.Return,

null, expression.Symbol);

}

If the expression is null, we check that the function returns void. If it does not, we report an error.

else {

Assert.Error(SymbolTable.CurrentFunction.Type.ReturnType.IsVoid(),

Message.Void\_returned\_from\_non\_\_void\_function);

codeList = new List<MiddleCode>();

AddMiddleCode(codeList, MiddleOperator.Return);

}

If the function is the main function, we exit the program execution.

if (SymbolTable.CurrentFunction.UniqueName.Equals

(AssemblyCodeGenerator.MainName)) {

AddMiddleCode(codeList, MiddleOperator.Exit);

}

return (new Statement(codeList));

}

### Optional Expression Statement

If there is an expression, we add its short list to the code list. We choose the short list rather than the long list since we are only interested in the side effects of the expression.

MiddleCodeGenerator.cs

public static Statement ExpressionStatement(Expression expression) {

List<MiddleCode> codeList = new List<MiddleCode>();

if (expression != null) {

codeList.AddRange(expression.ShortList);

}

return (new Statement(codeList));

}

### Jump Register Statements

The JumpRegisterStatement method adds an instruction for jumping to the address stores in a register. It is used by the setjmp function in the standard library.

public static Statement JumpRegisterStatement(Register register) {

List<MiddleCode> codeList = new List<MiddleCode>();

AddMiddleCode(codeList, MiddleOperator.JumpRegister, register);

return (new Statement(codeList));

}

### Interrupt Statements

The InterruptStatement method adds an instruction for performing an interrupt call. It is used by the standard library for the Windows environment.

public static Statement InterruptStatement(Expression expression) {

List<MiddleCode> codeList = new List<MiddleCode>();

AddMiddleCode(codeList, MiddleOperator.Interrupt,

expression.Symbol.Value);

return (new Statement(codeList));

}

### System Call Statements

The SystemCallStatement method adds an instruction for performing a system call. It is used by the standard library for the Linux environment.

public static Statement SystemCallStatement() {

List<MiddleCode> codeList = new List<MiddleCode>();

AddMiddleCode(codeList, MiddleOperator.SysCall);

return (new Statement(codeList));

}

## Expressions

The next section handles the middle code generation of the expressions. Similar to the parsing, we start with the operator of lowest precedence and continue with the operators of higher precedence.

### The Comma Expression

The value of a comma expression is the value of the right expression. The value of the left expression is discarded, we only keep the side effects of the left expression.

public static Expression CommaExpression(Expression leftExpression,

Expression rightExpression) {

List<MiddleCode> shortList = new List<MiddleCode>();

shortList.AddRange(leftExpression.ShortList);

shortList.AddRange(rightExpression.ShortList);

Note that we add the short list of the left expression to the middle code long list, because we are noted in the value of the left expression, only its side effects.

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(leftExpression.ShortList);

longList.AddRange(rightExpression.LongList);

We return the symbol of the right expression and discard the symbol of the left expression.

return (new Expression(rightExpression.Symbol, shortList, longList));

}

### The Assignment Expression

In C, there are both simple and compound assignment. The AssignmentExpression method calls Assignment with different arguments depending on the operator.

public static Expression AssignmentExpression(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

rightExpression =

TypeCast.ImplicitCast(rightExpression, leftExpression.Symbol.Type);

switch (middleOp) {

case MiddleOperator.Assign:

return Assignment(leftExpression, rightExpression, true);

In cases of compound assignment, we call the suitable method to perform the operation, and the Assignment method for the assignment.

case MiddleOperator.BinaryAdd:

case MiddleOperator.BinarySubtract:

return Assignment(leftExpression,

AdditionExpression(middleOp, leftExpression,

rightExpression));

case MiddleOperator.SignedMultiply:

case MiddleOperator.SignedDivide:

case MiddleOperator.SignedModulo:

return Assignment(leftExpression,

MultiplyExpression(middleOp, leftExpression,

rightExpression));

case MiddleOperator.BitwiseAnd:

case MiddleOperator.BitwiseOr:

case MiddleOperator.BitwiseXOr:

return Assignment(leftExpression,

BitwiseExpression(middleOp, leftExpression,

rightExpression));

default: // shift left, shift right

return Assignment(leftExpression,

ShiftExpression(middleOp, leftExpression,

rightExpression));

}

}

The Assignment method performs simple or compound assignment.

public static Expression Assignment(Expression leftExpression,

Expression rightExpression,

bool simpleAssignment = false) {

In cases of system calls, a specific register is assigned a value.

Register? register = leftExpression.Register;

if (register != null) {

Symbol rightSymbol = rightExpression.Symbol;

We check that the register has the same size (in bytes) as the expression it is assigned to. If it is not, we report an error.

Assert.Error(AssemblyCode.SizeOfRegister(register.Value) ==

rightExpression.Symbol.Type.Size(),

Message.Unmatched\_register\_size);

We add the long list of the right expression, since we need its value, to the final middle code list.

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(rightExpression.LongList);

We add the assignment of the value of right expression to the register.

AddMiddleCode(longList, MiddleOperator.AssignRegister,

register, rightExpression.Symbol);

Since this an assignment, the code list holds both the assignment of the register and the side effect of the assignment. Therefore, we add the long list as both the short list and the long list of the result expression.

return (new Expression(rightExpression.Symbol, longList, longList));

}

If the left expression is not a register, we first check that the left expression is assignable. If it is not, we report an error.

else {

Assert.Error(leftExpression.Symbol.Assignable,

leftExpression, Message.Not\_assignable);

List<MiddleCode> longList = new List<MiddleCode>();

In case of a simple assignment we add the long list of the left expression to the code list. In case of a compound assignment, the long list of the left expression has already been added to long list of the right expression. If we added the long list in case of a compound assignment, we would add the same code twice.

if (simpleAssignment) {

longList.AddRange(leftExpression.LongList);

If the left expression is of floating type, we need to pop the floating value stack in case of a simple assignment. In case of a compound assignment, we let the value stay on the stack.

if (leftExpression.Symbol.Type.IsFloating()) {

AddMiddleCode(longList, MiddleOperator.PopFloat);

}

}

We type cast the right expression to the type of the left expression and add its long list to the final code list.

rightExpression = TypeCast.ImplicitCast(rightExpression,

leftExpression.Symbol.Type);

longList.AddRange(rightExpression.LongList);

If the expressions hold floating type, we really do not perform an assignment, we simply top the current value on the floating value stack to the left expression.

if (leftExpression.Symbol.Type.IsFloating()) {

AddMiddleCode(longList, MiddleOperator.TopFloat,

leftExpression.Symbol);

List<MiddleCode> shortList = new List<MiddleCode>();

shortList.AddRange(longList);

In case of floating type, there is a difference between the short list and long list. In the long list, the resulting value of the assignment (which equals the new value of the left expression) shall be used. Therefore, it is preserved on the stack. In the short list, the value is popped from the stack, since the value shall not be used, only the side effect of the assignment.

AddMiddleCode(shortList, MiddleOperator.PopFloat);

return (new Expression(leftExpression.Symbol, shortList, longList));

}

In case of integral type, we add an assign instruction to the final code.

else {

if (leftExpression.Symbol.Type.IsStructOrUnion()) {

AddMiddleCode(longList, MiddleOperator.AssignInit,

leftExpression.Symbol, rightExpression.Symbol);

}

AddMiddleCode(longList, MiddleOperator.Assign,

leftExpression.Symbol, rightExpression.Symbol);

BigInteger? bitFieldMask =

leftExpression.Symbol.Type.GetBitfieldMask();

In case of bitfield, we add an add operation to set the non-relevant bits to zero.

if (bitFieldMask != null) {

Symbol maskSymbol = new Symbol(leftExpression.Symbol.Type,

bitFieldMask);

AddMiddleCode(longList, MiddleOperator.BitwiseAnd,

leftExpression.Symbol, leftExpression.Symbol,

maskSymbol);

}

In the integral case, we set the code list as both the short list and long list in the resulting expression since the assignment is a side effect.

return (new Expression(leftExpression.Symbol, longList, longList));

}

}

}

### The Condition Expression

The condition operator is rather complicated.

public static Expression ConditionalExpression(Expression testExpression,

Expression trueExpression,

Expression falseExpression) {

We type cast the test expression to logical type.

testExpression = TypeCast.ToLogical(testExpression);

If the test expression is constant, we simple return the true expression if the test expression is true and the false expression if it is false.

if (ConstantExpression.IsConstant(testExpression)) {

return ConstantExpression.IsTrue(testExpression)

? falseExpression : trueExpression;

}

If both the true and false expressions hold logical types, we keep their types.

if (trueExpression.Symbol.Type.IsLogical() &&

falseExpression.Symbol.Type.IsLogical()) {

We start by backpatching the true-set and false-set of the test expression to the beginning of the true and false expression’s code list.

Backpatch(testExpression.Symbol.TrueSet, trueExpression.LongList);

Backpatch(testExpression.Symbol.FalseSet, falseExpression.LongList);

The resulting true-set is the union of the true-sets of the true and false expression, and the resulting false-set is the union of the false-sets of the true and false expression.

ISet<MiddleCode> trueSet = new HashSet<MiddleCode>(),

falseSet = new HashSet<MiddleCode>();

trueSet.UnionWith(trueExpression.Symbol.TrueSet);

trueSet.UnionWith(falseExpression.Symbol.TrueSet);

falseSet.UnionWith(trueExpression.Symbol.FalseSet);

falseSet.UnionWith(falseExpression.Symbol.FalseSet);

List<MiddleCode> shortList = new List<MiddleCode>();

If the short lists of both the true and false expression is empty, it does not matter if the test expression is true or false, and we let the resulting short list be the short list of the true expression (which may be empty).

if (IsEmpty(trueExpression.ShortList) &&

IsEmpty(falseExpression.ShortList)) {

shortList.AddRange(testExpression.ShortList);

}

If the short list of the test expression is not empty, the situation becomes a bit more complicated. We add the long list, rather than the short list, of the test expression to the final short list since we need the value of the test expression in order to jump to the beginning of the short list of either the true of false expression.

else {

shortList.AddRange(testExpression.LongList);

shortList.AddRange(trueExpression.ShortList);

shortList.AddRange(falseExpression.ShortList);

}

We add the long list of the test, true, and false expression to the final long list.

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(testExpression.LongList);

longList.AddRange(trueExpression.LongList);

longList.AddRange(falseExpression.LongList);

Finally, we create a new symbol with logical type and the resulting true-sets and false-sets.

Symbol symbol = new Symbol(trueSet, falseSet);

return (new Expression(symbol, shortList, longList));

}

If at least one of the true or false expression does not hold logical type, we define maxType as their largest type, and type cast both expressions to that type.

else {

Type maxType = TypeCast.MaxType(trueExpression.Symbol.Type,

falseExpression.Symbol.Type);

trueExpression = TypeCast.ImplicitCast(trueExpression, maxType);

Backpatch(testExpression.Symbol.TrueSet, trueExpression.LongList);

We create a new temporary symbol to hold the result of the condition expression.

Symbol symbol = new Symbol(maxType);

In case of non-floating type, we add the assignment instruction. In case of a floating type, the value is already placed at the floating value stack and we do not need to do anything.

if (maxType.IsFloating()) {

AddMiddleCode(trueExpression.LongList,

MiddleOperator.DecreaseStack);

}

else { // XXX

if (trueExpression.Symbol.IsTemporary()) {

foreach (MiddleCode middleCode in trueExpression.LongList) {

if (middleCode[0] == trueExpression.Symbol) {

middleCode[0] = symbol;

}

}

}

else {

AddMiddleCode(trueExpression.LongList, MiddleOperator.Assign,

symbol, trueExpression.Symbol);

}

}

We add the jump to a target code to both the short and long list of the true expression. The target code will be added after the code of the false expression. Its purpose is to jump over the false expression code.

MiddleCode targetCode = new MiddleCode(MiddleOperator.Empty);

AddMiddleCode(trueExpression.ShortList,

MiddleOperator.Jump, targetCode);

AddMiddleCode(trueExpression.LongList,

MiddleOperator.Jump, targetCode);

Similar to the true expression, we type cast and false expression, and backpatch the true-set of the test expression to the beginning of the false expression code.

falseExpression = TypeCast.ImplicitCast(falseExpression, maxType);

Backpatch(testExpression.Symbol.FalseSet, falseExpression.LongList);

We also assign the value of the false expression to the temporary symbol if it does not hold floating type. If it holds floating type, the value is already placed at the top of the floating value stack, and we do nothing.

if (!maxType.IsFloating()) {

if (falseExpression.Symbol.IsTemporary()) {

foreach (MiddleCode middleCode in falseExpression.LongList) {

if (middleCode[0] == falseExpression.Symbol) {

middleCode[0] = symbol;

}

}

}

else {

AddMiddleCode(falseExpression.LongList, MiddleOperator.Assign,

symbol, falseExpression.Symbol);

}

}

If both the short lists of the true and false expressions are empty, we just add the short list of the test expression (which may be empty).

List<MiddleCode> shortList = new List<MiddleCode>();

if (IsEmpty(trueExpression.ShortList) &&

IsEmpty(falseExpression.ShortList)) {

shortList.AddRange(testExpression.ShortList);

}

If not both the short lists of the true and false expressions are empty, we add the long list of the test expression, rather than the short list, since we need to value of the test expression, as well as the short lists of the true and false expressions.

else {

shortList.AddRange(testExpression.LongList);

shortList.AddRange(trueExpression.ShortList);

shortList.AddRange(falseExpression.ShortList);

shortList.Add(targetCode);

}

Finally, add the long lists and return the expression.

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(testExpression.LongList);

longList.AddRange(trueExpression.LongList);

longList.AddRange(falseExpression.LongList);

longList.Add(targetCode);

return (new Expression(symbol, shortList, longList));

}

}

### Constant Expression

The ConstantIntegralExpression methods makes sure the expression is indeed constant and either integral or pointer.

public static Expression ConstantIntegralExpression(Expression expression)

{ expression = ConstantExpression.Cast(expression,

Type.SignedLongIntegerType);

Assert.Error(expression != null, expression,

Message.Non\_\_constant\_expression);

Assert.Error(expression.Symbol.Type.IsIntegralOrPointer(),

expression.Symbol, Message.Non\_\_integral\_expression);

return expression;

}

### The Logical Or Expression

The LogicalOrExpression method backpatches the true-set of the first operand to the beginning of the right operand. This can be interpreted as if the left expression is true, we try the right expression thereafter. If both the expressions are true, the result is true. However, if the left expression is false, the result is determined to be false without evaluating the right expression.

MiddleCodeGenerator.cs

public static Expression LogicalOrExpression(Expression leftExpression,

Expression rightExpression) {

We check if the expression is constant. If it is constant, we return the constant expression.

Expression constantExpression =

ConstantExpression.Logical(MiddleOperator.LogicalOr,

leftExpression, rightExpression);

if (constantExpression != null) {

return constantExpression;

}

We type cast both the expressions to the logical type.

leftExpression = TypeCast.ToLogical(leftExpression);

rightExpression = TypeCast.ToLogical(rightExpression);

For the resulting expression to be true, it is enough that one of the left or right expression is true. Therefore, the true-set of the resulting expression is the union of the true-set of the left and right expression. If the left expression is evaluated to true, the right expression (including its side effects) shall not be evaluated. The false-sets, on the other hand, are different. If the left expression is evaluated to false, we need to evaluate the right expression. Therefore, we backpatch the false-set of the left expression to the beginning of the right expression code. The false-set of the resulting expression is the false-set of the right expression.

ISet<MiddleCode> trueSet = new HashSet<MiddleCode>();

trueSet.UnionWith(leftExpression.Symbol.TrueSet);

trueSet.UnionWith(rightExpression.Symbol.TrueSet);

Backpatch(leftExpression.Symbol.FalseSet, rightExpression.LongList);

Symbol symbol = new Symbol(trueSet, rightExpression.Symbol.FalseSet);

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(leftExpression.LongList);

longList.AddRange(rightExpression.LongList);

List<MiddleCode> shortList = new List<MiddleCode>();

shortList.AddRange(leftExpression.ShortList);

shortList.AddRange(rightExpression.ShortList);

return (new Expression(symbol, shortList, longList));

}

### The Logical And Expression

The LogicalAndExpression method works in the opposite way compared to the logical or expression case above. If the left expression is true, we shall continue to evaluate the false expression. The resulting expression is true only if both the left and right expression is true. However, if the left expression is false the resulting expression is false, without evaluation of the right expression. Therefore, we backpatch the false-set of the first operand to the beginning of the right operand. The true-set of the resulting expression is the true-set of the right expression, and the false set ois the union of the false sets of the left and right expression.

public static Expression LogicalAndExpression(Expression leftExpression,

Expression rightExpression) {

Expression constantExpression =

ConstantExpression.Logical(MiddleOperator.LogicalAnd,

leftExpression, rightExpression);

if (constantExpression != null) {

return constantExpression;

}

leftExpression = TypeCast.ToLogical(leftExpression);

rightExpression = TypeCast.ToLogical(rightExpression);

ISet<MiddleCode> falseSet = new HashSet<MiddleCode>();

falseSet.UnionWith(leftExpression.Symbol.FalseSet);

falseSet.UnionWith(rightExpression.Symbol.FalseSet);

Backpatch(leftExpression.Symbol.TrueSet, rightExpression.LongList);

Symbol symbol = new Symbol(rightExpression.Symbol.TrueSet, falseSet);

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(leftExpression.LongList);

longList.AddRange(rightExpression.LongList);

List<MiddleCode> shortList = new List<MiddleCode>();

shortList.AddRange(leftExpression.ShortList);

shortList.AddRange(rightExpression.ShortList);

return (new Expression(symbol, shortList, longList));

}

### Bitwise Expressions

There are three bitwise operators: inclusive or, exclusive or (xor), and and.

public static Expression BitwiseExpression(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

If the expression can be evaluated to a constant value, we return the constant expression.

Expression constantExpression = ConstantExpression.

Arithmetic(middleOp, leftExpression, rightExpression);

if (constantExpression != null) {

return constantExpression;

}

We find the maximal type of the left and right expression.

Type maxType = TypeCast.MaxType(leftExpression.Symbol.Type,

rightExpression.Symbol.Type);

Symbol resultSymbol = new Symbol(maxType);

The type of the left and right expression must be integral or pointer. Array, string, and functions are converted to pointers.

Assert.Error(maxType.IsIntegralPointerArrayStringOrFunction(),

maxType, Message.Invalid\_type\_in\_bitwise\_expression);

We type cast both expressions into the maximal type.

leftExpression = TypeCast.ImplicitCast(leftExpression, maxType);

rightExpression = TypeCast.ImplicitCast(rightExpression, maxType);

The short list of the resulting expression is simply the short list of the expressions.

List<MiddleCode> shortList = new List<MiddleCode>();

shortList.AddRange(leftExpression.ShortList);

shortList.AddRange(rightExpression.ShortList);

The long list of the resulting expression is the long list of the expressions, and the bitwise operation.

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(leftExpression.LongList);

longList.AddRange(rightExpression.LongList);

AddMiddleCode(longList, middleOp, resultSymbol,

leftExpression.Symbol, rightExpression.Symbol);

return (new Expression(resultSymbol, shortList, longList));

}

### Shift Expression

In C, there are both left shift and right shift.

public static Expression ShiftExpression(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

First, we check that the left expression holds integral or pointer type. An array, a string, or a function is cast to a pointer.

Assert.Error(leftExpression.Symbol.Type.

IsIntegralPointerArrayStringOrFunction(),

leftExpression, Message.Invalid\_type\_in\_shift\_expression);

If the expression can be evaluated to a constant expression, we return the constant expression.

Expression constantExpression =

ConstantExpression.Arithmetic(middleOp, leftExpression,

rightExpression);

if (constantExpression != null) {

return constantExpression;

}

The right expression is type cast to an unsigned character, which always have a size of one byte.

rightExpression =

TypeCast.ImplicitCast(rightExpression, Type.UnsignedCharType);

The final short list is simple the short lists of the left and right expression.

List<MiddleCode> shortList = new List<MiddleCode>();

shortList.AddRange(leftExpression.ShortList);

shortList.AddRange(rightExpression.ShortList);

The final short list is simple the short lists of the left and right expression.

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(leftExpression.LongList);

longList.AddRange(rightExpression.LongList);

The final long list is the short lists of the left and right expression, and the shift operation.

Symbol resultSymbol = new Symbol(leftExpression.Symbol.Type);

AddMiddleCode(longList, middleOp, resultSymbol,

leftExpression.Symbol, rightExpression.Symbol);

return (new Expression(resultSymbol, shortList, longList));

}

### Equality and Relation Expressions

In C, there are the two equality operators equal and not equal, as well as the relational operators less than, less than or equal, greater than, and greater than or equal.

public static Expression RelationalExpression(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression){

First, we check the types of the expression. Everything except struct or union can be compared.

Assert.Error(!leftExpression.Symbol.Type.IsStructOrUnion(),

leftExpression,

Message.Invalid\_type\_in\_expression);

Assert.Error(!rightExpression.Symbol.Type.IsStructOrUnion(),

rightExpression,

Message.Invalid\_type\_in\_expression);

Expression constantExpression =

ConstantExpression.Relation(middleOp, leftExpression,

rightExpression);

if (constantExpression != null) {

return constantExpression;

}

The find the maximal type of the left and right expression types, and type cast both the expressions to that type.

Type maxType = TypeCast.MaxType(leftExpression.Symbol.Type,

rightExpression.Symbol.Type);

leftExpression = TypeCast.ImplicitCast(leftExpression, maxType);

rightExpression = TypeCast.ImplicitCast(rightExpression, maxType);

The final short list is simply the short lists of the left end right expression.

List<MiddleCode> shortList = new List<MiddleCode>();

shortList.AddRange(leftExpression.ShortList);

shortList.AddRange(rightExpression.ShortList);

If the maximal type is unsigned, we change to operator from signed to unsigned.

if (maxType.IsUnsigned()) {

string name = Enum.GetName(typeof(MiddleOperator), middleOp);

middleOp = (MiddleOperator) Enum.Parse(typeof(MiddleOperator),

name.Replace("Signed", "Unsigned"));

}

The final long list is the long lists of the left end right expression, to begin with.

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(leftExpression.LongList);

longList.AddRange(rightExpression.LongList);

We add two instruction to the long list. The first is an if-goto instruction that we add to the true-set. If the expression is true, this instruction will jump to the target that later will be backpatched into the instruction. In the same way, we add a goto instruction to the false-set. If the expression is false, the instruction will jump to the target later backpatched to the instruction.

ISet<MiddleCode> trueSet = new HashSet<MiddleCode>(),

falseSet = new HashSet<MiddleCode>();

trueSet.Add(AddMiddleCode(longList, middleOp, null,

leftExpression.Symbol,

rightExpression.Symbol));

falseSet.Add(AddMiddleCode(longList, MiddleOperator.Jump));

The final symbol holds logical type with true and false-set.

Symbol symbol = new Symbol(trueSet, falseSet);

return (new Expression(symbol, shortList, longList));

}

### Addition and Subtraction Expression

The addition and subtraction operators are a bit complicated, since it is possible to add and subtract pointers in C.

MiddleCodeGenerator.cs

public static Expression AdditionExpression(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

Type leftType = leftExpression.Symbol.Type,

rightType = rightExpression.Symbol.Type;

Similar to the other cases above, the return the possible constant expression.

Expression constantExpression =

ConstantExpression.Arithmetic(middleOp, leftExpression,

rightExpression);

if (constantExpression != null) {

return constantExpression;

}

In the addition case, we also check if the expression is a static expression, in which case we return the expression.

Expression staticExpression =

StaticExpression.Binary(MiddleOperator.BinarySubtract,

leftExpression, rightExpression);

if (staticExpression != null) {

return staticExpression;

}

The final short list is simple the short lists of the left and right expression.

List<MiddleCode> shortList = new List<MiddleCode>();

shortList.AddRange(leftExpression.ShortList);

shortList.AddRange(rightExpression.ShortList);

If at least on the expressions has pointer or array type, we call the PointerArithmetic methods that handles pointer arithmetic.

if (leftType.IsPointerOrArray()) {

return PointerArithmetics(middleOp, leftExpression, rightExpression);

}

In the right expression has pointer type, the operator must be addition, not subtraction. Note that we swap the expressions in the call to PointerArithmetic, so that the pointer is the first expression. XXX

else if (rightType.IsPointerOrArray()) {

Assert.Error(middleOp == MiddleOperator.BinaryAdd, middleOp,

Message.Invalid\_types\_in\_subtraction\_expression);

return PointerArithmetics(middleOp, rightExpression, leftExpression);

}

If none of the left or right expressions are pointers, we perform an arithmetic operation. First, we make sure that both expressions have arithmetic types (integral or floating).

else {

Assert.Error(leftExpression.Symbol.Type.IsArithmetic(),

leftExpression, Message.Non\_\_arithmetic\_expression);

Assert.Error(rightExpression.Symbol.Type.IsArithmetic(),

rightExpression, Message.Non\_\_arithmetic\_expression);

The we find the maximal type and cast both expressions to the type. The resulting expression does also have the maximal type.

Type maxType = TypeCast.MaxType(leftType, rightType);

leftExpression = TypeCast.ImplicitCast(leftExpression, maxType);

rightExpression = TypeCast.ImplicitCast(rightExpression, maxType);

Symbol resultSymbol = new Symbol(maxType);

The final long list is the long lists of the two expressions, and the operation (addition or subtraction).

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(leftExpression.LongList);

longList.AddRange(rightExpression.LongList);

AddMiddleCode(longList, middleOp, resultSymbol,

leftExpression.Symbol, rightExpression.Symbol);

return (new Expression(resultSymbol, shortList, longList));

}

}

The PointerArithmetic methods performs pointer arithmetic. We have two cases: addition of a pointer and an integral value, and subtraction of two pointer values.

private static Expression PointerArithmetic(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

List<MiddleCode> shortList = new List<MiddleCode>();

shortList.AddRange(leftExpression.LongList);

shortList.AddRange(rightExpression.LongList);

Type leftType = leftExpression.Symbol.Type,

rightType = rightExpression.Symbol.Type;

If the left type is pointer and the right type is integral, we add an integer values to and pointer. What makes this different compared to regular addition is that we need take the size of the pointer type into consideration.

if (leftType.IsPointerOrArray() && rightType.IsIntegral()) {

We make sure that the left pointer type is not void, since pointer arithmetic is not allowed on void pointers.

Assert.Error(!leftType.PointerOrArrayType.IsVoid(),

leftExpression, Message.Pointer\_to\_void);

We type cast the integral right expression to the left pointer type for the arithmetic to be performed on values of the same size.

rightExpression = TypeCast.ImplicitCast(rightExpression, leftType);

The final long list is the long lists of the left and right expression, and the pointer arithmetic.

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(leftExpression.LongList);

longList.AddRange(rightExpression.LongList);

The result symbol has the same type as the pointer expression.

Symbol resultSymbol =

new Symbol(new Type(leftType.PointerOrArrayType));

The pointer arithmetic is performed in two steps. First, we multiply the integral value with the pointer size, and then we add the product to the

int pointerSize = leftType.PointerOrArrayType.Size();

Symbol multSymbol = new Symbol(Type.PointerTypeX),

sizeSymbol = new Symbol(Type.PointerTypeX,

(BigInteger) pointerSize);

We multiply the integral value with the size of the pointer type and add the product to the pointer value. The multSymbol value be the result of two constant values, in case of a constant integral value, or the integral value times one, in case of a pointer size of one, which is inefficient. However, the Middle Code Optimizer in Chapter 11 will take care of those cases.

AddMiddleCode(longList, MiddleOperator.UnsignedMultiply, multSymbol,

rightExpression.Symbol, sizeSymbol);

AddMiddleCode(longList, middleOp, resultSymbol,

leftExpression.Symbol, multSymbol);

return (new Expression(resultSymbol, shortList, longList));

}

If it is not the case of a pointer and an integral value, then there must be two pointer values. We make sure that none of them pointes at void. Moreover, we also make sure that their pointer types have the same size.

else {

Assert.Error(!leftType.PointerOrArrayType.IsVoid(),

leftExpression, Message.Pointer\_to\_void);

Assert.Error(!rightType.PointerOrArrayType.IsVoid(),

rightExpression, Message.Pointer\_to\_void);

Assert.Error(leftType.PointerOrArrayType.Size() ==

rightType.PointerOrArrayType.Size(),

leftType + " and " + rightType,

Message.Invalid\_expression);

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(leftExpression.LongList);

longList.AddRange(rightExpression.LongList);

Symbol resultSymbol = new Symbol(Type.VoidPointerType);

int pointerSize = rightType.PointerOrArrayType.Size();

Symbol additionSymbol = new Symbol(Type.IntegerPointerType),

sizeSymbol = new Symbol(Type.IntegerPointerType,

(BigInteger) pointerSize);

We first subtract the pointer values, and then divide the difference by the size of the pointer types. There may be a division by one, in case the pointer types have size one, which is inefficient. However, the Middle Code Optimizer in Chapter 11 will take care of the case.

AddMiddleCode(longList, middleOp, additionSymbol,

leftExpression.Symbol, rightExpression.Symbol);

AddMiddleCode(longList, MiddleOperator.UnsignedDivide, resultSymbol,

additionSymbol, sizeSymbol);

Expression resultExpression = new Expression(resultSymbol,

shortList, longList);

Finally, we type cast the result of the division into signed integer.

return TypeCast.ImplicitCast(resultExpression,

Type.SignedIntegerType);

}

}

### Multiplication Expressions // XXX

public static Expression MultiplyExpression(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

List<MiddleCode> constLongList = new List<MiddleCode>();

constLongList.AddRange(leftExpression.LongList);

constLongList.AddRange(rightExpression.LongList);

Expression constantExpression =

ConstantExpression.Arithmetic(middleOp, leftExpression,

rightExpression);

if (constantExpression != null) {

return constantExpression;

}

Type maxType = TypeCast.MaxType(leftExpression.Symbol.Type,

rightExpression.Symbol.Type);

if ((middleOp == MiddleOperator.SignedModulo) ||

(middleOp == MiddleOperator.UnsignedModulo)) {

Assert.Error(maxType.IsIntegralPointerArrayStringOrFunction(),

maxType, Message.Invalid\_type\_in\_expression);

}

else {

Assert.Error(maxType.IsArithmeticPointerArrayStringOrFunction(),

maxType, Message.Invalid\_type\_in\_expression);

}

leftExpression = TypeCast.ImplicitCast(leftExpression, maxType);

rightExpression = TypeCast.ImplicitCast(rightExpression, maxType);

Symbol resultSymbol = new Symbol(maxType);

if (maxType.IsUnsigned()) {

string name = Enum.GetName(typeof(MiddleOperator), middleOp);

middleOp = (MiddleOperator) Enum.Parse(typeof(MiddleOperator),

name.Replace("Signed", "Unsigned"));

}

List<MiddleCode> shortList = new List<MiddleCode>(),

longList = new List<MiddleCode>();

shortList.AddRange(leftExpression.ShortList);

shortList.AddRange(rightExpression.ShortList);

longList.AddRange(leftExpression.LongList);

longList.AddRange(rightExpression.LongList);

AddMiddleCode(longList, middleOp, resultSymbol,

leftExpression.Symbol, rightExpression.Symbol);

return (new Expression(resultSymbol, shortList, longList));

}

### Cast Expressions

The TypeName method returns the type of the expression. The specifier parameter holds the resulting type of the declaration specifier list.

public static Type TypeName(Specifier specifier, Declarator declarator) {

Type specifierType = specifier.Type;

If the declarator is not null, we add the specifier’s type to the declarator, and returns its type.

if (declarator != null) {

declarator.Add(specifierType);

return declarator.Type;

}

If the declarator is null, we just return the specifier’s type.

else {

return specifierType;

}

}

### Unary Addition Expressions

The unary addition operators are plus and minus. The operand must be a logical or arithmetic value.

public static Expression UnaryExpression(MiddleOperator middleOp,

Expression expression) {

Type type = expression.Symbol.Type;

Assert.Error(type.IsLogical() || type.IsArithmetic (),

expression, Message.Non\_\_arithmetic\_expression);

Expression constantExpression =

ConstantExpression.Arithmetic(middleOp, expression);

if (constantExpression != null) {

return constantExpression;

}

Symbol resultSymbol = new Symbol(expression.Symbol.Type);

AddMiddleCode(expression.LongList, middleOp,

resultSymbol, expression.Symbol);

return (new Expression(resultSymbol, expression.ShortList,

expression.LongList));

}

### Logical Not Expression

The LogicalNotExpression method swap the true-set and false-set of the operand. Note that the method does not actually add any middle code, it just swaps the sets.

public static Expression LogicalNotExpression(Expression expression) {

Expression constantExpression =

ConstantExpression.LogicalNot(expression);

if (constantExpression != null) {

return constantExpression;

}

The resulting expression is the original expression with swapped true and false-sets.

expression = TypeCast.ToLogical(expression);

Symbol notSymbol =

new Symbol(expression.Symbol.FalseSet, expression.Symbol.TrueSet);

return (new Expression(notSymbol, expression.ShortList,

expression.LongList));

}

### Bitwise Not Expression

The BitwiseNotExpression method generates code for the bitwise not unary operator.

public static Expression BitwiseNotExpression(Expression expression) {

expression = TypeCast.LogicalToIntegral(expression);

Assert.Error(expression.Symbol.Type.IsIntegral(),

Message.Only\_integral\_values\_for\_bitwise\_not);

Expression constantExpression =

ConstantExpression.Arithmetic(MiddleOperator.BitwiseNot, expression);

if (constantExpression != null) {

return constantExpression;

}

expression = TypeCast.LogicalToIntegral(expression);

Symbol resultSymbol = new Symbol(expression.Symbol.Type);

AddMiddleCode(expression.LongList, MiddleOperator.BitwiseNot,

resultSymbol, expression.Symbol);

return (new Expression(resultSymbol, expression.ShortList,

expression.LongList));

}

### The sizeof Expression

The first sizeof operator takes a type or an expression as operand.

public static Expression SizeOfExpression(Expression expression) {

We check that the storage of the expression is not register, since it is not allowed.

Assert.Error(!expression.Symbol.IsRegister(), expression,

Message.Register\_storage\_not\_allowed\_in\_sizof\_expression);

Moreover, we check that it is not a function or a bitfield, since they do not have sizes.

Type type = expression.Symbol.Type;

Assert.Error(!type.IsFunction(),

Message.Sizeof\_applied\_to\_function\_not\_allowed);

Assert.Error(!type.IsBitfield(),

Message.Sizeof\_applied\_to\_bitfield\_not\_allowed);

We create a symbol of signed integer types with the size as its value.

Symbol symbol = new Symbol(Type.SignedIntegerType,

(BigInteger) (expression.Symbol.Type.Size()));

symbol.StaticSymbol =

ConstantExpression.Value(symbol.UniqueName, Type.SignedIntegerType,

(BigInteger) (expression.Symbol.Type.Size()));

SymbolTable.StaticSet.Add(symbol.StaticSymbol);

return (new Expression(symbol, new List<MiddleCode>(),

new List<MiddleCode>()));

}

In the case of the type name, we also check that the expression is not a function or a bitfield.

public static Expression SizeOfType(Type type) {

Assert.Error(!type.IsFunction(),

Message.Sizeof\_applied\_to\_function\_not\_allowed);

Assert.Error(!type.IsBitfield(),

Message.Sizeof\_applied\_to\_bitfield\_not\_allowed);

Symbol symbol =

new Symbol(Type.SignedIntegerType, (BigInteger) type.Size());

return (new Expression(symbol, new List<MiddleCode>(),

new List<MiddleCode>()));

}

### Address Expression // XXX

public static Expression AddressExpression(Expression expression) {

Assert.Error(!symbol.IsRegister() && !symbol.Type.IsBitfield(),

expression, Message.Not\_addressable);

Assert.Error(!expression.Symbol.IsRegister(), expression,

Message.Invalid\_address\_of\_register\_storage);

The address operator may result in a static address.

Expression staticExpression =

StaticExpression.Unary(MiddleOperator.Address, expression);

if (staticExpression!= null) {

return staticExpression ;

}

If the expression has floating type, we need to pop the value from the floating value stack.

if (expression.Symbol.Type.IsFloating()) {

AddMiddleCode(expression.LongList, MiddleOperator.PopFloat);

}

The type of the resulting expression is a pointer to the type of the original expression

Type pointerType = new Type(expression.Symbol.Type);

Symbol resultSymbol = new Symbol(pointerType);

AddMiddleCode(expression.LongList, MiddleOperator.Address,

resultSymbol, expression.Symbol);

return (new Expression(resultSymbol, expression.ShortList,

expression.LongList));

}

### Dereference Expression

The dereference operator applies to pointers, arrays, strings, and functions. In the function case, the expression is regarded as a pointer to a function.

MiddleCodeGenerator.cs

public static Expression DereferenceExpression(Expression expression) {

Assert.Error(expression.Symbol.Type.IsPointerArrayStringOrFunction(),

Message.Invalid\_dereference\_of\_non\_\_pointer);

Symbol resultSymbol =

new Symbol(expression.Symbol.Type.PointerOrArrayType);

return Dereference(expression, resultSymbol, 0);

}

### Prefix Increment and Decrement Expression

MiddleCodeGenerator.cs

private static IDictionary<MiddleOperator,MiddleOperator> m\_incrementMap =

new Dictionary<MiddleOperator, MiddleOperator>();

private static IDictionary<MiddleOperator,MiddleOperator>

m\_incrementInverseMap = new Dictionary<MiddleOperator,MiddleOperator>();

static MiddleCodeGenerator() {

m\_incrementMap.Add(MiddleOperator.Increment,

MiddleOperator.BinaryAdd);

m\_incrementMap.Add(MiddleOperator.Decrement,

MiddleOperator.BinarySubtract);

m\_incrementInverseMap.Add(MiddleOperator.Increment,

MiddleOperator.BinarySubtract);

m\_incrementInverseMap.Add(MiddleOperator.Decrement,

MiddleOperator.BinaryAdd);

}

public static Expression PrefixIncrementExpression

(MiddleOperator middleOp, Expression expression){

Symbol symbol = expression.Symbol;

Assert.Error(symbol.Assignable, Message.Not\_assignable);

Assert.Error(symbol.Type.IsArithmeticOrPointer(),

expression, Message.Invalid\_type\_in\_increment\_expression);

if (symbol.Type.IsIntegralOrPointer()) {

AddMiddleCode(expression.ShortList, middleOp, null, symbol);

AddMiddleCode(expression.LongList, middleOp, null, symbol);

BigInteger? bitFieldMask = symbol.Type.GetBitfieldMask();

if (bitFieldMask != null) {

Symbol maskSymbol = new Symbol(symbol.Type, bitFieldMask.Value);

MiddleCode maskCode = new MiddleCode(MiddleOperator.BitwiseAnd,

symbol, symbol, maskSymbol);

expression.ShortList.Add(maskCode);

expression.LongList.Add(maskCode);

}

Symbol resultSymbol = new Symbol(symbol.Type);

AddMiddleCode(expression.LongList, MiddleOperator.Assign,

resultSymbol, symbol);

return (new Expression(resultSymbol, expression.ShortList,

expression.LongList));

}

The increment and decrement operator apply not only to integral values, but also to floating values. The operation. We start by pushing the value one at the floating value stack, the value to be incremented or decremented has already been pushed at the stack by earlier operations. We perform the operation, which is addition or subtraction. We preform the operation on both the short list and list of the expression. The difference is that in the short list case we pop the value off the stack since we do not need it anymore. In the long list case we do nothing, we just let the value stay on the stack to be used by later operations.

else {

AddMiddleCode(expression.ShortList, MiddleOperator.PushOne);

Symbol oneSymbol = new Symbol(symbol.Type, (decimal) 1);

AddMiddleCode(expression.ShortList, m\_incrementMap[middleOp],

symbol, symbol, oneSymbol);

AddMiddleCode(expression.ShortList, MiddleOperator.PopFloat, symbol);

AddMiddleCode(expression.LongList, MiddleOperator.PushOne);

AddMiddleCode(expression.LongList, m\_incrementMap[middleOp],

symbol, symbol, oneSymbol);

AddMiddleCode(expression.LongList, MiddleOperator.TopFloat, symbol);

Symbol resultSymbol = new Symbol(symbol.Type);

return (new Expression(resultSymbol, expression.ShortList,

expression.LongList));

}

}

### Postfix Increment and Decrement Expression

MiddleCodeGenerator.cs

public static Expression PostfixIncrementExpression

(MiddleOperator middleOp, Expression expression) {

Symbol symbol = expression.Symbol;

Assert.Error(symbol.Assignable, Message.Not\_assignable);

Assert.Error(symbol.Type.IsArithmeticOrPointer(),

expression, Message.Invalid\_type\_in\_increment\_expression);

if (symbol.Type.IsIntegralOrPointer()) {

Symbol resultSymbol = new Symbol(symbol.Type);

AddMiddleCode(expression.LongList, MiddleOperator.Assign,

resultSymbol, symbol);

AddMiddleCode(expression.ShortList, middleOp, null, symbol);

AddMiddleCode(expression.LongList, middleOp, null, symbol);

BigInteger? bitFieldMask = symbol.Type.GetBitfieldMask();

if (bitFieldMask != null) {

Symbol maskSymbol = new Symbol(symbol.Type, bitFieldMask.Value);

AddMiddleCode(expression.ShortList, MiddleOperator.BitwiseAnd,

symbol, symbol, maskSymbol);

AddMiddleCode(expression.LongList, MiddleOperator.BitwiseAnd,

symbol, symbol, maskSymbol);

}

return (new Expression(resultSymbol, expression.ShortList,

expression.LongList));

}

else {

AddMiddleCode(expression.ShortList, MiddleOperator.PushOne);

Symbol oneSymbol = new Symbol(symbol.Type, (decimal) 1);

AddMiddleCode(expression.ShortList, m\_incrementMap[middleOp],

symbol, symbol, oneSymbol);

AddMiddleCode(expression.ShortList, MiddleOperator.PopFloat, symbol);

In the long list case, the situation becomes a little more complicated, since the result of the operations shall be the original value, not the resulting value. Therefore, we must perform the inverse operation on the value on the stack in order to return to the original value.

AddMiddleCode(expression.LongList, MiddleOperator.PushOne);

AddMiddleCode(expression.LongList, m\_incrementMap[middleOp],

symbol, symbol, oneSymbol);

AddMiddleCode(expression.ShortList, MiddleOperator.TopFloat, symbol);

AddMiddleCode(expression.LongList, MiddleOperator.PushOne);

AddMiddleCode(expression.LongList, m\_incrementInverseMap[middleOp],

symbol, symbol, oneSymbol);

Symbol resultSymbol = new Symbol(symbol.Type);

return (new Expression(resultSymbol, expression.ShortList,

expression.LongList));

}

}

### Arrow Expression

MiddleCodeGenerator.cs

public static Expression ArrowExpression(Expression expression,

string memberName) {

Assert.Error(expression.Symbol.Type.IsPointer() &&

expression.Symbol.Type.PointerType.IsStructOrUnion(),

expression,

Message.Not\_a\_pointer\_to\_a\_struct\_or\_union\_in\_arrow\_expression);

Symbol memberSymbol =

expression.Symbol.Type.PointerType.MemberMap[memberName];

Assert.Error(memberSymbol != null, memberName,

Message.Unknown\_member\_in\_arrow\_expression);

Symbol resultSymbol = new Symbol(memberSymbol.Type);

return Dereference(expression, resultSymbol, memberSymbol.Offset);

}

### Index Expression

MiddleCodeGenerator.cs

public static Expression IndexExpression(Expression leftExpression,

Expression rightExpression) {

Expression staticExpression =

StaticExpression.Binary(MiddleOperator.Index, leftExpression,

rightExpression);

if (staticExpression != null) {

return staticExpression;

}

Expression arrayExpression, indexExpression;

if (leftExpression.Symbol.Type.IsPointerOrArray()) {

arrayExpression = leftExpression;

indexExpression = rightExpression;

}

else {

indexExpression = leftExpression;

arrayExpression = rightExpression;

}

Type arrayType = arrayExpression.Symbol.Type,

indexType = indexExpression.Symbol.Type;

Assert.Error(arrayType.IsPointerOrArray() &&

!arrayType.PointerOrArrayType.IsVoid(),

arrayType, Message.Invalid\_type\_in\_index\_expression);

Assert.Error(indexType.IsIntegral(), indexExpression,

Message.Invalid\_type\_in\_index\_expression);

List<MiddleCode> shortList = new List<MiddleCode>();

shortList.AddRange(arrayExpression.ShortList);

shortList.AddRange(indexExpression.ShortList);

Symbol resultSymbol =

new Symbol(arrayExpression.Symbol.Type.PointerOrArrayType);

if (indexExpression.Symbol.Value is BigInteger) {

int indexValue = (int) ((BigInteger)indexExpression.Symbol.Value),

indexSize = arrayExpression.Symbol.Type.PointerOrArrayType.Size();

return Dereference(arrayExpression, resultSymbol,

indexValue \* indexSize);

}

else {

indexExpression =

TypeCast.ImplicitCast(indexExpression, arrayExpression.Symbol.Type);

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(arrayExpression.LongList);

longList.AddRange(indexExpression.LongList);

Symbol arraySymbol = arrayExpression.Symbol,

indexSymbol = indexExpression.Symbol;

Symbol sizeSymbol = new Symbol(indexSymbol.Type, (BigInteger)

arraySymbol.Type.PointerOrArrayType.Size()),

multSymbol = new Symbol(arraySymbol.Type);

AddMiddleCode(longList, MiddleOperator.UnsignedMultiply,

multSymbol, indexSymbol, sizeSymbol);

Symbol addSymbol = new Symbol(arraySymbol.Type);

AddMiddleCode(longList, MiddleOperator.BinaryAdd,

addSymbol, arraySymbol, multSymbol);

Expression addExpression =

new Expression(addSymbol, shortList, longList);

return Dereference(addExpression, resultSymbol, 0);

}

}

private static Expression Dereference(Expression expression,

Symbol resultSymbol, int offset) {

resultSymbol.AddressSymbol = expression.Symbol;

resultSymbol.AddressOffset = offset;

AddMiddleCode(expression.LongList, MiddleOperator.Dereference,

resultSymbol, expression.Symbol, 0);

If the resulting expression holds floating type, we need to pop the value from the floating value stack.

if (resultSymbol.Type.IsFloating()) {

AddMiddleCode(expression.LongList, MiddleOperator.PushFloat,

resultSymbol);

}

return (new Expression(resultSymbol, expression.ShortList,

expression.LongList));

}

### Dot Expression

MiddleCodeGenerator.cs

public static Expression DotExpression(Expression expression,

string memberName) {

Symbol parentSymbol = expression.Symbol;

Assert.Error(parentSymbol.Type.IsStructOrUnion(), expression,

Message.Not\_a\_struct\_or\_union\_in\_dot\_expression);

Assert.Error(parentSymbol.Type.MemberMap != null, expression,

Message.Member\_access\_of\_uncomplete\_struct\_or\_union);

Symbol memberSymbol;

Assert.Error(parentSymbol.Type.MemberMap.TryGetValue(memberName,

out memberSymbol), memberName,

Message.Unknown\_member\_in\_dot\_expression);

Symbol resultSymbol;

if (parentSymbol.AddressSymbol != null) {

string name = parentSymbol.Name + "." + memberSymbol.Name +

Symbol.SeparatorId + memberSymbol.Offset;

resultSymbol = new Symbol(name, parentSymbol.ExternalLinkage,

parentSymbol.Storage, memberSymbol.Type,

parentSymbol.IsParameter());

resultSymbol.UniqueName = parentSymbol.UniqueName;

resultSymbol.AddressSymbol = parentSymbol.AddressSymbol;

resultSymbol.AddressOffset = parentSymbol.AddressOffset;

resultSymbol.Offset = parentSymbol.Offset + memberSymbol.Offset;

}

else {

resultSymbol = new Symbol(memberSymbol.Type);

resultSymbol.Name = parentSymbol.Name + Symbol.SeparatorId +

memberName;

resultSymbol.UniqueName = parentSymbol.UniqueName;

resultSymbol.Storage = parentSymbol.Storage;

resultSymbol.Offset = parentSymbol.Offset + memberSymbol.Offset;

}

return (new Expression(resultSymbol, expression.ShortList,

expression.LongList));

}

### Function Call Expression

The FunctionPreCall method is called before the argument list is parsed. It checks that the expression is either a function or a pointer to a function, adds the parameter type list of the function to the type list stack, adds zero to the current offset stack, and adds the call header instruction to the short and long list of the expression.

MiddleCodeGenerator.cs

public static void FunctionPreCall(Expression expression) {

Type type = expression.Symbol.Type;

Assert.Error(type.IsFunction() || type.IsFunctionPointer(),

expression.Symbol, Message.Not\_a\_function);

Type functionType = type.IsFunction() ? type : type.PointerType;

TypeListStack.Push(functionType.TypeList);

ParameterOffsetStack.Push(0);

AddMiddleCode(expression.ShortList, MiddleOperator.FunctionPreCall,

SymbolTable.CurrentTable.CurrentOffset);

AddMiddleCode(expression.LongList, MiddleOperator.FunctionPreCall,

SymbolTable.CurrentTable.CurrentOffset);

}

The CallExpression method is called after the argument list is parsed. It checks that the expression is either a function or a pointer to a function, adds the parameter type list of the function to the type list stack, adds zero to the current offset stack, and adds the call header instruction to the short and long list of the expression.

public static Expression CallExpression(Expression functionExpression,

List<Expression> argumentList){

List<Type> typeList = TypeListStack.Pop();

ParameterOffsetStack.Pop();

The reason we keep track of the current offset is that we in case of nested function calls need to allocate space for the previous arguments of the functions call outside the nested function. If the current offset is more than zero, we add it together with the size of the standard function top the total offset. If the current offset is zero, no argument has yet been passed, and we do not need to increase the total offset.

int totalOffset = 0;

foreach (int currentOffset in ParameterOffsetStack) {

if (currentOffset > 0) {

totalOffset += (SymbolTable.FunctionHeaderSize + currentOffset);

}

}

Type functionType = functionExpression.Symbol.Type.IsPointer() ?

functionExpression.Symbol.Type.PointerType :

functionExpression.Symbol.Type;

We check that there are not too few parameters in the call. If the type list is null, the number of arguments does not matter. However, if the type list is not null, the number of arguments must be at least the number of parameters in the type list.

Assert.Error((typeList == null) ||

(argumentList.Count >= typeList.Count),

functionExpression,

Message.Too\_few\_actual\_parameters\_in\_function\_call);

We also check that there are not too many parameters in the call. If the type list is null or if the callee function has a variadic parameter list, the number of arguments does not matter. There may be more arguments than parameters, we already know that the arguments are not fewer than the parameters. However, if the type list is not null and the function does not have a variadic parameter list, the number of arguments must equal the number of parameters.

Assert.Error(functionType.IsVariadic() || (typeList == null) ||

(argumentList.Count == typeList.Count),

functionExpression,

Message.Too\_many\_parameters\_in\_function\_call);

List<MiddleCode> longList = new List<MiddleCode>();

longList.AddRange(functionExpression.LongList);

When iterate through the, we keep track of the extra size. In variadic function calls we need to allocate memory for the extra parameters that is not included in the regular parameter list.

int count = 0, offset = SymbolTable.FunctionHeaderSize, extra = 0;

foreach (Expression argumentExpression in argumentList) {

longList.AddRange(argumentExpression.LongList);

Type type;

if ((typeList != null) && (count++ < typeList.Count)) {

type = typeList[count];

}

else {

type = ParameterType(argumentExpression.Symbol);

extra += type.Size();

}

We add the argument to the parameter list and update the current offset.

if (type.IsStructOrUnion()) {

AddMiddleCode(longList, MiddleOperator.ParameterInitSize,

SymbolTable.CurrentTable.CurrentOffset + totalOffset +

offset, type, argumentExpression.Symbol);

}

AddMiddleCode(longList, MiddleOperator.Parameter,

SymbolTable.CurrentTable. CurrentOffset + totalOffset +

offset, type, argumentExpression.Symbol);

offset += type.Size();

}

We add both the function call and post call instructions to the long list. The post call instruction is for

Symbol functionSymbol = functionExpression.Symbol;

AddMiddleCode(longList, MiddleOperator.Call,

SymbolTable.CurrentTable.CurrentOffset + totalOffset,

functionSymbol, extra);

AddMiddleCode(longList, MiddleOperator.PostCall,

SymbolTable.CurrentTable.CurrentOffset + totalOffset);

Type returnType = functionType.ReturnType;

Symbol returnSymbol = new Symbol(returnType);

We make the short list be a copy of the long list. However, there is one difference between the lists if the function returns a floating value. In that case, we pop the value off the floating value stack in the short list, since the value is not interesting in the short list, only the side effect of the function call. However, we keep the value on the stack in the long list, since the value shall be used in the long list.

List<MiddleCode> shortList = new List<MiddleCode>();

shortList.AddRange(longList);

if (!returnType.IsVoid()) {

if (returnType.IsStructOrUnion()) {

Type pointerType = new Type(returnType);

Symbol addressSymbol = new Symbol(pointerType);

returnSymbol.AddressSymbol = addressSymbol;

}

AddMiddleCode(longList, MiddleOperator.GetReturnValue,

returnSymbol);

if (returnType.IsFloating()) {

AddMiddleCode(shortList, MiddleOperator.PopEmpty);

}

}

return (new Expression(returnSymbol, shortList, longList));

}

### Argument Expression List

The ArgumentExpression method type casts the argument into an appropriate type.

MiddleCodeGenerator.cs

public static Expression ArgumentExpression(int index,

Expression expression) {

We need to type list on top of the type list stack to compare the argument expression with the parameter type.

List<Type> typeList = TypeListStack.Peek();

If the argument is within the parameter list, we type cast the argument into the parameter type.

if ((typeList != null) && (index < typeList.Count)) {

expression = TypeCast.ImplicitCast(expression, typeList[index]);

}

else {

If the argument is not within the parameter list (variadic call), we type cast the argument into an appropriate type: character and short integer are cast into (signed or unsigned) integer, and float is cast into double. These type casts apply to the variadic function call rules of ANSI C.

Type type = expression.Symbol.Type;

if (type.IsChar() || type.IsShort()) {

if (type.IsSigned()) {

expression =

TypeCast.ImplicitCast(expression, Type.SignedIntegerType);

}

else {

expression =

TypeCast.ImplicitCast(expression, Type.UnsignedIntegerType);

}

}

else if (type.IsFloat()) {

expression = TypeCast.ImplicitCast(expression, Type.DoubleType);

}

}

We need to update current offset to allocate space in case of nested function calls.

int offset = ParameterOffsetStack.Pop();

ParameterOffsetStack.Push(offset +

ParameterType(expression.Symbol).Size());

return (new Expression(expression.Symbol, expression.LongList,

expression.LongList));

}

The ParameterType method returns the type of the parameter. In case of array, string, or function, it returns pointer types. Otherwise, it returns the regular type.

private static Type ParameterType(Symbol symbol) {

switch (symbol.Type.Sort) {

case Sort.Array:

return (new Type(symbol.Type.ArrayType));

case Sort.Function:

return (new Type(symbol.Type));

case Sort.String:

return (new Type(new Type(Sort.SignedChar)));

default:

return symbol.Type;

}

}

### Primary Expressions

The value holds floating type, we push it at the floating value stack.

MiddleCodeGenerator.cs

public static Expression ValueExpression(Symbol symbol) {

List<MiddleCode> longList = new List<MiddleCode>();

if (symbol.Type.IsFloating()) {

AddMiddleCode(longList, MiddleOperator.PushFloat, symbol);

}

return (new Expression(symbol, new List<MiddleCode>(), longList));

}

Given a name, we look it up in the current symbol table. If there is no symbol, we add a function without a parameter list, that returns a signed integer, to the symbol table.

MiddleCodeGenerator.cs

public static Expression NameExpression(string name) {

Symbol symbol = SymbolTable.CurrentTable.LookupSymbol(name);

Assert.Error(symbol != null, name, Message.Unknown\_name);

List<MiddleCode> shortList = new List<MiddleCode>(),

longList = new List<MiddleCode>();

if (symbol.Type.IsFloating()) {

AddMiddleCode(shortList, MiddleOperator.PushFloat, symbol);

AddMiddleCode(longList, MiddleOperator.PushFloat, symbol);

}

return (new Expression(symbol, shortList, longList));

}

The registers are only used internally, in conjunction with system calls. On some occasion the interrupt call returns information stored in a register.

The closed statements do also include a set of internal statements; that is, statements not included in standard C, but necessary for the standard library functionality. The first statement is the load register statement, which stores a value in a register. The operands are the name of a register and an expression of integral or pointer type with same size as the register.

MiddleCodeGenerator.cs

public static Expression RegisterExpression(Register register) {

List<MiddleCode> longList = new List<MiddleCode>();

int size = AssemblyCode.SizeOfRegister(register);

Type type = TypeSize.SizeToUnsignedType(size);

Symbol symbol = new Symbol(type);

AddMiddleCode(longList, MiddleOperator.InspectRegister,

symbol, register);

return (new Expression(symbol, new List<MiddleCode>(),

longList, register));

}

MiddleCodeGenerator.cs

public static Expression CarryFlagExpression() {

ISet<MiddleCode> trueSet = new HashSet<MiddleCode>(),

falseSet = new HashSet<MiddleCode>();

List<MiddleCode> longList = new List<MiddleCode>();

trueSet.Add(AddMiddleCode(longList, MiddleOperator.Carry));

falseSet.Add(AddMiddleCode(longList, MiddleOperator.Jump));

Symbol symbol = new Symbol(trueSet, falseSet);

return (new Expression(symbol, new List<MiddleCode>(), longList));

}

MiddleCodeGenerator.cs

public static Expression StackTopExpression() {

List<MiddleCode> longList = new List<MiddleCode>();

Symbol symbol = new Symbol(new Type(Type.UnsignedCharType));

AddMiddleCode(longList, MiddleOperator.StackTop, symbol);

return (new Expression(symbol, new List<MiddleCode>(), longList));

}

# Declaration Specifiers and Declarators

A symbol definition is made up by a specifier and a declarator. The Mask class is used to distinguish between the different type specifiers and is used by the Specifier class below.

Mask.cs

namespace CCompiler {

public enum Mask {StorageMask = 0x0000FFF,

Auto = 0x0000010,

Register = 0x0000020,

Static = 0x0000040,

Extern = 0x0000080,

Typedef = 0x0000100,

QualifierMask = 0x000F000,

Constant = 0x0001000,

Volatile = 0x0002000,

SortMask = 0xFFF0000,

Signed = 0x0010000,

Unsigned = 0x0020000,

Char = 0x0040000,

Short = 0x0100000,

Int = 0x0200000,

Long = 0x0400000,

Float = 0x0800000,

Double = 0x1000000,

Void = 0x2000000,

We use the bitwise or operator to mask together simply mask to compound masks.

SignedChar = Signed | Char,

UnsignedChar = Unsigned | Char,

ShortInt = Short | Int,

SignedShort = Signed | Short,

SignedShortInt = Signed | Short | Int,

UnsignedShort = Unsigned | Short,

UnsignedShortInt = Unsigned | Short | Int,

SignedInt = Signed | Int,

UnsignedInt = Unsigned | Int,

LongInt = Long | Int,

SignedLong = Signed | Long,

SignedLongInt = Signed | Long | Int,

UnsignedLong = Unsigned | Long,

UnsignedLongInt = Unsigned | Long | Int,

LongDouble = Long | Double};

}

The Specifier class is used to generate the type specified by a list of specifiers. The declaration specifiers are made up by keywords or struct, union, or enumeration declarations.

Specifier.cs

using System;

using System.Text;

using System.Collections.Generic;

namespace CCompiler {

public class Specifier {

The specifier decides the storage and type of the symbol, it also decide whether the symbol has external linkage.

private bool m\_externalLinkage;

private Storage? m\_storage;

private Type m\_type;

The m\_maskToSortMap map maps the masks of the Mask enumeration to the types of the Sort enumeration. However, the masks are represented by unsigned integer values since we use unsigned integer values to mask the keywords of the declaration specifiers in the SpecifierList method below.

private static IDictionary<int, Sort> m\_maskToSortMap =

new Dictionary<int, Sort>() {

{(int) Mask.Void, Sort.Void},

{(int) Mask.Char, Sort.SignedChar},

{(int) Mask.SignedChar, Sort.SignedChar},

{(int) Mask.UnsignedChar, Sort.UnsignedChar},

{(int) Mask.Short, Sort.SignedShortInt},

{(int) Mask.ShortInt, Sort.SignedShortInt},

{(int) Mask.SignedShort, Sort.SignedShortInt},

{(int) Mask.SignedShortInt, Sort.SignedShortInt},

{(int) Mask.UnsignedShort, Sort.UnsignedShortInt},

{(int) Mask.UnsignedShortInt, Sort.UnsignedShortInt},

{(int) Mask.Int, Sort.SignedInt},

{(int) Mask.Signed, Sort.SignedInt},

{(int) Mask.SignedInt, Sort.SignedInt},

{(int) Mask.Unsigned, Sort.UnsignedInt},

{(int) Mask.UnsignedInt, Sort.UnsignedInt},

{(int) Mask.Long, Sort.SignedLongInt},

{(int) Mask.LongInt, Sort.SignedLongInt},

{(int) Mask.SignedLong, Sort.SignedLongInt},

{(int) Mask.SignedLongInt, Sort.SignedLongInt},

{(int) Mask.UnsignedLong, Sort.UnsignedLongInt},

{(int) Mask.UnsignedLongInt, Sort.UnsignedLongInt},

{(int) Mask.Float, Sort.Float},

{(int) Mask.Double, Sort.Double},

{(int) Mask.LongDouble, Sort.LongDouble}

};

A specifier is made up by the external linkage, storage, and type.

public Specifier(bool externalLinkage, Storage? storage, Type type) {

m\_externalLinkage = externalLinkage;

m\_storage = storage;

m\_type = type;

}

public bool ExternalLinkage {

get { return m\_externalLinkage; }

}

public Storage Storage {

get { Assert.ErrorXXX(m\_storage != null);

return m\_storage.Value; }

}

public Type Type {

get { return m\_type; }

}

The static SpecifierList method takes a list of declaration specifiers and returns a Specifier object, holding the storage, type, and external linkage of the symbol.

public static Specifier SpecifierList(List<object> specifierList) {

int totalMaskValue = 0;

Type compoundType = null;

We iterate through the declaration specifier list and gather the keywords of the Mask enumeration. If a keyword occurs twice, we report an error.

foreach (object obj in specifierList) {

if (obj is Mask) {

int maskValue = (int) obj;

if ((maskValue & totalMaskValue) != 0) {

Assert.Error(MaskToString(maskValue),

Message.Keyword\_defined\_twice);

}

totalMaskValue |= maskValue;

}

If the element of the list is not a Mask enumeration, it must be a type. If more than one type occurs in the list, we report an error.

else {

if (compoundType != null) {

Assert.Error(MaskToString(totalMaskValue),

Message.Invalid\_specifier\_sequence);

}

compoundType = (Type) obj;

}

}

When we have the final mask, we begin by extracting the storage. The storage is initially set to null, and we mask out the storage part of the total mask. It is possible that there was no storage in the specifier list. However, if there is a storage it must match one of the storage masks. If it does not match, more than one storage specifiers were present in the declaration specifier list, and we report an error.

Storage? storage = null;

{ int totalStorageValue = totalMaskValue & ((int) Mask.StorageMask);

if (totalStorageValue != 0) {

Assert.Error(Enum.IsDefined(typeof(Mask), totalStorageValue),

MaskToString(totalStorageValue),

Message.Invalid\_specifier\_sequence);

storage = (Storage) totalStorageValue;

}

}

The next step is to determine the external linkage. The symbol has externa linkage if the scope of the current symbol table is global (the symbol is defined in global space). Moreover, the storage shall be null (is has not been stated in the declaration specifier list) or extern. If the symbol is static in global scope it has not external linkage.

bool externalLinkage = (SymbolTable.CurrentTable.Scope == Scope.Global)

&& ((storage == null) ||

(storage == CCompiler.Storage.Extern));

if (storage == null) {

if (SymbolTable.CurrentTable.Scope == Scope.Global) {

storage = Storage.Static;

}

else {

storage = Storage.Auto;

}

}

When the storage and external linkage has been taken care of, we perform a series of error checks. If the symbol is a function parameter (function scope) the storage must be null, auto or register.

if (SymbolTable.CurrentTable.Scope == Scope.Parameter) {

Assert.Error((storage == Storage.Auto) ||

(storage == Storage.Register), storage, Message.

Only\_auto\_or\_register\_storage\_allowed\_in\_parameter\_declaration);

}

If the scope of the current symbol table is a struct or union, the storage must also be null, auto, or register.

else if ((SymbolTable.CurrentTable.Scope == Scope.Struct) ||

(SymbolTable.CurrentTable.Scope == Scope.Union)) {

Assert.Error((storage == Storage.Auto) ||

(storage == Storage.Register), storage, Message.

Only\_auto\_or\_register\_storage\_allowed\_for\_struct\_or\_union\_scope);

}

If the scope of the current symbol table is global, the storage must also be null, extern, static, or typedef.

else if (SymbolTable.CurrentTable.Scope == Scope.Global) {

Assert.Error((storage == Storage.Extern) ||

(storage == Storage.Static) ||

(storage == Storage.Typedef), storage, Message.

Only\_extern\_\_\_\_static\_\_\_\_or\_typedef\_storage\_allowed\_in\_global\_scope);

}

If there is a compound type and if it is an enumeration, we need to add its members as signed integers to the symbol table.

if ((compoundType != null) && (compoundType.EnumItemSet != null)){

foreach (Pair<Symbol,bool> pair in compoundType.EnumItemSet){

Symbol enumSymbol = pair.First;

switch (enumSymbol.Storage = storage.Value) {

case CCompiler.Storage.Static:

SymbolTable.StaticSet.Add(ConstantExpression.

Value(enumSymbol));

break;

case CCompiler.Storage.Extern: {

bool enumInitializer = pair.Second;

Assert.Error(!enumInitializer,

enumSymbol + " = " + enumSymbol.Value,

Message.Extern\_enumeration\_item\_cannot\_be\_initialized);

}

break;

case CCompiler.Storage.Auto:

case CCompiler.Storage.Register:

SymbolTable.CurrentTable.SetOffset(enumSymbol);

break;

}

}

}

We check if the total mask holds the masks representing the constant or volatile keywords.

{ bool isConstant = (totalMaskValue & ((int) Mask.Constant)) != 0;

bool isVolatile = (totalMaskValue & ((int) Mask.Volatile)) != 0;

If the type is constant or volatile, we face a rather complicated situation. We need to mark the compound type as constant or volatile. However, we do not want to mark the struct or union specification. Therefore, we create a new compound type.

if ((isConstant || isVolatile) && (compoundType != null) &&

compoundType.IsStructOrUnion() /\*&& compoundType.HasTag()\*/) {

compoundType = new Type(compoundType.Sort, compoundType.MemberMap,

compoundType.MemberList);

}

Finally, we extract the sort mask and test whether it represent a valid type. If the declaration specifier list does not represent a valid combination, we report an error. For instance, signed short double is an invalid combination of declaration specifiers.

Sort? sort = null;

int sortMaskValue = totalMaskValue & ((int) Mask.SortMask);

if (sortMaskValue != 0) {

if (!m\_maskToSortMap.ContainsKey(sortMaskValue)) {

Assert.Error(MaskToString(sortMaskValue),

Message.Invalid\_specifier\_sequence);

}

sort = m\_maskToSortMap[sortMaskValue];

}

When defining the final type, there are four cases. If the compound type is not null and the sort is null, the result type is the compound type, with the possible addition of constant or volatile qualifiers.

Type type = null;

if ((compoundType != null) && (sort == null)) {

compoundType.Constant = (compoundType.Constant || isConstant);

compoundType.Volatile = (compoundType.Volatile || isVolatile);

type = compoundType;

}

If the compound type is null and the sort is not null, the result type is based on the sort.

else if ((compoundType == null) && (sort != null)) {

type = new Type(sort.Value);

type.Constant = isConstant;

type.Volatile = isVolatile;

}

If the compound type and sort are both null, there is no type. This is valid in function definitions, in which case the type shall be a signed integer.

else if ((compoundType == null) && (sort == null)) {

type = new Type(Sort.SignedInt);

type.Constant = isConstant;

type.Volatile = isVolatile;

}

If both the compound type and the sort are not null, the type specification is invalid. For instance, unsigned short struct {int i;} is an invalid combination.

else {

Assert.Error(MaskToString((int) sortMaskValue), Message.

Invalid\_specifier\_sequence\_together\_with\_type);

}

Finally, the specifier with the external linkage, storage, and type is returned.

return (new Specifier(externalLinkage, storage, type));

}

}

The QualifierList method is called when a pointer is declared. It exams whether the pointer is constant or volatile.

public static Type QualifierList(List<Mask> qualifierList) {

int totalMaskValue = 0;

foreach (Mask mask in qualifierList) {

int maskValue = (int) mask;

if ((maskValue & totalMaskValue) != 0) {

Assert.Error(MaskToString(maskValue),

Message.Keyword\_defined\_twice);

}

totalMaskValue |= maskValue;

}

Type type = Type.VoidPointerType;

type.Constant = (totalMaskValue & ((int) Mask.Constant)) != 0;

type.Volatile = (totalMaskValue & ((int) Mask.Volatile)) != 0;

return type;

}

The MaskToString method is called when reporting an error. It returns a string holding the declaration specifiers as text.

private static string MaskToString(int totalMaskValue) {

StringBuilder maskBuffer = new StringBuilder();

We iterate through the bits of the mask and add the text of the Mask enumeration matching each true bit (bit with value one).

for (int maskValue = 1; maskValue != 0; maskValue <<= 1) {

if ((maskValue & totalMaskValue) != 0) {

string maskName = Enum.GetName(typeof(Mask), maskValue).ToLower();

maskBuffer.Append(((maskBuffer.Length > 0) ? " " : "") + maskName);

}

}

return maskBuffer.ToString();

}

}

}

## Declarators

A declarator follows the declaration specifiers, it can be a simple variable, a pointer, an array, or a function with old-style or new-style parameter list. A declarator can be initialized with a value or marked with its size in bits. The bold part of the following code are examples of declarators. Only members of a struct can be marked with bits. Struct or union members, as well as function parameters, cannot be initialized.

int i, j = 2;

double x = 3.14;

int f(a, b, c), g(int i, double x);

char s[] = "Hello";

int \*p, \*q = &i;

long int value : 7; // Struct members only

The Declarator class handles a declarator. For each new element of the declarator, the add method is called. For instance, if the declaration specifier is an integer and the declarator is a pointer (int \*p;), add is called with a pointer first, and then with an integer. The second call attaches the integer to the pointer, so that the pointer points at the integer. In the same way, the array type becomes attached to an array and the return type becomes attached to a function.

namespace CCompiler {

public class Declarator {

private string m\_name;

private Type m\_firstType, m\_lastType;

public Declarator(string name) {

m\_name = name;

}

public string Name {

get { return m\_name; }

set { m\_name = value; }

}

public Type Type {

get { return m\_firstType; }

}

If we add a type to the declarator where there is no previous type, we set both the first and last type to that type.

public void Add(Type newType) {

if (m\_firstType == null) {

m\_firstType = m\_lastType = newType;

}

else {

switch (m\_lastType.Sort) {

If the last type of the declarator is a pointer, we set it to point at the new type.

case Sort.Pointer:

m\_lastType.PointerType = newType;

m\_lastType = newType;

break;

If the last type of the declarator is an array, we set the type of the array to be the new type. However, the array must be complete; that is, it must have a stated size. Moreover, the new type cannot be a function since arrays of functions are not allowed.

case Sort.Array:

Assert.Error(newType.IsComplete(),

Message.Array\_of\_incomplete\_type\_not\_allowed);

Assert.Error(!newType.IsFunction(),

Message.Array\_of\_function\_not\_allowed);

m\_lastType.ArrayType = newType;

m\_lastType = newType;

break;

If the last type of the declarator is a function, set the return type of the function to be the new type. However, the new type cannot be an array or a function since functions are not allowed to return an arrays or functions.

case Sort.Function:

Assert.Error(!newType.IsArray(),

Message.Function\_cannot\_return\_array);

Assert.Error(!newType.IsFunction(),

Message.Function\_cannot\_return\_function);

m\_lastType.ReturnType = newType;

m\_lastType = newType;

break;

}

}

}

}

}

# The Symbol Table

The symbol table keeps track of the values, types, functions, and variables of the code, variables defined by the programmer and well as temporary variables introduced by the compiler. It also holds struct and union tags (enumerations do not have tags). There are actually several symbol tables. There are symbol tables for the global space, for each function, for each block in the functions, and for the members of each struct or union. In this way, the symbol tables form a hierarchy where each table holds a reference to its parent table (except the table for the global space, which table parent reference is null).

For example, let us look at the following code.

int globalInt;

int i;

void main() {

register int i;

char mainChar;

if (globalCount > 0) { // Block 1

int i;

float ifFloat;

{ int i; // Block 2

double blockDouble;

// Point p

}

}

}

When the parsing reaches point p, the symbol table has the form below. Note that the variable i defined in every block and is present in every table. When a symbol is looked up, we start searching in the current table and continue to search up through the hierarchy until we found a symbol with the given name or the parent symbol reference is null, in which case the table representing the global space.

Static int globalInt

Static int i

m\_parentTable null

Global Space

Register i : int

Auto mainChar : char

m\_parentTable

main

auto int i

auto float ifFloat

m\_parentTable

Block 1

auto int i

auto double blockDouble

m\_parentTable

Block 2

SymbolTable.CurrentTable

null

The Scope enumeration holds the possible scopes of the symbol table. Each symbol table in the hierarchy holds a specific scope. The Block scope corresponds to a block inside a function.

Scope.cs

namespace CCompiler {

public enum Scope { Struct = (int) Sort.Struct,

Union = (int) Sort.Union,

Function, Global, Block};

}

The Storage enumeration holds the storage specifiers of a symbol.

Storage.cs

namespace CCompiler {

public enum Storage {Auto = (int) Mask.Auto,

Register = (int) Mask.Register,

Static = (int) Mask.Static,

Extern = (int) Mask.Extern,

Typedef = (int) Mask.Typedef};

}

The SymbolTable class holds a symbol table. The scope of a table is Global, Function, Block, Struct, or Union as described above.

SymbolTable.cs

using System;

using System.Collections.Generic;

namespace CCompiler {

public class SymbolTable {

private Scope m\_scope;

For each function call, an activation record is allocated at the call stack. The first three entries are the return jump address (the address of the instruction following the function call), the regular frame pointer (the address of the current activation record), and the variadic frame pointer (the regular frame pointer plus the size of potential extra parameters in a variadic call). The function header size is the sum of the size of those three entries. As the compiler in this book can be set to generate code for different target machines, the offsets are also given different values as the pointer size varies. See chapter 12 for more information about the frame pointers.

public static int ReturnAddressOffset = 0;

public static int RegularFrameOffset = TypeSize.PointerSize;

public static int VariadicFrameOffset = 2 \* TypeSize.PointerSize;

public static int FunctionHeaderSize = 3 \* TypeSize.PointerSize;

Each table holds a reference to its parent table, except the table of global space, which table parent reference is null.

private SymbolTable m\_parentTable;

The activation record is organized in the following way:

* The offsets of the return jump address, the regular frame pointer, and the variadic frame pointer.
* The regular function parameters.
* The extra parameters in case of a variadic call.
* Variables of auto or register storage.
* Temporary integral values are stored are loaded from the registers and stored to the activation record during function calls.
* Temporary floating-point values are loaded from the floating-point stack and to the activation record during function calls.

private int m\_currentOffset;

All parameters and variables of a function or a block, as well as the members of a struct or a union, are stored in the m\_entryMap and m\_entryList. We use m\_entryMap to look up symbols and m\_entryList when initializing values in a struct or union, in which case the symbols need to be stored in the order they were declared in the code. The standard C# Dictionary class provides fast searching but does not guarantee any order among its key-value pairs.

private IDictionary<string,Symbol> m\_entryMap =

new Dictionary<string,Symbol>();

private List<Symbol> m\_entryList = new List<Symbol>();

The tag map holds the types of the structs or unions that are marked with a tag.

private IDictionary<string,Type> m\_tagMap = new Dictionary<string,Type>();

CurrentTable and CurrentFunction hold references to the current symbol table and the function currently processed by the parser. We need the current symbol table when adding or looking up symbols, and the current function when generating code for calling a function or returning a value. In global scope, the CurrentFunction is null.

public static SymbolTable CurrentTable = null;

public static Symbol CurrentFunction = null;

GlobalStaticSet holds all static symbols. Only symbols of auto or register storage are given offsets. Symbols of static, extern, or typedef storage are not given offsets, since they are not stored at the activation record in runtime.

public static ISet<StaticSymbol> GlobalStaticSet;

public static StaticSymbolLinux InitSymbol, ArgsSymbol;

The constructor of the SymbolTable class takes the parent table (which is null in global space) and the scope of the table.

public SymbolTable(SymbolTable parentTable, Scope scope) {

m\_parentTable = parentTable;

In case of global scope, we create the global static set. There is only one table of global scope and one static set in each source code file. In this case we do not need to set the offset field since there are no variables of auto or register storage in global scope.

switch (m\_scope = scope) {

case Scope.Global:

GlobalStaticSet = new HashSet<StaticSymbol>();

InitSymbol = ArgsSymbol = null;

break;

In case of struct of union scope, the current offset is initialized to zero.

case Scope.Struct:

case Scope.Union:

m\_currentOffset = 0;

break;

In case of function scope, the current offset is initialized to the size of the header size. The return address, regular frame pointer, and variadic frame pointer holds the first entries, and the parameters are given the offsets following the initialization.

case Scope.Function:

m\_currentOffset = FunctionHeaderSize;

break;

In case of block scope, the current offset is initialized to the current offset of its parent table. When the parsing leaves a block the parent table again becomes the current table, and a following block is again initialized to the current offset of the parent table. In this way, symbols of different blocks can share the same space on the activation record.

case Scope.Block:

m\_currentOffset = m\_parentTable.m\_currentOffset;

break;

}

}

public Scope Scope {

get { return m\_scope; }

}

public SymbolTable ParentTable {

get { return m\_parentTable; }

}

public IDictionary<string,Symbol> EntryMap {

get { return m\_entryMap; }

}

public List<Symbol> EntryList {

get { return m\_entryList; }

}

public int CurrentOffset {

get { return m\_currentOffset; }

}

The AddSymbol method adds a symbol to the symbol table. It is possible to add two symbols with the same name if they have equal types and at least one them holds extern storage.

When adding a symbol with a name we have to check a few things. If the symbol is extern it cannot be initialized (its value must be null).

public void AddSymbol(Symbol newSymbol) {

string name = newSymbol.Name;

The name of the symbol may be null in case of unnamed function parameters.

if (name != null) {

Symbol oldSymbol;

We look up the entry list, and if we find another symbol with the same name, we check whether at least one of the symbols is extern and they have equals types. Otherwise, we have two non-extern symbols with the same name, and we report an error.

if (m\_entryMap.TryGetValue(name, out oldSymbol)) {

if (oldSymbol.Type.IsFunction() && newSymbol.Type.IsFunction()) {

Assert.Error(!oldSymbol.FunctionDefinition ||

!newSymbol.FunctionDefinition, name,

Message.Name\_already\_defined);

}

else {

Assert.Error(oldSymbol.IsExtern() || newSymbol.IsExtern(),

name, Message.Name\_already\_defined);

}

Assert.Error(oldSymbol.Type.Equals(newSymbol.Type),

name, Message.Different\_types\_in\_redeclaration);

If the new symbol is not extern, we remove the old symbol and replace it with the new symbol.

if (!newSymbol.IsExtern()) {

m\_entryMap[name] = newSymbol;

m\_entryList.Remove(oldSymbol);

m\_entryList.Add(newSymbol);

}

}

If there is no previous symbol with the same name, we just add the new symbol to the entry map and entry list.

else {

m\_entryMap[name] = newSymbol;

m\_entryList.Add(newSymbol);

}

}

If the variable is auto or register, it shall be given an offset in the activation record. Unless in union scope, in which each member always has offset zero.

if (!newSymbol.Type.IsFunction()) {

if (newSymbol.IsAutoOrRegister()) {

if (m\_scope == Scope.Union) {

newSymbol.Offset = 0;

}

If the new symbol is an enumeration constant item, is shall not be given an offset. This is because we do not know at this point if the item holds auto or register scope. An item is given auto storage from the beginning, but that may change when the whole enumeration is given its storage, which is done after the item is stored in the symbol table.

else if (!newSymbol.Type.EnumeratorItem) {

newSymbol.Offset = m\_currentOffset;

m\_currentOffset += newSymbol.Type.Size();

}

}

}

}

The SetOffset method sets the offset of a symbol. It is called when an enumeration constant item has finally been given auto or register storage.

public void SetOffset(Symbol symbol) {

symbol.Offset = m\_currentOffset;

m\_currentOffset += symbol.Type.Size();

}

The LookupSymbol looks up the symbol with the given in the current symbol table and its parent tables, recursively. If no symbol with the given name is found, null is returned. Note that this method searches the parent tables while AddSymbol only checks for symbols with the same name in the current table.

public Symbol LookupSymbol(string name) {

Symbol symbol;

if (m\_entryMap.TryGetValue(name, out symbol)) {

return symbol;

}

else if (m\_parentTable != null) {

return m\_parentTable.LookupSymbol(name);

}

return null;

}

The AddTag method adds a struct or union tag to the current symbol table. It is possible to add a new tag with an already taken name. However, in that case, the old and the new tag must have the same sort (they must both be structs or both be unions) and the member map of at least one of them must be null. This order is necessary to provide linked lists in C.

public void AddTag(string name, Type newType) {

if (m\_tagMap.ContainsKey(name)) {

Type oldType = m\_tagMap[name];

Assert.Error(!oldType.IsEnumerator() &&

(oldType.Sort == newType.Sort), name,

Message.Name\_already\_defined);

If the member map of the old tag is null, we assign it the member map of the new tag.

if (oldType.MemberMap == null) {

oldType.MemberMap = newType.MemberMap;

oldType.MemberList = newType.MemberList;

}

If the neither the old nor the new member map is null, we report an error since it is not allowed to have two structs or unions with the same name tags with non-null member maps.

else {

Assert.Error(newType.MemberMap == null, name,

Message.Name\_already\_defined);

}

}

If there is no struct or union tag previous added to the tag map, we simply add the new tag to the map.

else {

m\_tagMap.Add(name, newType);

}

}

The LookUpTag method looks up a tag in the current table or its parent tables. If the tag is not found, null is returned.

public Type LookupTag(string name, Sort sort) {

Type type;

if (m\_tagMap.TryGetValue(name, out type) && (type.Sort == sort)) {

return type;

}

else if (m\_parentTable != null) {

return m\_parentTable.LookupTag(name, sort);

}

return null;

}

}

}

## The Symbol

The symbol table is made up of symbols, which have a storage, type, and potential value. If the storage specifier is omitted, a variable becomes static in global scope and auto in a function while a function becomes extern, unless it has a body in which case is becomes static. The register storage has been included for the sake of completeness. The only difference between an auto or register symbol is that the address operator cannot be applied to an auto symbol only.

The Symbol class describes a variable, a value, a function, or a type specified with typedef.

Symbol.cs

using System;

using System.IO;

using System.Text;

using System.Numerics;

using System.Globalization;

using System.Collections.Generic;

namespace CCompiler {

public class Symbol {

public const string NumberId = "#";

public const string TemporaryId = "£";

public const string SeparatorId = "$";

public const string SeparatorDot = ".";

public const string FileMarker = "@";

A symbol may have external linkage; that is, is visible from other source code files. A function may have a definition (a body with code) or be just a declaration (just the name, return type, and parameter list).

private bool m\_externalLinkage, m\_functionDefinition;

A symbol actually has two names: the regular name it has been defined with and is looked up by the symbol table, and a unique name. For symbols with external linkage, the regular and unique names are the same. For symbols without external linkage, the unique name is really unique. The idea is that the no two names without external linkage in any file shall have the same name. For instance, two static symbols in two different functions, or two global symbols in two different files, shall have different unique names.

private string m\_name, m\_uniqueName;

A symbol may be a parameter in a function definition.

private bool m\_parameter;

A symbol has a storage and a type, it may also have a value.

private Storage m\_storage;

private Type m\_type;

private object m\_value;

Symbols with auto and register storage are given an offset on the current activation record.

private int m\_offset;

When dereferencing a symbol with the dereference, arrow, and index (with a constant index value) expression, the dereferenced symbol is stored in the address symbol, with its offset. The offset is zero in the dereference case, the offset of the field in the arrow case, and the index value times the size of the array type in the index case. In case of the index operator with a non-constant index value, the value of the address offset is calculated in run-time.

private Symbol m\_addressSymbol;

private int m\_addressOffset;

A symbol may hold logical type, in which case the true-set and false-set are stored in the symbol.

private ISet<MiddleCode> m\_trueSet, m\_falseSet;

Finally, we have static counters for unique names and temporary names.

private static int UniqueNameCount = 0, TemporaryNameCount = 0;

The first constructor is used for defined symbols, variables, constants, and functions. In case of enumeration constants, they also hold a value.

public Symbol(string name, bool externalLinkage, Storage storage,

Type type, bool parameter = false, object value = null) {

m\_name = name;

m\_externalLinkage = externalLinkage;

m\_storage = storage;

If the symbol has external linkage its unique name is its regular name.

if (m\_externalLinkage) {

m\_uniqueName = m\_name;

}

If the symbol does not have external linkage, its unique name is a file marker, a unique number, and the regular name. The file marker will, by the linker, be replaced by a number unique for the object file. In this way, every symbol of the same file will have a unique number and each file will also have a unique number.

else {

m\_uniqueName = Symbol.FileMarker + (UniqueNameCount++) +

Symbol.SeparatorId + m\_name;

}

m\_parameter = parameter;

m\_value = CheckValue(m\_type = type, value);

}

If the value of the symbol is a BigInteger value, we check that it does not exceed its limits, which is defined for each integral type.

private static object CheckValue(Type type, object value) {

if (value is BigInteger) {

BigInteger bigValue = (BigInteger) value;

if (type.IsUnsigned() && (bigValue < 0)) {

bigValue += TypeSize.GetMaxValue(type.Sort) + 1;

}

Assert.Error((bigValue >= TypeSize.GetMinValue(type.Sort)) &&

(bigValue <= TypeSize.GetMaxValue(type.Sort)),

type + ": " + value, Message.Value\_overflow);

return bigValue;

}

return value;

}

In many expressions, the resulting value is a temporary symbol. In case of a dereference, arrow, dot, or index expression, the resulting symbol may be assignable and addressable.

public Symbol(Type type) {

m\_name = Symbol.TemporaryId + "temporary" + (TemporaryNameCount++);

m\_externalLinkage = false;

m\_storage = Storage.Auto;

m\_type = type;

m\_parameter = false;

}

The resulting symbol of a logical expression.

public Symbol(ISet<MiddleCode> trueSet, ISet<MiddleCode> falseSet) {

m\_name = Symbol.TemporaryId + "logical" + (TemporaryNameCount++);

m\_storage = Storage.Auto;

m\_type = new Type(Sort.Logical);

m\_trueSet = (trueSet != null) ? trueSet : (new HashSet<MiddleCode>());

m\_falseSet = (falseSet != null) ? falseSet

: (new HashSet<MiddleCode>());

m\_parameter = false;

}

We need to store a constant value as the denominator in an integral multiplication and division, and when pushing an integral or floating value to the floating value stack.

public Symbol(Type type, object value) {

Assert.ErrorXXX(!(value is bool));

m\_name = ValueName(type, value);

m\_uniqueName = Symbol.FileMarker + (UniqueNameCount++) +

Symbol.SeparatorId + m\_name;

m\_storage = Storage.Static;

m\_parameter = false;

m\_value = CheckValue(m\_type = type, value);

}

The value is given a name the reflects its type and its actual value.

public static string ValueName(CCompiler.Type type, object value) {

Assert.ErrorXXX(value != null);

If the value is a stgring, its name becomes the text with each character that is not a letter or digit replaced by its hex code.

if (value is string) {

string text = (string) value;

StringBuilder buffer = new StringBuilder();

for (int index = 0; index < text.Length; ++index) {

if (char.IsLetterOrDigit(text[index]) ||

(text[index] == '\_')) {

buffer.Append(text[index]);

}

else if (text[index] != '\0') {

int asciiValue = (int) text[index];

char hex1 = "0123456789ABCDEF"[asciiValue / 16],

hex2 = "0123456789ABCDEF"[asciiValue % 16];

buffer.Append(hex1.ToString() + hex2.ToString());

}

}

return "string\_" + buffer.ToString() + Symbol.NumberId;

}

If the value is a static address, the name is the text “staticaddress” with the unique name and offset.

else if (value is StaticAddress) {

StaticAddress staticAddress = (StaticAddress) value;

return "staticaddress" + Symbol.SeparatorId + staticAddress.UniqueName

+ Symbol.SeparatorId + staticAddress.Offset + Symbol.NumberId;

}

An array, floating, or integral value is given the name of the values’s string with the minus sign replaced by the text “minus”.

else if (type.IsArray()) {

return "Array\_" + value.ToString() + Symbol.NumberId;

}

else if (type.IsFloating()) {

return "float" + type.Size().ToString() + Symbol.SeparatorId +

value.ToString().Replace("-", "minus") + Symbol.NumberId;

}

else {

return "int" + type.Size().ToString() + Symbol.SeparatorId +

value.ToString().Replace("-", "minus") + Symbol.NumberId;

}

}

public string Name {

get { return m\_name; }

set { m\_name = value; }

}

public string UniqueName {

get { return m\_uniqueName; }

set { m\_uniqueName = value; }

}

public bool ExternalLinkage {

get { return m\_externalLinkage; }

}

public bool FunctionDefinition {

get { return m\_functionDefinition; }

set { m\_functionDefinition = value; }

}

public Storage Storage {

get { return m\_storage; }

set { m\_storage = value; }

}

public Type Type {

get { return m\_type; }

set { m\_type = value; }

}

public int Offset {

get { return m\_offset; }

set { m\_offset = value; }

}

public bool IsExtern() {

return (m\_storage == Storage.Extern);

}

public bool IsStatic() {

return (m\_storage == Storage.Static);

}

public bool IsExternOrStatic() {

return IsExtern() || IsStatic();

}

public bool IsTypedef() {

return (m\_storage == Storage.Typedef);

}

public bool IsAuto() {

return (m\_storage == Storage.Auto);

}

public bool IsRegister() {

return (m\_storage == Storage.Register);

}

public bool IsAutoOrRegister() {

return IsAuto() || IsRegister();

}

public ISet<MiddleCode> TrueSet {

get { return m\_trueSet; }

}

public ISet<MiddleCode> FalseSet {

get { return m\_falseSet; }

}

public bool Parameter {

get { return m\_parameter; }

}

public object Value {

get { return m\_value; }

set { m\_value = value; }

}

public Symbol AddressSymbol {

get { return m\_addressSymbol; }

set { m\_addressSymbol = value; }

}

public int AddressOffset {

get { return m\_addressOffset; }

set { m\_addressOffset = value; }

}

A symbol is a value if its name is not null (it may be null if it is parameter) and contains the number identifier.

public bool IsValue() {

return (m\_name != null) && m\_name.Contains(NumberId);

}

A symbol is temporary if its name is not null (it may be null if it is parameter) and contains the temporary identifier, and its address symbol is null. Technically, it may still be a temporary symbol if its address symbol is not null. However, in that case it shall not be regarded as a temporary symbol.

public bool IsTemporary() {

return (m\_name != null) && m\_name.Contains(TemporaryId) &&

(m\_addressSymbol == null);

}

The symbol is assignable if it is not a value, its type not recursively constant, and is not an array, function, or string. A type is recursively constant if it is constant or, in case of a struct or union, any of its members is recursively constant.

public bool IsAssignable() {

return !IsValue() && !m\_type.IsConstantRecursive() &&

!m\_type.IsArrayFunctionOrString();

}

public override string ToString() {

if (m\_value != null) {

return m\_value.ToString();

}

else if (m\_name != null) {

if (m\_addressSymbol != null) {

return m\_name + " -> " + m\_addressSymbol.ToString();

}

else {

return m\_name;

}

}

else {

if (m\_addressSymbol != null) {

return m\_addressSymbol.ToString();

}

else {

return "";

}

}

}

public static string SimpleName(string name) {

int index = name.LastIndexOf(Symbol.SeparatorId);

return (index != -1) ? name.Substring(index + 1).Replace("#", "")

: name;

}

}

}

## The Static Symbol

The StaticSymbol class is the base class of StaticSymbolLinux and StaticSymbolWindows.

StaticSymbol.cs

using System.IO;

namespace CCompiler {

public abstract class StaticSymbol {

The class holds a unique name.

private string m\_uniqueName;

public StaticSymbol() {

// Empty.

}

public StaticSymbol(string uniqueName) {

m\_uniqueName = uniqueName;

}

public string UniqueName {

get { return m\_uniqueName; }

}

Two static symbols are considered equal if they share the same unique name. This may happen if they hold values, in which case only the first of them shall be added to the global static set.

public override bool Equals(object obj) {

if (obj is StaticSymbol) {

return m\_uniqueName.Equals(((StaticSymbol) obj).m\_uniqueName);

}

return false;

}

public override int GetHashCode() {

return m\_uniqueName.GetHashCode();

}

The Write and Read methods as intended to be overridden by sub classes. In this class, they only write and read the unique name.

public virtual void Write(BinaryWriter outStream) {

outStream.Write(m\_uniqueName);

}

public virtual void Read(BinaryReader inStream) {

m\_uniqueName = inStream.ReadString();

}

}

}

The StaticSymbolLinux class is a sub class of StaticSymbol. The m\_textList field hold the assembly code instructions, and m\_externSet holds the extern references of the assembly code.

StaticSymbolLinux.cs

using System;

using System.IO;

using System.Collections.Generic;

namespace CCompiler {

public class StaticSymbolLinux : StaticSymbol {

private List<String> m\_textList;

private ISet<string> m\_externSet;

The constructor calls the base class constructor with the unique name.

public StaticSymbolLinux(string uniqueName, List<string> textList,

ISet<string> externSet)

:base(uniqueName) {

m\_textList = textList;

m\_externSet = externSet;

}

public List<string> TextList {

get { return m\_textList; }

}

public ISet<string> ExternSet {

get { return m\_externSet; }

}

}

}

There is also the StaticSymbolWindows class for the Windows environment. We will look into it in Chapter 13.

StaticSymbolWindows.cs

namespace CCompiler {

public class StaticSymbolWindows : StaticSymbol {

// ...

}

}

# The Type System

C has a rather large set of types. The simple types are made up of the integral types signed and unsigned char, short int, int, and long int as well as the floating types float, double, and long double. The compound types are pointers, arrays, structs, unions, and functions. Values of enumeration types (enum) are stored as signed integer. Moreover, a type can also be constant or volatile. Internally, we introduce the logical and string types, even though there are no such types in C. We also have the void type, which technically is not a type, but rather mark the absence of a type. It is used to mark the absence of a function return type, parameter list, or pointer type. However, array of void is not allowed.

The Sort enumeration holds the simple and compound types of C. Note that String and Logical are present, even though they are not types in C. However, temporary vale may hold these types.

## The Sort Enumeration

Sort.cs

namespace CCompiler {

public enum Sort {Void, SignedChar, UnsignedChar, SignedShortInt,

UnsignedShortInt, SignedInt, UnsignedInt,

SignedLongInt, UnsignedLongInt, Float, Double,

LongDouble, String, Pointer, Array, Struct, Union,

Function, Logical};

}

## The Type Class

The Type class holds methods for create types with different features. A pointer has a pointer type, an array has an array size and type, a function has a return type and parameter list, and a struct or union has a member map. A struct or union member may have a bitfield.

Type.cs

using System;

using System.Linq;

using System.Numerics;

using System.Collections.Generic;

namespace CCompiler {

public class Type {

private Sort m\_sort;

### Integral, Floating, and String

Each type holds a sort, it may be an integral sort (signed or unsigned character, short integer, integer, or long integer) a floating sort (float, double, or long double), a pointer or an array, a struct or a union, or a function.

public Sort Sort {

get { return m\_sort; }

}

The following constructor takes an integral or arithmetic type, or void.

public Type(Sort sort) { // arithmetic or logical

m\_sort = sort;

}

The IsSigned method return true if the type a signed character, short integer, integer, or long integer.

public static bool IsSigned(Sort sort) {

return (sort == Sort.SignedChar) || (sort == Sort.SignedShortInt) ||

(sort == Sort.SignedInt) || (sort == Sort.SignedLongInt);

}

public bool IsVoid() {

return (m\_sort == Sort.Void);

}

public bool IsChar() {

return (m\_sort == Sort.SignedChar) || (m\_sort == Sort.UnsignedChar);

}

public bool IsShort() {

return (m\_sort == Sort.SignedShortInt) ||

(m\_sort == Sort.UnsignedShortInt);

}

public bool IsInteger() {

return (m\_sort == Sort.SignedInt) || (m\_sort == Sort.UnsignedInt);

}

public bool IsIntegral() {

return IsSigned() || IsUnsigned();

}

public bool IsSigned() {

switch (m\_sort) {

case Sort.SignedChar:

case Sort.SignedShortInt:

case Sort.SignedInt:

case Sort.SignedLongInt:

return true;

default:

return false;

}

}

public bool IsUnsigned() {

switch (m\_sort) {

case Sort.UnsignedChar:

case Sort.UnsignedShortInt:

case Sort.UnsignedInt:

case Sort.UnsignedLongInt:

return true;

default:

return false;

}

}

public bool IsFloat() {

return (m\_sort == Sort.Float);

}

public bool IsFloating() {

switch (m\_sort) {

case Sort.Float:

case Sort.Double:

case Sort.LongDouble:

return true;

default:

return false;

}

}

public bool IsArithmetic() {

return IsIntegral() || IsFloating();

}

public bool IsString() {

return (m\_sort == Sort.String);

}

### Logical Types

Contrary to some other languages, there is no logical type in C. However, as C applies lazy evaluation, we need a logical type. Lazy evaluation means that an expression shall be not be evaluated more than necessary to determine its value. More specifically, if the left operand of the logical or operator is true the right operand shall not be evaluated. In the same way, if the left operand of the logical and operator is false the right operand shall not be evaluated. The same goes for the conditional operator, depending on whether the first expression is true or false, only the second or the third expression becomes evaluated. The logical type holds the two sets m\_trueSet and m\_falseSet, holding the middle code addresses to jumps instructions that will later be filled with the addresses to jump to if the expression is true or false.

public bool IsLogical() {

return (m\_sort == Sort.Logical);

}

### Bitfields

In case of a bitfield member in a struct or union, we store its bitfield mask. The mask is technically two to the power of the bits minus one or. Informally, it is the bits number of ones, counting from the least significant bit. The bitfield type is an integral type, with the addition of the bitfield mask. The mask is used to set the unused bits to zero when a bitfield variable is assigned a value.

private BigInteger? m\_bitfieldMask = null;

public void SetBitfieldMask(int bits) {

m\_bitfieldMask = (BigInteger) (Math.Pow(2, bits) - 1);

}

public bool IsBitfield() {

return (m\_bitfieldMask != null);

}

public BigInteger? GetBitfieldMask() {

return m\_bitfieldMask;

}

### Pointers

The type pointed at is null when the type is created, it will later be set by the Declarator class. The pointer type is given as a constructor parameter.

private Type m\_pointerType;

public Type(Type pointerType) {

m\_sort = Sort.Pointer;

m\_pointerType = pointerType;

}

public Type PointerType {

get { return m\_pointerType; }

set { m\_pointerType = value; }

}

public bool IsPointer() {

return (m\_sort == Sort.Pointer);

}

public bool IsIntegralOrPointer() {

return IsIntegral() || IsPointer();

}

public bool IsArithmeticOrPointer() {

return IsArithmetic() || IsPointer();

}

public bool IsIntegralLogicalOrPointer() {

return IsIntegralOrPointer() || IsLogical();

}

### Arrays

In case of an array, the constructor takes its size and the array type. When the type is created, the array size can be zero. In that case it will later be set by the length of its initialization list or become marked as uncomplete, in which case is can only be used as function parameter.

private int m\_arraySize;

private Type m\_arrayType;

public Type(int arraySize, Type arrayType) {

m\_sort = Sort.Array;

m\_arraySize = arraySize;

m\_arrayType = arrayType;

}

public int ArraySize {

get { return m\_arraySize; }

set { m\_arraySize = value; }

}

public Type ArrayType {

get { return m\_arrayType; }

set { m\_arrayType = value; }

}

public bool IsArray() {

return (m\_sort == Sort.Array);

}

public bool IsPointerOrArray() {

return IsPointer() || IsArray();

}

public bool IsIntegralPointerOrArray() {

return IsIntegral() || IsPointerOrArray();

}

The PointerOrArrayType property gets the pointer or array type, which case it may be.

public Type PointerOrArrayType {

get { return (m\_sort == Sort.Pointer) ? m\_pointerType : m\_arrayType; }

}

### Functions

C support both old-style and new-style function declarations. The old-style declaration takes a parameter list of identifiers, that is matched to a set of declarations, while the new-style declaration takes a parameter list of declarations. Therefore, function constructor is a bit complicated. However, both styles have in common that the function has a return type.

public enum FunctionStyle {Old, New};

private FunctionStyle m\_functionStyle;

private Type m\_returnType;

The old-style function has name list, while the new style function has a parameter list, made up by name-symbol pairs.

private List<string> m\_nameList;

private List<Symbol> m\_parameterList;

The parameter list is transformed into a type list, holding the types of each parameter.

private List<Type> m\_typeList;

Some functions, like printf and scanf, are variadic, which means that they can take a various number of parameters.

private bool m\_variadic;

The following constructor takes an old-style function. The names in the list must be unique; that is, the same name may not appear twice in the list.

public Type(Type returnType, List<string> nameList) {

m\_sort = Sort.Function;

m\_functionStyle = FunctionStyle.Old;

m\_returnType = returnType;

m\_nameList = nameList;

m\_parameterList = null;

m\_variadic = false;

We add the names of the list to a set, and check that they hold the same size. If they do not, there is at least two names with the same name, which is not allowed. If the same name is added to set twice, only the first is stored in the set, and the set will be smaller than the list.

Assert.Error(nameList.Count == new HashSet<string>(nameList).Count,

null, Message.Duplicate\_name\_in\_parameter\_list);

}

The following construct takes a new-style function, with return type, parameter list, and variadic status.

public Type(Type returnType, List< Symbol> parameterList,

bool variadic) {

m\_sort = Sort.Function;

m\_functionStyle = FunctionStyle.New;

m\_returnType = returnType;

m\_nameList = null;

m\_parameterList = parameterList;

m\_variadic = variadic;

m\_typeList = null;

If the parameter list is not null, we iterate through it and add the types to the type list. If the parameter list is null, the type list also becomes null. In that case the function lacks a parameter list and, when called, accepts all arguments.

if (parameterList != null) {

m\_typeList = new List<Type>();

foreach (Symbol symbol in parameterList) {

m\_typeList.Add(symbol.Type);

}

}

}

public bool IsFunction() {

return (m\_sort == Sort.Function);

}

public bool IsFunctionPointer() {

return (m\_sort == Sort.Pointer) && m\_pointerType.IsFunction();

}

public bool IsArrayFunctionOrString() {

return IsArray() || IsFunction() || IsString();

}

public bool IsIntegralPointerOrFunction() {

return IsIntegralOrPointer() || IsFunction();

}

public bool IsPointerArrayStringOrFunction() {

return IsPointer() || IsArrayFunctionOrString ();

}

public bool IsIntegralPointerOrFunction() {

return IsIntegralOrPointer() || IsFunction();

}

public bool IsIntegralPointerArrayOrFunction() {

return IsIntegralPointerOrArray() || IsFunction();

}

public bool IsIntegralPointerArrayStringOrFunction() {

return IsIntegralPointerArrayOrFunction() || IsString();

}

public bool IsArithmeticPointerArrayStringOrFunction() {

return IsArithmeticOrPointer() || IsArray() ||

IsFunction() || IsString();

}

The following properties gets the style, name list, parameter list, type list, return type, and whther the function is variadic.

public FunctionStyle Style {

get { return m\_functionStyle; }

}

public List<string> NameList {

get { return m\_nameList; }

}

public List<Symbol> ParameterList {

get { return m\_parameterList; }

}

public List<Type> TypeList {

get { return m\_typeList; }

}

public Type ReturnType {

get { return m\_returnType; }

set { m\_returnType = value; }

}

public bool IsVariadic() {

return m\_variadic;

}

### Structs and Unions

A struct and union takes a list of symbols. A struct or union is tagged if is given a name:

|  |  |
| --- | --- |
| struct {  int a, b;  }; | struct s {  int a, b;  }; |
| (a) An untagged struct | (b) A struct tagged with the name s |

A struct or union definition can be meaningful or meaningless. The declaration struct s {int i;}; is meaningful, since s can later be used to define variables. The declaration struct {int i;} t; is also meaningful, since t is a defined variable. However, the declaration struct {int i;}; is meaningless, but allowed.

The member map may be null, in which case the declaration is uncomplete. Then it can only be used when defining a pointer to it, which is useful when constructing linked lists.

private IDictionary<string,Symbol> m\_memberMap;

private List<Symbol> m\_memberList;

public Type(Sort sort, IDictionary<string,Symbol> symbolMap,

List<Symbol> symbolList) {

m\_sort = sort;

m\_memberMap = symbolMap;

m\_memberList = symbolList;

}

public IDictionary<string,Symbol> MemberMap {

get { return m\_memberMap; }

set { m\_memberMap = value; }

}

public List<Symbol> MemberList {

get { return m\_memberList; }

set { m\_memberList = value; }

}

public bool IsStruct() {

return (m\_sort == Sort.Struct);

}

public bool IsUnion() {

return (m\_sort == Sort.Union);

}

public bool IsStructOrUnion() {

return IsStruct() || IsUnion();

}

public bool IsArrayFunctionStringStructOrUnion() {

return IsArray() || IsFunction() || IsString() || IsStructOrUnion();

}

### Enumerations

The enumeration type (enum) is stored as an integer with a value, explicitly stated or implicitly assigned. However, the Specifier class of Chapter 5 needs to know if the type is enum to initialize its value. Therefore, we add the m\_enum field.

private ISet<Pair<Symbol,bool>> m\_enumeratorItemSet;

private bool m\_enumeratorItem;

public Type(Sort sort, bool enumeratorItem) {

m\_sort = sort;

m\_enumeratorItem = enumeratorItem;

}

public bool EnumeratorItem {

get { return m\_enumeratorItem; }

}

public Type(Sort sort, ISet<Pair<Symbol,bool>> enumSet) {

m\_sort = sort;

m\_enumeratorItemSet = enumSet;

}

public ISet<Pair<Symbol,bool>> EnumItemSet {

get { return m\_enumeratorItemSet; }

}

public bool IsEnumerator() {

return (m\_enumeratorItemSet != null);

}

### Type Size

Each type has a size, even though void and function as well as incomplete arrays, struct, and union sizes are defined to zero. The size of an array is its size times the size of its type, the size of a struct is the sum of the sizes of its members, and the size of a union is the size of its largest member. Note that a pointer always has the same size, regardless of what it points at.

public int Size() {

switch (m\_sort) {

case Sort.Array:

return m\_arraySize \* m\_arrayType.Size();

case Sort.Struct:

if (m\_memberMap != null) {

int size = 0;

foreach (Symbol symbol in m\_memberMap.Values) {

size += symbol.Type.Size();

}

return size;

}

else {

return 0;

}

case Sort.Union:

if (m\_memberMap == null) {

int size = 0;

foreach (Symbol symbol in m\_memberMap.Values) {

size = Math.Max(size, symbol.Type.Size());

}

return size;

}

else {

return 0;

}

case Sort.Logical:

return TypeSize.SignedIntegerSize;

default:

return TypeSize.Size(m\_sort);

}

}

The SizeArray works as Size above, with the difference that arrays, functions, and strings is given pointer size. This method is called when arguments are passed in function calls.

public int SizeArray() {

switch (m\_sort) {

case Sort.Array:

case Sort.Function:

case Sort.String:

return TypeSize.PointerSize;

default:

return Size();

}

}

### Complete Types

It is possible to define an array without stating its size, in which case the array is given the size zero. In that case, the array size must be determined by the size of its initialization list. However, if the array definition lacks an initialization list, the array keeps the size zero and is considered incomplete. In the same way, it is possible to define only the name tag of a struct or union, with its member map to be defined later. In that case, the member map is given the value null, which is keep if the member map is not defined, and the struct or union is consider incomplete. Variables can only have complete types, and the pointer type, array type or function return type must be also complete.

public bool IsComplete() {

switch (m\_sort) {

case Sort.Array:

return (m\_arraySize > 0);

case Sort.Struct:

case Sort.Union:

return (m\_memberMap != null);

default:

return true;

}

}

### Constant and Volatile

The idea of the volatile qualifier is to prevent optimization, and since this book is focused on optimization techniques, we have no real use for the volatile qualifier. However, for the sake of completeness we include the m\_volatile field in the Type class. However, note that there makes no difference if a type is volatile or not.

private bool m\_constant, m\_volatile;

public bool Constant {

get { return m\_constant; }

set { m\_constant = value; }

}

public bool Volatile {

get { return m\_volatile; }

set { m\_volatile = value; }

}

A type is constant if its field m\_constant is true. However, a struct or union is constant if it is constant in itself, or if is any of its members if (recursively) constant.

public bool IsConstantRecursive() {

if (m\_constant) {

return true;

}

else if (IsStructOrUnion() && (m\_memberMap != null)) {

foreach (Symbol symbol in m\_memberMap.Values) {

if (symbol.Type.IsConstantRecursive()) {

return true;

}

}

}

return false;

}

### Hash Code and Equals

The GetHashCode method simply return the hash code of its base class. It has been included for the sake of completeness since we also include the Equals method.

public override int GetHashCode() {

return base.GetHashCode();

}

The Equals method return true if the types are equal.

public override bool Equals(object obj) {

if (obj is Type) {

Type type = (Type) obj;

if ((m\_constant == type.m\_constant) &&

(m\_volatile == type.m\_volatile) && (m\_sort == type.m\_sort)) {

Two pointers are considered to be equal if the types they point at are equal.

switch (m\_sort) {

case Sort.Pointer:

return m\_pointerType.Equals(type.m\_pointerType);

Two arrays are equal if their types are equal and they have the same size, or if at least one of the arrays have size zero.

case Sort.Array:

return ((m\_arraySize == 0) || (type.m\_arraySize == 0) ||

(m\_arraySize == type.m\_arraySize)) &&

m\_arrayType.Equals(type.m\_arrayType);

Two structs or unions are equals if they both are incomplete (their member maps are null) or if their member maps are equal. Note that they must not only have the same members, the members must also appear in the same order.

case Sort.Struct:

case Sort.Union:

return ((m\_memberMap == null) && (type.m\_memberMap == null)) ||

((m\_memberMap != null) && (type.m\_memberMap != null) &&

m\_memberMap.Equals(type.m\_memberMap));

Two functions are equal if their return types are equal, they have the same style (old or new) and their type lists, respectively, are equal. Note that two functions can be equal even if their parameters have different names, at long as their types are equal.

case Sort.Function:

return m\_returnType.Equals(type.m\_returnType) &&

(((m\_typeList == null) && (type.m\_typeList == null)) ||

((m\_typeList != null) && (type.m\_typeList != null) &&

m\_typeList.SequenceEqual(type.m\_typeList))) &&

(m\_variadic == type.m\_variadic);

Finally, if none of the cases above apply, we compare the sorts of the types.

default:

return true;

}

}

}

return false;

}

### Predefined Types

We have a set of predefined types.

public static Type SignedShortIntegerType =

new Type(Sort.SignedShortInt);

public static Type UnsignedShortIntegerType =

new Type(Sort.UnsignedShortInt);

public static Type SignedIntegerType = new Type(Sort.SignedInt);

public static Type UnsignedIntegerType = new Type(Sort.UnsignedInt);

public static Type SignedLongIntegerType = new Type(Sort.SignedLongInt);

public static Type UnsignedLongIntegerType =

new Type(Sort.UnsignedLongInt);

public static Type FloatType = new Type(Sort.Float);

public static Type DoubleType = new Type(Sort.Double);

public static Type LongDoubleType = new Type(Sort.LongDouble);

public static Type SignedCharType = new Type(Sort.SignedChar);

public static Type UnsignedCharType = new Type(Sort.UnsignedChar);

public static Type StringType = new Type(Sort.String);

public static Type VoidType = new Type(Sort.Void);

public static Type PointerTypeX = new Type(SignedIntegerType);

public static Type VoidPointerType = new Type(new Type(Sort.Void));

public static Type LogicalType = new Type(Sort.Logical);

### ToString

Finally, we the ToString method. It returns the sort of the type as a string, with underlines replaced bu spaces (‘ ‘).For instance, The sort signed\_short\_int becomes the string “signed short int”. It is only called in error messages.

public override string ToString() {

return Enum.GetName(typeof(Sort), m\_sort).Replace("\_", " ").ToLower();

}

}

}

## Type Size

The TypeSize class holds the sizes of the types and the minimum and maximum values of each type.

TypeSize.cs

using System.Numerics;

using System.Collections.Generic;

namespace CCompiler {

class TypeSize {

public static int PointerSize;

public static int SignedIntegerSize;

The m\_maskMap map holds the bit masks for integer values of one, two, and four bytes. Note that its values are the same for the Linux and the Windows environment.

private static IDictionary<int, BigInteger>

m\_maskMap = new Dictionary<int, BigInteger>() {

{1, (BigInteger) 0x000000FF},

{2, (BigInteger) 0x0000FFFF},

{4, (BigInteger) 0xFFFFFFFF}};

The m\_sizeMap map holds the size of type.

public static IDictionary<Sort,int> m\_sizeMap =

new Dictionary<Sort,int>();

The m\_signedMap and m\_unsignedMap maps hold the signed and unsigned integer types of each size.

public static IDictionary<int,Sort>

m\_signedMap = new Dictionary<int,Sort>(),

m\_unsignedMap = new Dictionary<int,Sort>();

The m\_minValueMap and m\_maxValueMap hold the minimum and maximum values of each integral type.

private static IDictionary<Sort,BigInteger>

m\_minValueMap = new Dictionary<Sort,BigInteger>(),

m\_maxValueMap = new Dictionary<Sort,BigInteger>();

The sizes and values of the types depends on whether the compiler generates code for the Linux or the Windows environment. In the Linux environment, a pointer is eight bytes and an integer is four bytes.

if (Start.Linux) {

PointerSize = 8;

SignedIntegerSize = 4;

m\_sizeMap.Add(Sort.Void, 0);

m\_sizeMap.Add(Sort.Function, 0);

m\_sizeMap.Add(Sort.Logical, 1);

m\_sizeMap.Add(Sort.Pointer, 8);

m\_sizeMap.Add(Sort.Array, 8);

m\_sizeMap.Add(Sort.String, 4);

m\_sizeMap.Add(Sort.SignedChar, 1);

m\_sizeMap.Add(Sort.UnsignedChar, 1);

m\_sizeMap.Add(Sort.SignedShortInt, 2);

m\_sizeMap.Add(Sort.UnsignedShortInt, 2);

m\_sizeMap.Add(Sort.SignedInt, 4);

m\_sizeMap.Add(Sort.UnsignedInt, 4);

m\_sizeMap.Add(Sort.SignedLongInt, 8);

m\_sizeMap.Add(Sort.UnsignedLongInt, 8);

m\_sizeMap.Add(Sort.Float, 4);

m\_sizeMap.Add(Sort.Double, 8);

m\_sizeMap.Add(Sort.LongDouble, 8);

m\_signedMap.Add(1, Sort.SignedChar);

m\_signedMap.Add(2, Sort.SignedShortInt);

m\_signedMap.Add(4, Sort.SignedInt);

m\_signedMap.Add(8, Sort.SignedLongInt);

m\_unsignedMap.Add(1, Sort.UnsignedChar);

m\_unsignedMap.Add(2, Sort.UnsignedShortInt);

m\_unsignedMap.Add(4, Sort.UnsignedInt);

m\_unsignedMap.Add(8, Sort.UnsignedLongInt);

m\_minValueMap.Add(Sort.Logical, 0);

m\_minValueMap.Add(Sort.SignedChar, -128);

m\_minValueMap.Add(Sort.UnsignedChar, 0);

m\_minValueMap.Add(Sort.SignedShortInt, -32768);

m\_minValueMap.Add(Sort.UnsignedShortInt, 0);

m\_minValueMap.Add(Sort.SignedInt, -2147483648);

m\_minValueMap.Add(Sort.UnsignedInt, 0);

m\_minValueMap.Add(Sort.Array, 0);

m\_minValueMap.Add(Sort.Pointer, 0);

m\_minValueMap.Add(Sort.SignedLongInt, -9223372036854775808);

m\_minValueMap.Add(Sort.UnsignedLongInt, 0);

m\_maxValueMap.Add(Sort.Logical, 1);

m\_maxValueMap.Add(Sort.SignedChar, 127);

m\_maxValueMap.Add(Sort.UnsignedChar, 255);

m\_maxValueMap.Add(Sort.SignedShortInt, 32767);

m\_maxValueMap.Add(Sort.UnsignedShortInt, 65535);

m\_maxValueMap.Add(Sort.SignedInt, 2147483647);

m\_maxValueMap.Add(Sort.UnsignedInt, 4294967295);

m\_maxValueMap.Add(Sort.Array, 4294967295);

m\_maxValueMap.Add(Sort.Pointer, 4294967295);

m\_maxValueMap.Add(Sort.SignedLongInt, 9223372036854775807);

m\_maxValueMap.Add(Sort.UnsignedLongInt, 18446744073709551615);

}

The type sizes for the Windows environment are defined in Chapter 13.

if (Start.Windows) {

// ...

}

}

public static BigInteger GetMinValue(Sort sort) {

return m\_minValueMap[sort];

}

public static BigInteger GetMaxValue(Sort sort) {

return m\_maxValueMap[sort];

}

public static BigInteger GetMask(Sort sort) {

return m\_maskMap[m\_sizeMap[sort]];

}

public static Type SizeToSignedType(int size) {

return new Type(m\_signedMap[size]);

}

public static Type SizeToUnsignedType(int size) {

return new Type(m\_unsignedMap[size]);

}

public static int Size(Sort sort) {

return m\_sizeMap[sort];

}

}

}

## Type Casting

Type casting occurs on several occasions in C. There are two types of cast: explicit (manual) conversation where type cast is stated and implicit (automatic) where it the type cast is omitted:

double x = 3.1;

int y = x, // implicit cast

z = (int) x; // explicit cast

TypeCast.cs

using System.Numerics;

using System.Collections.Generic;

namespace CCompiler {

public class TypeCast {

The LogicalToIntegral method makes sure the expression holds integral type. If it holds logical type, it is cast from logical type to signed integer type.

public static Expression LogicalToIntegral(Expression expression) {

if (expression.Symbol.Type.IsLogical()) {

return ImplicitCast(expression, Type.SignedIntegerType);

}

return expression;

}

The LogicalToFloating method makes sure the expression holds floating type. If it holds logical type, it is cast from logical type to double type.

public static Expression LogicalToFloating(Expression expression) {

if (expression.Symbol.Type.IsLogical()) {

return ImplicitCast(expression, Type.DoubleType);

}

return expression;

}

The ToLogical method makes sure the expression has a logical type. If it does not already hold logical type, the expression is cast to logical type.

public static Expression ToLogical(Expression expression) {

if (!expression.Symbol.Type.IsLogical()) {

return ImplicitCast(expression, Type.LogicalType);

}

return expression;

}

### Implicit Cast

The ImplicitCast method performs an implicit cast. To begin with, if the expression is constant, we return the constant casted expression.

public static Expression ImplicitCast(Expression fromExpression,

Type toType) {

Expression constantExpression =

ConstantExpression.Cast(fromExpression, toType);

if (constantExpression != null) {

return constantExpression;

}

Type fromType = fromExpression.Symbol.Type;

if (fromType.Equals(toType) ||

(fromType.IsLogical() && toType.IsLogical()) ||

(fromType.IsPointerArrayStringOrFunction() &&

toType.IsPointerOrArray()) ||

(fromType.IsPointerArrayStringOrFunction() &&

toType.IsIntegral() && (fromType.Size() == toType.Size())) ||

(((fromType.IsFloating() && toType.IsFloating()) ||

(fromType.IsIntegralPointerOrFunction() &&

toType.IsIntegralPointerArrayOrFunction())) &&

(fromType.SizeArray() == toType.SizeArray()))) {

return fromExpression;

}

else {

return ExplicitCast(fromExpression, toType);

}

}

### Explicit Cast

The ExplicitCast method perform an explicit cast between the different types of C.

public static Expression ExplicitCast(Expression sourceExpression,

Type targetType) {

Expression constantExpression =

ConstantExpression.ConstantCast(sourceExpression, targetType);

if (constantExpression != null) {

return constantExpression;

}

Symbol sourceSymbol = sourceExpression.Symbol, targetSymbol = null;

Type sourceType = sourceSymbol.Type;

List<MiddleCode> shortList = sourceExpression.ShortList,

longList = sourceExpression.LongList;

if (targetType.IsVoid()) {

targetSymbol = new Symbol(targetType);

}

else if (sourceType.IsStructOrUnion() && targetType.IsStructOrUnion()) {

Assert.Error(sourceType.Equals(targetType), sourceType + " to " +

targetType, Message.Invalid\_type\_cast);

targetSymbol = new Symbol(targetType);

}

If the source type is logical and the target type is integral, array, or pointer, we need to backpatch the true-set and false-set of the expression’s symbol.

The true-set becomes backpatched to a statement that assign the value one to a temporary variable in the

else if (sourceType.IsLogical() &&

targetType.IsIntegralPointerOrArray()) {

targetSymbol = new Symbol(targetType);

We backpatch the true-set to a statement that assign the value one to a temporary variable.

Symbol oneSymbol = new Symbol(targetType, BigInteger.One);

MiddleCode trueCode =

new MiddleCode(MiddleOperator.Assign, targetSymbol, oneSymbol);

MiddleCodeGenerator.Backpatch(sourceSymbol.TrueSet, trueCode);

longList.Add(trueCode);

In the middle of backpatching of the true-set and false-set, we add code that jumps to the zero value assignment.

MiddleCode targetCode = new MiddleCode(MiddleOperator.Empty);

longList.Add(new MiddleCode(MiddleOperator.Jump, targetCode));

We backpatch the false-set to a statement that assign the value one to a temporary variable.

Symbol zeroSymbol = new Symbol(targetType, BigInteger.Zero);

MiddleCode falseCode =

new MiddleCode(MiddleOperator.Assign, targetSymbol, zeroSymbol);

MiddleCodeGenerator.Backpatch(sourceSymbol.FalseSet, falseCode);

longList.Add(falseCode);

Finally, we add the target of the jump code.

longList.Add(targetCode);

}

If the source type is logical and the target type is floating, we push the values one and zero on the floating-point stack instead of assigning a variable.

else if (sourceType.IsLogical() && targetType.IsFloating()) {

targetSymbol = new Symbol(targetType);

We backpatch the true-set of the source symbol to the code that pushes one at the stack.

MiddleCode trueCode = new MiddleCode(MiddleOperator.PushOne);

MiddleCodeGenerator.Backpatch(sourceSymbol.TrueSet, trueCode);

longList.Add(trueCode);

Similar to the integral case above, in the middle of backpatching of the true-set and false-set, we add code that jumps to the push-zero assignment.

MiddleCode targetCode = new MiddleCode(MiddleOperator.Empty);

longList.Add(new MiddleCode(MiddleOperator.Jump, targetCode));

We backpatch the true-set of the source symbol to the code that pushes one at the stack.

MiddleCode falseCode = new MiddleCode(MiddleOperator.PushZero);

MiddleCodeGenerator.Backpatch(sourceSymbol.FalseSet, falseCode);

longList.Add(falseCode);

Finally, we add the target of the jump code.

longList.Add(targetCode);

}

In case of an arithmetic (integral or floating), array, string, or function source type, and logical target type, we test whether the value of the source expression is non-zero, and generate a true-set and a false-set from the test.

else if (sourceType.IsArithmeticPointerArrayStringOrFunction() &&

targetType.IsLogical()) {

We begin by creating the zero value to compare the source value with. It is decimal in case of floating type, and BigInteger in case of integral, pointer, array, string, of function.

object zeroValue = sourceType.IsLogical() ? ((object) decimal.Zero)

: ((object)BigInteger.Zero);

Symbol zeroSymbol = new Symbol(targetType, zeroValue);

We add code for comparing the source value with the zero value. If the source value is not zero we jump to yet unknown address. We add that first instruction to the true-set.

MiddleCode testCode =

new MiddleCode(MiddleOperator.NotEqual, null,

sourceSymbol, zeroSymbol);

ISet<MiddleCode> trueSet = new HashSet<MiddleCode>();

trueSet.Add(testCode);

longList.Add(testCode);

If the source value is zero, we jump to another yet unknown address. We add that instruction to the false-set.

MiddleCode gotoCode = new MiddleCode(MiddleOperator.Jump);

ISet<MiddleCode> falseSet = new HashSet<MiddleCode>();

falseSet.Add(gotoCode);

longList.Add(gotoCode);

Finally, the target symbol holds the true-set and false-set.

targetSymbol = new Symbol(trueSet, falseSet);

}

In case of a cast from an integral type of one byte (pointers are always two bytes) to a floating type, it becomes a little bit more complicated since there is no assembly converting operator. Instead, we need to first convert the small integral value (one byte) to a larger integral value (two bytes), which we in turn convert to a floating value.

In case of cast between two floating values, we only create the target symbol. We do not need to do anything else, since the floating-point value is already stored at the floating-point stack.

else if (sourceType.IsFloating() && targetType.IsFloating()) {

targetSymbol = new Symbol(targetType);

}

The same goes for the other way around, when converting a floating value to a small (one byte) integral value, we first need to convert the floating value to an integral value of two bytes, which we in turn convert to an integral value of one byte.

else if (sourceType.IsFloating() &&

targetType.IsIntegralPointerOrArray()) {

targetSymbol = new Symbol(targetType);

Due to technical limitations, there is no assembly instruction for popping a one-byte integral value from the floating-point stack. Therefore, when casting a floating value to a one-byte integral value, we first need to add a floating-to-integral instruction to a temporary integer value, which we in turn cast to the one-byte integral value.

if (targetType.Size() == 1) {

Type tempType = sourceType.IsSigned() ? Type.SignedIntegerType

: Type.UnsignedIntegerType;

Symbol tempSymbol = new Symbol(tempType);

MiddleCode tempCode =

new MiddleCode(MiddleOperator.FloatingToIntegral, tempSymbol,

sourceSymbol);

longList.Add(tempCode);

MiddleCode resultCode =

new MiddleCode(MiddleOperator.IntegralToIntegral, targetSymbol,

tempSymbol);

longList.Add(resultCode);

}

If the target type size is more than one byte, we simply add a floating-to-integral instruction.

else {

MiddleCode resultCode =

new MiddleCode(MiddleOperator.FloatingToIntegral, targetSymbol,

sourceSymbol);

longList.Add(resultCode);

}

}

If the source type is integral, pointer, array, string, or function, and the target type is floating, we have the opposite situation if the source type size is one. We cast the one-byte value to a temporary integer value, which we in turn cast to a floating value.

else if (sourceType.IsIntegralPointerArrayStringOrFunction () &&

targetType.IsFloating()) {

targetSymbol = new Symbol(targetType);

if (sourceType.Size() == 1) {

Type tempType = sourceType.IsSigned() ? Type.SignedIntegerType

: Type.UnsignedIntegerType;

Symbol tempSymbol = new Symbol(tempType);

MiddleCodeGenerator.

AddMiddleCode(longList, MiddleOperator.IntegralToIntegral,

tempSymbol, sourceSymbol);

MiddleCodeGenerator.

AddMiddleCode(longList, MiddleOperator.IntegralToFloating,

targetSymbol, tempSymbol);

}

If the target type size is more than one byte, we simply add an integral-to-floating instruction.

else {

MiddleCodeGenerator.

AddMiddleCode(longList, MiddleOperator.IntegralToFloating,

targetSymbol, sourceSymbol);

}

}

If the source type is integral, pointer, array, string, or function, and the target type is integral, array, or pointer, we simply add an integral-to-integral instruction.

else if (sourceType.IsIntegralPointerArrayStringOrFunction () &&

targetType.IsIntegralPointerOrArray()) {

targetSymbol = new Symbol(targetType);

MiddleCodeGenerator.

AddMiddleCode(longList, MiddleOperator.IntegralToIntegral,

targetSymbol, sourceSymbol);

}

In none of the cases above applies, the target symbol is still null, and we report an error.

Assert.Error(targetSymbol != null, sourceType + " to " +

targetType, Message.Invalid\_type\_cast);

return (new Expression(targetSymbol, shortList, longList));

}

### Type Promotion

Type promotion is the process of converting a “smaller” type to a “larger” type in a binary expression. For instance, in the expression below, the integer i shall be converted to a double value, since double is larger than integer.

int i = 3;

double x = 3.14, y;

y = x + i;

Below is a table holding the relations of the types. Note that float is considered to be larger than long int, even if they have the same size. In the same way, short is considered to the larger than char even if they have the same size.

|  |  |
| --- | --- |
| Largest type | long double |
|  | double |
|  | float |
|  | signed long int  unsigned long int |
|  | signed int  unsigned int |
|  | signed short int  unsigned short int |
| Smallest type | signed char  unsigned char |

In accordance with the Ansi C standard the signed and unsigned types have the same size, and a character is always one byte. This due to historical reasons. In the early versions of C, there was no void type. Instead, char was used a generic type in pointer expressions. Therefore, a character always holds one byte. However, the standard does not state whether a signed value shall be converted to an unsigned value or the other way around in the expression below. However, in this book we always convert the unsigned value to the matching signed type (if the hold the same size).

signed int i = -3;

unsigned int u = 6;

i + u;

TypeCast.cs

public static Type MaxType(Type leftType, Type rightType) {

if ((leftType.IsFloating() && !rightType.IsFloating()) ||

((leftType.Size() == rightType.Size()) &&

leftType.IsSigned() && !rightType.IsSigned())) {

return leftType;

}

else if ((!leftType.IsFloating() && rightType.IsFloating()) ||

((leftType.Size() == rightType.Size()) &&

!leftType.IsSigned() && rightType.IsSigned())) {

return rightType;

}

else {

return (leftType.Size() > rightType.Size()) ? leftType : rightType;

}

}

}

}

# Constant Expression

There are advantages by calculating the values of constant expression during the compilation rather than the execution. Besides the optimization benefits, resulting in smaller and faster code, there are some occasions when we need to calculate the value. In array definitions, we need to evaluate the size of the array, in initialized enumeration values we need to evaluate the value to be able to assign the next uninitialized enumeration its correct value, in case statement we need to decide the value in order to make sure the same case value does not occur twice in the same switch statement, and the preprocessor need to evaluate the values of if directives.

## Unary and Binary Expressions

The ConstantExpression class holds methods that calculates the value of unary and binary logical and arithmetic expressions, if their operands are constant.

ConstantExpression.cs

using System.Numerics;

using System.Collections.Generic;

namespace CCompiler {

public class ConstantExpression {

The IsConstant method returns true if the expression can be evaluated to a value, which it can if it holds logical type and either the true-set or the false-set is empty (both sets cannot be empty). If one set is empty, the expression holds no if statement, only a goto statement. It can also be evaluated to a constant value if it holds integral, pointer type, or floating type with a not-null value.

public static bool IsConstant(Expression expression) {

Symbol resultSymbol = expression.Symbol;

Type type = resultSymbol.Type;

return (type.IsLogical() && ((resultSymbol.TrueSet.Count == 0) ||

(resultSymbol.FalseSet.Count == 0))) ||

(type.IsIntegralOrPointer() &&

(resultSymbol.Value is BigInteger)) ||

(type.IsFloating() && (resultSymbol.Value is decimal));

}

The IsTrue method returns true if the (assumed constant) expression can be evaluated to a true value., which it can if the value is a reference to a BigInteger object that is not zero, or a decimal object that is not zero, or if the true-set is not empty.

public static bool IsTrue(Expression expression) {

return ((expression.Symbol.Value is BigInteger) &&

!((BigInteger) expression.Symbol.Value).IsZero) ||

((expression.Symbol.Value is decimal) &&

!((decimal) expression.Symbol.Value).Equals((decimal) 0)) ||

(expression.Symbol.TrueSet.Count > 0);

}

### Relation Expressions

The Relation method returns the calculated value of the expression, if possible. Otherwise, it returns null.

public static Expression Relation(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

If at least one of the expressions is not constant, we just return null.

if (!IsConstant(leftExpression) || !IsConstant(rightExpression)) {

return null;

}

If at least one of the expressions have floating type, we call the RelationFloating method. If not at least one of the expressions have floating type, both expressions must have integral type and we call RelationIntegral instead.

else if (leftExpression.Symbol.Type.IsFloating() ||

rightExpression.Symbol.Type.IsFloating()) {

return RelationFloating(middleOp, leftExpression, rightExpression);

}

else {

return RelationIntegral(middleOp, leftExpression, rightExpression);

}

}

The RelationIntegral method evaluates the constant value of a relation expression. At this point, we assume that both operands are constant.

private static Expression RelationIntegral(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

First, we need to cast the expression from a possible logical type into the integral type: signed integer.

leftExpression = TypeCast.LogicalToIntegral(leftExpression);

rightExpression = TypeCast.LogicalToIntegral(rightExpression);

This method is called only if both expressions have values. Therefore, we do not have to check that the values are not null.

BigInteger leftValue = (BigInteger) leftExpression.Symbol.Value,

rightValue = (BigInteger) rightExpression.Symbol.Value;

Since the expression has a constant value, we only add one unconditional jump instruction.

List<MiddleCode> longList = new List<MiddleCode>();

MiddleCode jumpCode = new MiddleCode(MiddleOperator.Jump);

longList.Add(jumpCode);

ISet<MiddleCode> jumpSet = new HashSet<MiddleCode>();

jumpSet.Add(jumpCode);

Then we need to evaluate the value of the expression. Since the values are object of the BigInteger class we can use the regular relation operators.

bool resultValue = false;

switch (middleOp) {

case MiddleOperator.Equal:

resultValue = (leftValue == rightValue);

break;

case MiddleOperator.NotEqual:

resultValue = (leftValue != rightValue);

break;

case MiddleOperator.SignedLessThan:

case MiddleOperator.UnsignedLessThan:

resultValue = (leftValue < rightValue);

break;

case MiddleOperator.SignedLessThanEqual:

case MiddleOperator.UnsignedLessThanEqual:

resultValue = (leftValue <= rightValue);

break;

case MiddleOperator.SignedGreaterThan:

case MiddleOperator.UnsignedGreaterThan:

resultValue = (leftValue > rightValue);

break;

case MiddleOperator.SignedGreaterThanEqual:

case MiddleOperator.UnsignedGreaterThanEqual:

resultValue = (leftValue >= rightValue);

break;

}

Finally, we define the logical symbol with the jump set, depending on the value of the expression. If the value is true the jump set becomes the true-set of the symbol, and if the value is false the jump set becomes the false-set of the symbol.

Symbol resultSymbol = resultValue ? (new Symbol(jumpSet, null))

: (new Symbol(null, jumpSet));

The resulting expression has no short list, and its long list is made up be one instruction only: the unconditional jump.

return (new Expression(resultSymbol, null, longList));

}

The RelationFloating method makes sure the expression holds integral type. If it holds logical type, it is cast from logical type to signed integer type.

private static Expression RelationFloating(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

First, we need to cast the potential logical or integral expressions to floating type (double).

leftExpression = TypeCast.LogicalToFloating(leftExpression);

rightExpression = TypeCast.LogicalToFloating(leftExpression);

Then we cast left value and right from BigInteger to decimal, if necessary.

decimal leftValue;

if (leftExpression.Symbol.Value is BigInteger) {

leftValue = (decimal) ((BigInteger) leftExpression.Symbol.Value);

}

else {

leftValue = (decimal) leftExpression.Symbol.Value;

}

decimal rightValue;

if (rightExpression.Symbol.Value is BigInteger) {

rightValue = (decimal) ((BigInteger) rightExpression.Symbol.Value);

}

else {

rightValue = (decimal) rightExpression.Symbol.Value;

}

We add an unconditional jump instruction to the long list and jump set. The jump set will take the place of true-set of false-set, depending on whether the value is true.

List<MiddleCode> longList = new List<MiddleCode>();

MiddleCode jumpCode = new MiddleCode(MiddleOperator.Jump);

longList.Add(jumpCode);

ISet<MiddleCode> jumpSet = new HashSet<MiddleCode>();

jumpSet.Add(jumpCode);

Since the values are object of the decimal class, we can use the regular relation operators to evaluate the value.

bool resultValue = false;

switch (middleOp) {

case MiddleOperator.Equal:

resultValue = (leftValue == rightValue);

break;

case MiddleOperator.NotEqual:

resultValue = (leftValue != rightValue);

break;

case MiddleOperator.SignedLessThan:

resultValue = (leftValue < rightValue);

break;

case MiddleOperator.SignedLessThanEqual:

resultValue = (leftValue <= rightValue);

break;

case MiddleOperator.SignedGreaterThan:

resultValue = (leftValue > rightValue);

break;

case MiddleOperator.SignedGreaterThanEqual:

resultValue = (leftValue >= rightValue);

break;

}

The resulting symbol has a logical value. The jump set becomes its true-set or false-set, depending on whether the result value is true.

Symbol resultSymbol = resultValue ? (new Symbol(jumpSet, null))

: (new Symbol(null, jumpSet));

return (new Expression(resultSymbol, null, longList));

}

The Logical method evaluate the value of two constant logical expressions, with the or or and operator.

public static Expression Logical(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

if (!IsConstant(leftExpression) || !IsConstant(rightExpression)) {

return null;

}

We make sure that the expressions have logical type by calling ToLogical.

Expression leftLogicalExpression = TypeCast.ToLogical(leftExpression),

rightLogicalExpression = TypeCast.ToLogical(rightExpression);

We decide the value of the expression by looking at their true-sets. If the value is true, the unconditional jump instruction is stored in its true-set. Otherwise, the value is stored in the false-set. In this way, we can conclude that the value is true if the true-set is not empty, and false if the true-set is empty.

bool leftValue = leftLogicalExpression.Symbol.TrueSet.Count > 0,

rightValue = rightLogicalExpression.Symbol.TrueSet.Count > 0;

The long list of the final expression holds only one unconditional jump instruction that is added to the jump set, which will be the true-set or false of the final expression.

List<MiddleCode> longList = new List<MiddleCode>();

MiddleCode jumpCode = new MiddleCode(MiddleOperator.Jump);

longList.Add(jumpCode);

ISet<MiddleCode> jumpSet = new HashSet<MiddleCode>();

jumpSet.Add(jumpCode);

We compare the expression with the given operator and receive the final value.

bool resultValue = false;

switch (middleOp) {

case MiddleOperator.LogicalAnd:

resultValue = leftValue && rightValue;

break;

case MiddleOperator.LogicalOr:

resultValue = leftValue || rightValue;

break;

}

The resulting symbol has a logical value. The jump set becomes its true-set or false-set, depending on whether the result value is true.

Symbol resultSymbol = resultValue ? (new Symbol(jumpSet, null))

: (new Symbol(null, jumpSet));

The final expression does not have a short list, and its long list is made up by the unconditional jump instruction only.

return (new Expression(resultSymbol, null, longList));

}

### Arithmetic Expressions

The Arithmetic method evaluates the value of a constant arithmetic expression.

public static Expression Arithmetic(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

If at least one of the operands is not constant, we just return null.

if (!IsConstant(leftExpression) || !IsConstant(rightExpression)) {

return null;

}

If at least one of the expressions has floating type, we call ArithmeticFloating.

else if (leftExpression.Symbol.Type.IsFloating() ||

rightExpression.Symbol.Type.IsFloating()) {

return ArithmeticFloating(middleOp, leftExpression, rightExpression);

}

If neither of the expressions has floating type, we call ArithmeticIntegral.

else {

return ArithmeticIntegral(middleOp, leftExpression, rightExpression);

}

}

The ArithmeticIntegral method evaluate the value of a constant integral arithmetic expression. It calls in turn the ArithmeticIntegral method that takes two symbols as parameters. The reason for using two methods is that the latter method is called by the middle code optimizer.

private static Expression ArithmeticIntegral(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression){

We cast potential logical expression to integral type (signed integer).

leftExpression = TypeCast.LogicalToIntegral(leftExpression);

rightExpression = TypeCast.LogicalToIntegral(rightExpression);

Symbol symbol = ArithmeticIntegral(middleOp, leftExpression.Symbol,

rightExpression.Symbol);

The integral arithmetic evaluation does not generate any middle code, which is why the expression has neither a short nor a long list.

return (new Expression(symbol, null, null));

}

The integral arithmetic evaluation is a bit more complicated since we must take pointer arithmetic in consideration.

public static Symbol ArithmeticIntegral(MiddleOperator middleOp,

Symbol leftSymbol,

Symbol rightSymbol) {

Type leftType = leftSymbol.Type,

rightType = rightSymbol.Type;

Since this method has been called, we know that none of the expression have floating type, and the potential logical types has been cast to integral types. Therefore, we can assume that their values are BigInteger objects.

BigInteger leftValue = (BigInteger) leftSymbol.Value,

rightValue = (BigInteger) rightSymbol.Value,

resultValue = 0;

In case of addition, we must check whether the type of the left expression is a pointer or array type. In that case we need to multiply the value of the right expression with the size of the pointer or array type.

switch (middleOp) {

case MiddleOperator.BinaryAdd:

if (leftType.IsPointerOrArray()) {

resultValue = leftValue +

(rightValue \* leftType.PointerOrArrayType.Size());

}

We must also check whether the right type of the right expression is a pointer or array, in which case we multiply the left value with the size of the pointer or array type.

else if (leftType.IsPointerOrArray()) {

resultValue = (leftValue \* rightType.PointerOrArrayType.Size()) +

rightValue;

}

If neither the left nor the right expression holds pointer or array types, we simply add the values.

else {

resultValue = leftValue + rightValue;

}

break;

In case of subtraction, to begin with we must check if both the expressions hold pointer or array types. If they do, we subtract their values and divide the difference with size of the pointer or array type.

case MiddleOperator.BinarySubtract:

if (leftType.IsPointerOrArray() && rightType.IsPointerOrArray()) {

resultValue = (leftValue - rightValue) /

leftType.PointerOrArrayType.Size();

}

If the left expression (but not the right expression) holds pointer or array type, we must multiply the right value with the size of the pointer or array type before we subtract t it from the left value.

else if (leftType.IsPointerOrArray()) {

resultValue = leftValue -

(rightValue \* leftType.PointerOrArrayType.Size());

}

If neither the left nor the right expression holds pointer or array type, we simply subtract the values.

else {

resultValue = leftValue - rightValue;

}

break;

In case of multiplication, division, or modulo, bitwise, or shift operators, we simply perform the operation.

case MiddleOperator.SignedMultiply:

case MiddleOperator.UnsignedMultiply:

resultValue = leftValue \* rightValue;

break;

case MiddleOperator.SignedDivide:

case MiddleOperator.UnsignedDivide:

resultValue = leftValue / rightValue;

break;

case MiddleOperator.SignedModulo:

case MiddleOperator.UnsignedModulo:

resultValue = leftValue % rightValue;

break;

case MiddleOperator.BitwiseOr:

resultValue = leftValue | rightValue;

break;

case MiddleOperator.BitwiseXOr:

resultValue = leftValue ^ rightValue;

break;

case MiddleOperator.BitwiseAnd:

resultValue = leftValue & rightValue;

break;

In case of shift operators, we need to type cast the right value from BigInteger to int, simply because the BigInteger operation wants an integer in those cases.

case MiddleOperator.ShiftLeft:

resultValue = leftValue << ((int) rightValue);

break;

case MiddleOperator.ShiftRight:

resultValue = leftValue >> ((int) rightValue);

break;

}

When we have evaluated the value, we need to find its type, which we do by calling the MaxType method in the TypeCast class.

Type maxType = TypeCast.MaxType(leftSymbol.Type, rightSymbol.Type);

Finally, we return a symbol holding the maximum type and the resulting value.

return (new Symbol(maxType, resultValue));

}

The evaluation of a floating constant expression is simpler than the integral expression, since we do not need to take pointer arithmetic in consideration.

private static Expression ArithmeticFloating(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

leftExpression = TypeCast.LogicalToFloating(leftExpression);

rightExpression = TypeCast.LogicalToFloating(rightExpression);

decimal leftValue, rightValue, resultValue = 0;

The value of each expression can be BigInteger or decimal. In the case of BigInteger, we first cast the value to BigInteger and then to decimal.

if (leftExpression.Symbol.Value is BigInteger) {

leftValue = (decimal) ((BigInteger) leftExpression.Symbol.Value);

}

else {

leftValue = (decimal) leftExpression.Symbol.Value;

}

if (rightExpression.Symbol.Value is BigInteger) {

rightValue = (decimal) ((BigInteger) rightExpression.Symbol.Value);

}

else {

rightValue = (decimal) rightExpression.Symbol.Value;

}

Since the values are object of the decimal class, we can user the regular operators.

switch (middleOp) {

case MiddleOperator.BinaryAdd:

resultValue = leftValue + rightValue;

break;

case MiddleOperator.BinarySubtract:

resultValue = leftValue - rightValue;

break;

case MiddleOperator.SignedMultiply:

resultValue = leftValue \* rightValue;

break;

case MiddleOperator.SignedDivide:

resultValue = leftValue / rightValue;

break;

}

When we have evaluated the value, we must decide its type, which we do by calling the MaxType method in the TypeCast class.

Type maxType = TypeCast.MaxType(leftExpression.Symbol.Type,

rightExpression.Symbol.Type);

The final symbol is a symbol with the maximum type and resulting value.

Symbol resultSymbol = new Symbol(maxType, resultValue);

In the floating evaluation case, we need to push the resulting value to the floating value stack.

List<MiddleCode> longList = new List<MiddleCode>();

MiddleCodeGenerator.AddMiddleCode(longList, MiddleOperator.PushFloat,

resultSymbol);

The final expression has no short list, and the long list hold the stack pushing instruction only.

return (new Expression(resultSymbol, null, longList));

}

The LogicalNot method evaluates the logical invers of the expression if it is constant.

public static Expression LogicalNot(Expression expression) {

if (IsConstant(expression)) {

expression = TypeCast.ToLogical(expression);

The resulting symbol is simply the original symbol with its true-set and false-set swapped.

Symbol resultSymbol = new Symbol(expression.Symbol.FalseSet,

expression.Symbol.TrueSet);

return (new Expression(resultSymbol, null, expression.LongList));

}

return null;

}

The Arithmetic method performs a unary arithmetic operation on a constant expression. If the expression is constant and floating ArithmeticFloating is called, and if it is constant and integral ArithmeticIntegral is called.

public static Expression Arithmetic(MiddleOperator middleOp,

Expression expression) {

if (!IsConstant(expression)) {

return null;

}

else if (expression.Symbol.Type.IsFloating()) {

return ArithmeticFloating(middleOp, expression);

}

else {

return ArithmeticIntegral(middleOp, expression);

}

}

The ArithmeticIntegral method performs unary integral arithmetic operations: unary addition and subtraction as well as bitwise not.

private static Expression ArithmeticIntegral(MiddleOperator middleOp,

Expression expression) {

expression = TypeCast.LogicalToIntegral(expression);

BigInteger value = (BigInteger) expression.Symbol.Value,

resultValue = 0;

The unary addition operator does in fact nothing. The unary subtraction operator and bitwise not-operator perform the corresponding operations.

switch (middleOp) {

case MiddleOperator.UnaryAdd:

resultValue = value;

break;

case MiddleOperator.UnarySubtract:

resultValue = -value;

break;

case MiddleOperator.BitwiseNot:

resultValue = ~value;

break;

}

The resulting symbol hold the type of the original expression and the resulting value.

Symbol resultSymbol = new Symbol(expression.Symbol.Type, resultValue);

The final expression does not have short list and long list, since the integral operations do not produce any middle code.

return (new Expression(resultSymbol, null, null));

}

The ArithmeticFloating method performs unary addition and subtraction. In case of addition, nothing happens, a new symbol with the same value is created.

private static Expression ArithmeticFloating(MiddleOperator middleOp,

Expression expression) {

expression = TypeCast.LogicalToFloating(expression);

decimal value = (decimal) expression.Symbol.Value;

Symbol resultSymbol = null;

switch (middleOp) {

case MiddleOperator.UnaryAdd:

resultSymbol = new Symbol(expression.Symbol.Type, value);

break;

case MiddleOperator.UnarySubtract:

resultSymbol = new Symbol(expression.Symbol.Type, -value);

break;

}

In the case of floating operators, we push the resulting value on the floating value stack.

List<MiddleCode> longList = new List<MiddleCode>();

MiddleCodeGenerator.AddMiddleCode(longList, MiddleOperator.PushFloat,

resultSymbol);

The resulting expression has no short list, and the long list holds only the floating stack push instruction.

return (new Expression(resultSymbol, null, longList));

}

## Constant Type Cast

The ConstantCast method cast a constant expression into a new type. If the expression if not constant, we just return null.

public static Expression ConstantCast(Expression sourceExpression,

Type targetType) {

Symbol sourceSymbol = sourceExpression.Symbol, targetSymbol;

Type sourceType = sourceSymbol.Type;

if (!IsConstant(sourceExpression) || sourceType.IsVoid()) {

return null;

}

object sourceValue = sourceSymbol.Value;

List<MiddleCode> longList = new List<MiddleCode>();

If the source type equals the target type, we just return the original expression.

if (targetType.IsVoid()) {

targetSymbol = new Symbol(targetType);

}

If the source and target types have the same size and they both are either integral, array, or pointers, or they both are floating, we just return a new expression with the target symbol and the short list and long list of the source expression.

else if (sourceType.IsIntegralPointerArrayStringOrFunction() &&

targetType.IsFloating()) {

decimal targetValue = ((decimal) ((BigInteger) sourceValue));

targetSymbol = new Symbol(targetType, targetValue);

}

else if (sourceType.IsFloating() &&

targetType.IsIntegralPointerOrArray()) {

BigInteger targetValue = ((BigInteger) ((decimal) sourceValue));

targetSymbol = new Symbol(targetType, targetValue);

}

If the source type is logical and the target type is in integral, array, or pointer. We look into the true and false-sets of the expression.

If the true-set is not empty, the expression is true and we return a new expression with the value one. We do not need to exam the false-set, since we know that the expression is constant, in which case one of the sets is always empty.

If the true-set is empty, we return a new expression with the value zero, since the false-set must be non-empty.

else if (sourceType.IsLogical() &&

targetType.IsIntegralPointerOrArray ()) {

bool isTrue = (sourceSymbol.TrueSet.Count > 0);

BigInteger targetValue = isTrue ? BigInteger.One : BigInteger.Zero;

targetSymbol = new Symbol(targetType, targetValue);

}

else if (sourceType.IsLogical() && targetType.IsFloating()) {

bool isTrue = (sourceSymbol.TrueSet.Count > 0);

decimal targetValue = isTrue ? decimal.One : decimal.Zero;

targetSymbol = new Symbol(targetType, targetValue);

}

If the source type is integral, array, pointer, or floating, and the target type is logical, we start by constructing a jump set holding one unconditional jump instruction.

If the value of the source symbol is zero, we create a target symbol with jump set as its false-set. If it not zero, we create the target symbol with the jump set as its true-set.

The long list holds the unconditional jump instruction of the jump set.

else if (sourceType.IsArithmeticPointerArrayStringOrFunction() &&

targetType.IsLogical()) {

bool isTrue = !sourceValue.Equals(BigInteger.Zero) &&

!sourceValue.Equals(decimal.Zero);

MiddleCode gotoCode = new MiddleCode(MiddleOperator.Jump);

longList.Add(gotoCode);

ISet<MiddleCode> trueSet = new HashSet<MiddleCode>(),

falseSet = new HashSet<MiddleCode>();

if (isTrue) {

trueSet.Add(gotoCode);

}

else {

falseSet.Add(gotoCode);

}

targetSymbol = new Symbol(trueSet, falseSet);

}

else {

targetSymbol = new Symbol(targetType, sourceValue);

}

if (targetType.IsFloating()) {

longList.Add(new MiddleCode(MiddleOperator.PushFloat, targetSymbol));

}

return (new Expression(targetSymbol, null, longList));

}

Finally, we have reached the point where both the source type and target type are integral, array, pointer, or floating. The only thing that remains to be done is to make sure the final value is an object of the correct class: BigInteger or decimal.

If the source type is integral, array, or pointer, and the target type is floating, we need to type cast the value from BigInteger to decimal.

if (sourceType.IsIntegralPointerArrayStringOrFunction () &&

targetType.IsFloating()) {

targetSymbol =

new Symbol(targetType, (decimal) ((BigInteger) sourceValue));

}

In the same way, if the source type is floating and the target type is integral, array, or pointer, we need to type cast the value from decimal to BigInteger.

else if (sourceType.IsFloating() &&

targetType.IsIntegralPointerOrArray()) {

targetSymbol =

new Symbol(targetType, (BigInteger) ((decimal) sourceValue));

}

Finally, if the both the source and target types are integral, array, or pointer, or if they both are floating, we just create a new symbol with target type and source value.

else {

targetSymbol = new Symbol(targetType, sourceValue);

}

List<MiddleCode> longList = new List<MiddleCode>();

if (targetType.IsFloating()) {

MiddleCodeGenerator.

AddMiddleCode(longList, MiddleOperator.PushFloat, targetSymbol);

}

return (new Expression(targetSymbol, null, longList));

}

}

### Constant Value

The Value method takes a symbol and returns a static symbol.

public static StaticSymbol Value(Symbol symbol) {

return Value(symbol.UniqueName, symbol.Type, symbol.Value);

}

public static StaticSymbol Value(string uniqueName, Type type,

object value) {

List<MiddleCode> middleCodeList = new List<MiddleCode>();

If the value is not null, we add the initializer instruction to the code list.

if (value != null) {

middleCodeList.Add(new MiddleCode(MiddleOperator.Initializer,

type.Sort, value));

}

If the value is null, we add the initializer ero instruction, with the type size.

else {

middleCodeList.Add(new MiddleCode(MiddleOperator.InitializerZero,

type.Size()));

}

We translate the middle code instructions to assembly code.

List<AssemblyCode> assemblyCodeList = new List<AssemblyCode>();

AssemblyCodeGenerator.GenerateAssembly(middleCodeList,

assemblyCodeList);

For the Linux target machine, we generate a list of text holding the final assembly code. For Windows target code generation, se Chapter 13.

if (Start.Linux) {

List<string> textList = new List<string>();

textList.Add("section .data");

textList.Add("\n" + uniqueName + ":");

ISet<string> externSet = new HashSet<string>();

AssemblyCodeGenerator.LinuxTextList(assemblyCodeList, textList,

externSet);

return (new StaticSymbolLinux(uniqueName, textList, externSet));

}

if (Start.Windows) {

// ...

}

return null;

}

}

}

# Static Address Expression

If an expression is not constant, the next step is to decide whether it is static.

### Static Value and Address

The StaticValue and StaticAddress classes are identical sub classes of StaticBase, which holds a name and an offset. In the end, only static addresses are allowed, but static values can hold intermediate values during the parsing. For instance, in the static address &a[3], a[3] is temporary stored as a static value.

StaticBase.cs

namespace CCompiler {

public abstract class StaticBase {

private string m\_uniqueName;

private int m\_offset;

public StaticBase(string name, int offset) {

m\_uniqueName = name;

m\_offset = offset;

}

public string UniqueName {

get { return m\_uniqueName; }

}

public int Offset {

get { return m\_offset; }

}

public override string ToString() {

if (m\_offset > 0) {

return m\_uniqueName + " + " + m\_offset;

}

else if (m\_offset < 0) {

return m\_uniqueName + " - " + (-m\_offset);

}

else {

return m\_uniqueName;

}

}

}

public class StaticValue : StaticBase {

public StaticValue(string name, int offset)

:base(name, offset) {

// Empty.

}

}

public class StaticAddress : StaticBase {

public StaticAddress(string name, int offset)

:base(name, offset) {

// Empty.

}

}

}

### Static Expression

A static expression is an expression that is located at a specific place in the memory. The address can be identified by its name and offset. The StaticExpression class holds the two methods Binary and Unary, which takes a binary or unary expression and returns a static expression if there is one, or null if there is no one.

StaticExpression.cs

using System.Numerics;

namespace CCompiler {

public class StaticExpression {

The Binary method exams the expressions if the operator is binary addition or subtraction, index, or dot.

public static Expression Binary(MiddleOperator middleOp,

Expression leftExpression,

Expression rightExpression) {

Type leftType = leftExpression.Symbol.Type,

rightType = rightExpression.Symbol.Type;

object leftValue = leftExpression.Symbol.Value,

rightValue = rightExpression.Symbol.Value;

In the addition case, the operand values must be a static address and a constant integer value, or an extern or static array and a constant value. For instance, &i + 2, 2 + &i, a + 2 or 2 + a, where a is an array and i an integer.

In case of left array or static address and an integral right value, we GenerateAddition to generate the resulting static address.

switch (middleOp) {

case MiddleOperator.BinaryAdd: // &i + 2, a + 2

if (((leftValue is StaticAddress) ||

(leftExpression.Symbol.IsExternOrStatic() &&

leftType.IsArray())) && (rightValue is BigInteger)) {

return GenerateAddition(leftExpression.Symbol,

(BigInteger) rightValue);

}

In case of an integral left value and a right array or static address, we just swap the operands in the GenerateAddition call.

else if ((leftValue is BigInteger) && // 2 + &i, 2 + a

((rightValue is StaticAddress) || (rightType.IsArray() &&

rightExpression.Symbol.IsExternOrStatic()))) {

return GenerateAddition(rightExpression.Symbol,

(BigInteger) leftValue);

}

break;

In the subtraction case, the left operand must a static address as value or be an extern of static array, and the right operand must be a constant integer value. For instance, &i - 2 or a - 2.

case MiddleOperator.BinarySubtract: // &i - 2, a - 2

if (((leftValue is StaticAddress) ||

(leftExpression.Symbol.IsExternOrStatic() &&

leftType.IsArray())) &&

(rightValue is BigInteger)) {

return GenerateAddition(leftExpression.Symbol,

-((BigInteger) rightValue));

}

break;

In the index case, the operands must be an extern or static array or a static address, and a constant integer value. We call GenerateIndex to generate the static address.

case MiddleOperator.Index:

if (((leftValue is StaticAddress) || (leftType.IsArray() &&

leftExpression.Symbol.IsExternOrStatic())) &&

(rightValue is BigInteger)) { // a[2]

return GenerateIndex(leftExpression.Symbol,

(BigInteger) rightValue);

}

In C, where are allowed to write an array index as i[a], where i is an integer and a is an array. In that case, we just swap the parameters in GenerateIndex.

else if ((leftValue is BigInteger) && ((rightType.IsArray() &&

rightExpression.Symbol.IsExternOrStatic()) ||

(rightValue is StaticAddress))) {

return GenerateIndex(rightExpression.Symbol,

(BigInteger) leftValue);

}

break;

In the dot case, the operands must be an extern or static struct or union, or a static address.

case MiddleOperator.Dot:

if (leftExpression.Symbol.IsExternOrStatic()) { // XXX

object resultValue =

new StaticValue(leftExpression.Symbol.UniqueName,

rightExpression.Symbol.Offset); // s.i

Symbol resultSymbol = new Symbol(leftType, resultValue);

return (new Expression(resultSymbol, null, null));

}

break;

}

return null;

}

The GenerateAddition method generates a static address for an addition expression.

private static Expression GenerateAddition(Symbol symbol,

BigInteger value) {

int offset = ((int) value) \* symbol.Type.PointerOrArrayType.Size();

StaticAddress resultValue;

if (symbol.Value is StaticAddress) {

StaticAddress staticAddress = (StaticAddress) symbol.Value;

resultValue = new StaticAddress(staticAddress.UniqueName,

staticAddress.Offset + offset);

}

else {

resultValue = new StaticAddress(symbol.UniqueName, offset);

}

Symbol resultSymbol = new Symbol(symbol.Type, resultValue);

return (new Expression(resultSymbol, null, null));

}

The GenerateIndex method generates the static address for an index expression.

private static Expression GenerateIndex(Symbol symbol,

BigInteger value) {

int offset = ((int) value) \* symbol.Type.ArrayType.Size();

StaticValue resultValue;

if (symbol.Value is StaticAddress) {

StaticAddress staticAddress = (StaticAddress) symbol.Value;

resultValue = new StaticValue(staticAddress.UniqueName,

staticAddress.Offset + offset);

}

else {

resultValue = new StaticValue(symbol.UniqueName, offset);

}

Symbol resultSymbol = new Symbol(symbol.Type, resultValue);

return (new Expression(resultSymbol, null, null));

}

Finally, we have to unary case. There is only one relevant operator: the address operator.

public static Expression Unary(MiddleOperator middleOp,

Expression expression) {

Symbol symbol = expression.Symbol;

If the symbol in the address case is a static value, we create a static address with the same name and offset.

switch (middleOp) {

case MiddleOperator.Address:

if (symbol.Value is StaticValue) { // &a[i], &s.i

StaticValue staticValue = (StaticValue) symbol.Value;

StaticAddress staticAddress =

new StaticAddress(staticValue.UniqueName, staticValue.Offset);

Symbol resultSymbol =

new Symbol(new Type(symbol.Type), staticAddress);

return (new Expression(resultSymbol, null, null));

}

If the symbol is extern or static, we create a static address, with the symbol name and offset zero.

else if (symbol.IsExternOrStatic()) { // &i

StaticAddress staticAddress =

new StaticAddress(symbol.UniqueName, 0);

Symbol resultSymbol =

new Symbol(new Type(symbol.Type), staticAddress);

return (new Expression(resultSymbol, null, null));

}

break;

}

}

return null;

}

}

}

# Initialization

In C, it is possible to initialize simple and compound variables. Therefore, we need to check that the initialized value has the correct type. There are two kinds of initialization: static and auto. They perform the same task: matching a type against an initializer. However, static initialization results in a memory block while auto initialization results in a sequence of assignment instructions. The initializer may be an expression or a recursive list of expressions; that is, the list may hold sub lists.

There are several cases to consider:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Initializer** | **Example** |
| Pointer | Address | static int array[3];  static int\* p = &a[2]; |
| Pointer to signed or unsigned char | String | static char \*p = “Hello”; |
| Pointer | List | int \*p = {1,2,3}; |
| Array | String | char s[] = “World”; |
| Array | List | Char st[] = {‘a’, ’b’, ‘c’}; |
| Struct | List |  |
| Union | List |  |
| Integral or Pointer |  |  |
| Floating |  |  |

If the variable is a pointer and the initializer is an address we add its offset to the block and store the name of the address in the access map, the address value will later be looked up and added by the linker.

If the variable type is a pointer to a (signed or unsigned) character and the initializator is a string, the address is stored in the access map and a zero address is stored in the block, it will later be properly looked up and set by the linker.

### Auto Initialization

Auto initialization occurs when an auto or register variable becomes initialized. Since the initialization value can be non-constant, no memory block is created. Instead, a sequence of assignment instructions is added to the middle code.

GenerateAutoInitializer.cs

using System.Numerics;

using System.Collections.Generic;

namespace CCompiler {

class GenerateAutoInitializer {

public static List<MiddleCode> GenerateAuto(Symbol toSymbol,

object fromInitializer) {

Assert.ErrorXXX((fromInitializer is Expression) ||

(fromInitializer is List<object>));

Type toType = toSymbol.Type;

List<MiddleCode> codeList = new List<MiddleCode>();

If the initializer is an expression, we have two cases. If the type is an array of characters and the initializer is a string, we change the initializer from a string to a list of characters, and call GenerateAutoInitializer recursively with the list of characters instream of the string. In C, when initializing an array of characters, a string or a list of characters are equivalent. For instance, the initializations char s[] = "Hi" and char s[] = {'H', 'i', '\0'} are equivalent. Note the terminating zero-character is added implicitly and the string case.

if (fromInitializer is Expression) {

Expression fromExpression = (Expression) fromInitializer;

if (toType.IsArray() && toType.ArrayType.IsChar() &&

fromExpression.Symbol.Type.IsString()) {

string text = ((string) fromExpression.Symbol.Value) + "\0";

List<object> list = new List<object>();

foreach (char c in text) {

Symbol charSymbol =

new Symbol(toType.ArrayType, (BigInteger) ((int) c));

Expression charExpression = new Expression(charSymbol, null, null);

list.Add(charExpression);

}

return GenerateAuto(toSymbol, list);

}

In all other cases, we type cast the expression into the defined type and generate middle code instructions. In case of a floating value, we pop the value from the floating value stack to the symbol to be initialized. In case of any other type we generate an assignment instruction.

else {

fromExpression = TypeCast.ImplicitCast(fromExpression, toType);

codeList.AddRange(fromExpression.LongList);

if (toSymbol.Type.IsFloating()) {

codeList.Add(new MiddleCode(MiddleOperator.PopFloat, toSymbol));

}

else {

if (fromExpression.Symbol.Type.IsStructOrUnion()) {

codeList.Add(new MiddleCode(MiddleOperator.AssignInit,

toSymbol, fromExpression.Symbol));

}

codeList.Add(new MiddleCode(MiddleOperator.Assign, toSymbol,

fromExpression.Symbol));

}

}

}

If the initializer is not an expression is must be a list of expressions or other lists, recursively. In that case the type must be an array, a struct, or a union, since only arrays, structs, and unions can be initialized with a list.

else {

Assert.Error(toType.IsArray() ||toType.IsStructOrUnion(),

toType, Message.

Only\_array\_struct\_or\_union\_can\_be\_initialized\_by\_a\_list);

List<object> fromList = (List<object>) fromInitializer;

If the type is an array, we set its size if it has not yet been defined or check that the list size does not exceed the array size if the size has been defined. Note that in the case of an array we do not add a terminating zero character.

switch (toType.Sort) {

case Sort.Array: {

fromList = ModifyInitializer.ModifyArray(toType, fromList);

if (toType.ArraySize == 0) {

toType.ArraySize = fromList.Count;

}

else {

Assert.Error(fromList.Count <= toType.ArraySize,

toType, Message.Too\_many\_initializers);

}

We iterate through the list and, for each element in the list, define a symbol for the array index and call GenerateAuto recursively, so that potential sub list are properly processed. If the list does not cover the whole array, the remaining part of the array remains uninitialized. The offset of each index symbol is sum of the offset of the symbol and the offset of the index symbol. Note that we must multiply the index of the array with the size of the array type to determine the correct index. However, it is not strictly necessary give the index symbol a name. That is for readability reasons only.

for (int index = 0; index < fromList.Count; ++index) {

Symbol indexSymbol = new Symbol(toType.ArrayType);

indexSymbol.Offset = toSymbol.Offset +

(index \* toType.ArrayType.Size());

indexSymbol.Name = toSymbol.Name + "[" + index + "]";

codeList.AddRange(GenerateAuto(indexSymbol, fromList[index]));

}

}

break;

In case of a struct or a union, we check that the list size does not exceed its allowed size. In case of a struct, the list size must not exceed the number of struct members. In case of a union, the list must hold exactly on element.

case Sort.Struct:

case Sort.Union: {

List<Symbol> memberList = toType.MemberList;

Assert.Error((toType.IsStruct() &&

(fromList.Count <= memberList.Count)) ||

(toType.IsUnion() && (fromList.Count == 1)),

toType, Message.Too\_many\_initializers);

Like the array case, we iterate through the list and, for each element in the list, create a symbol for the member and call GenerateAuto recursively, so that potential sub lists are properly processed. The offset of each member symbol is the sum of the offset of the struct or union and the member.

IEnumerator<Symbol> enumerator = memberList.GetEnumerator();

foreach (object fromInitializor in fromList) {

enumerator.MoveNext();

Symbol memberSymbol = enumerator.Current;

Symbol subSymbol = new Symbol(memberSymbol.Type);

subSymbol.Name = toSymbol.Name + "." + memberSymbol.Name;

subSymbol.Offset = toSymbol.Offset + memberSymbol.Offset;

codeList.AddRange(GenerateAuto(subSymbol, fromInitializor));

}

}

break;

}

}

return codeList;

}

}

}

### Static Initialization

Static initialization occurs when a static variable becomes initialized. The initialization value has to be constant and known at compile time. No code is generated, instead a memory block holding the value is created.

The GenerateStaticInitializer class basically performs the same task as the GenerateAutoInitializer class above: it interprets the initialization and compare it to the defined type. However, it does not generate assignment instruction, but rather initialization instructions. One difference compared to the auto initialization case is that if the array, struct, or union is not completely initialized we need add zeros for the remaining part.

GenerateStaticInitializer.cs

using System.Collections.Generic;

namespace CCompiler {

public class GenerateStaticInitializer {

public static List<MiddleCode> GenerateStatic(Type toType,

object fromInitializer) {

Assert.ErrorXXX((fromInitializer is Expression) ||

(fromInitializer is List<object>));

List<MiddleCode> codeList = new List<MiddleCode>();

If the initializer is an expression, we have two special cases: that the defined type is a character array and the initializer is a string, the defined type is a pointer and the initializer is an array, a function, or a string.

if (fromInitializer is Expression) {

Expression fromExpression = (Expression) fromInitializer;

Symbol fromSymbol = fromExpression.Symbol;

Assert.Error(fromSymbol.IsExternOrStatic(), fromSymbol,

Message.Non\_\_static\_initializer);

Type fromType = fromSymbol.Type;

If the defined type is a character array and the initializer is a string, we first check the size of the string. If the array size is undefined, we set its size to the size of the string, plus one for the terminating zero-character. If the array size is defined, we check that the string (including the terminating zero-character) fits withing the array.

if (toType.IsArray() && toType.ArrayType.IsChar() &&

fromType.IsString()) {

string text = (string) fromSymbol.Value;

Note that we add space for an extra character, the terminating zero-character.

if (toType.ArraySize == 0) {

toType.ArraySize = text.Length + 1;

}

Note that the string length must be strictly less than the array size, for the terminating zero-character to fit in the string.

else {

Assert.Error(text.Length < toType.ArraySize,

toType, Message.Too\_many\_initializers);

}

codeList.Add(new MiddleCode(MiddleOperator.Initializer,

fromSymbol.Type.Sort, text));

}

If the defined type is a pointer and the initializer is an array, a function, or a string, we create a static address and add an initialization middle code instruction.

else if (toType.IsPointer() && fromType.IsArrayFunctionOrString()) {

Assert.ErrorXXX((fromType.IsString() && toType.PointerType.IsChar())

||(fromType.IsArray() &&

fromType.ArrayType.Equals(toType.PointerType)) ||

(fromType.IsFunction() &&

fromType.Equals(toType.PointerType)));

StaticAddress staticAddress =

new StaticAddress(fromSymbol.UniqueName, 0);

codeList.Add(new MiddleCode(MiddleOperator.Initializer,

toType.Sort, staticAddress));

}

In all other cases, we type cast the expression into the defined type, check that the expression is constant, and add an initialization middle code instruction.

else {

Expression toExpression =

TypeCast.ImplicitCast(fromExpression, toType);

Symbol toSymbol = toExpression.Symbol;

If the value is not null, the expression is constant. However, the value may be a static address, which we consider to be constant in this context.

Assert.Error(toSymbol.Value != null, toSymbol,

Message.Non\_\_constant\_expression);

codeList.Add(new MiddleCode(MiddleOperator.Initializer,

toSymbol.Type.Sort, toSymbol.Value));

}

}

If the initializer is a list, the defined type must be an array, a struct, or a union.

else {

Assert.Error(toType.IsArray() || toType.IsStructOrUnion(),

toType, Message.

Only\_array\_struct\_or\_union\_can\_be\_initialized\_by\_a\_list);

List<object> fromList = (List<object>) fromInitializer;

In case of an array, we set the array size to the list size if undefined. If the array is defined, we check that the list size does not exceed the array size.

switch (toType.Sort) {

case Sort.Array: {

fromList = ModifyInitializer.ModifyArray(toType, fromList);

if (toType.ArraySize == 0) {

toType.ArraySize = fromList.Count;

}

else {

Assert.Error(fromList.Count <= toType.ArraySize,

toType, Message.Too\_many\_initializers);

}

We iterate the array values and call GenerateStatic for each value. In the static case we do not need to create index symbol since the code to be generated are initializations rather than assignments.

foreach (object value in fromList) {

codeList.AddRange(GenerateStatic(toType.ArrayType, value));

}

However, in the static case we must add an instruction for the initialization of the potential remaining part of the array. The InitializerZero instruction adds a sequence of zero-bytes.

int restSize = toType.Size() -

(fromList.Count \* toType.ArrayType.Size());

if (restSize > 0) {

codeList.Add(new MiddleCode(MiddleOperator.InitializerZero,

restSize));

}

}

break;

Like the auto case, the initializer list size must not exceed the number of members in a struct. The initializer list must hold exactly one element in case of a union.

case Sort.Struct:

case Sort.Union: {

List<Symbol> memberList = toType.MemberList;

Assert.Error((toType.IsStruct() &&

(fromList.Count <= memberList.Count)) ||

(toType.IsUnion() && (fromList.Count == 1)),

toType, Message.Too\_many\_initializers);

Like the auto case, we iterate through the initialization list and call GenerateStatic recursively for each element in the list. However, in the static case we must sum the size of the initializer list in order to add zero-bytes if the list does not cover the struct or union.

int size = 0;

IEnumerator<Symbol> enumerator = memberList.GetEnumerator();

foreach (object fromInitializor in fromList) {

enumerator.MoveNext();

Symbol memberSymbol = enumerator.Current;

codeList.AddRange(GenerateStatic(memberSymbol.Type,

fromInitializor));

size += memberSymbol.Type.Size();

}

int restSize = toType.Size() - size;

if (restSize > 0) {

codeList.Add(new MiddleCode(MiddleOperator.InitializerZero,

restSize));

}

}

break;

}

}

return codeList;

}

}

}

### **Modify Initializer**

The ModifyInitializer class changes an initialization list into the form of the defined type. For instance: int [3][2] = {1, 2, 3, 4, 5, 6} is changed to int [3][2] = {{1, 2}, {3, 4}, {5, 6}}.

ModifyInitializer.cs

using System;

using System.IO;

using System.Collections.Generic;

namespace CCompiler {

class ModifyInitializer {

public static List<object> ModifyArray(Type type, List<object> list) {

First, we define the dimension-to-size map; that is, each dimension in the defined array is assigned an array size. For instance, in int [2][3][4] dimension 1 has size 4, dimension 2 has size 3, and dimension 3 has size 2. Dimension zero refers to the final array type, and its size is always zero. In int [][3][4], dimension 3 is undefined and set to zero, but that does not matter since we do not exam the highest dimension.

IDictionary<int,int> dimensionToSizeMap = new Dictionary<int,int>();

int maxDimension = DimensionToSizeMap(type, dimensionToSizeMap);

Then we define the initializer-to-dimension map; that is, each list and sub list in the initializer is assigned a dimension.

IDictionary<object,int> initializerToDimensionMap =

new Dictionary<object,int>();

InitializerToDimensionMap(list, initializerToDimensionMap);

Then we iterate through the dimensions of the array, from dimension one to the second-highest dimension, inclusive. For each dimension we define the total list, where each element holds the current dimension.

for (int dimension = 1; dimension < maxDimension; ++dimension) {

List<object> totalList =

new List<object>(), currentList = new List<object>();

int arraySize = dimensionToSizeMap[dimension];

Assert.ErrorXXX(arraySize > 0);

foreach (object member in list) {

if (initializerToDimensionMap[member] < dimension) {

currentList.Add(member);

}

else {

if (currentList.Count > 0) {

initializerToDimensionMap[currentList] = dimension;

totalList.Add(currentList);

}

totalList.Add(member);

currentList = new List<object>();

}

if (currentList.Count == arraySize) {

initializerToDimensionMap[currentList] = dimension;

totalList.Add(currentList);

currentList = new List<object>();

}

}

if (currentList.Count > 0) {

initializerToDimensionMap[currentList] = dimension;

totalList.Add(currentList);

}

list = totalList;

}

return list;

}

The DimensionToSizeMap method assigns the array size of each dimension in the nested array.

private static int DimensionToSizeMap(Type type,

IDictionary<int,int> dimensionToSizeMap) {

if (type.IsArray()) {

int dimension =

DimensionToSizeMap(type.ArrayType, dimensionToSizeMap) + 1;

dimensionToSizeMap[dimension] = type.ArraySize;

}

return 0;

}

The InitializerToDimensionMap method assigns the dimension to each list in the nested list.

private static int InitializerToDimensionMap(object initializer,

IDictionary<object,int> initializerToDimensionMap) {

if (initializer is List<object>) {

List<object> list = (List<object>) initializer;

int maxDimension = 0;

If the initializer is a list, we iterate through the list, and assign the list the dimension of the sub list with the highest dimension, plus one.

foreach (object member in list) {

int dimension =

InitializerToDimensionMap(member, initializerToDimensionMap);

maxDimension = Math.Max(maxDimension, dimension);

}

initializerToDimensionMap[list] = maxDimension + 1;

return maxDimension + 1;

}

return 0;

}

}

}

# Middle Code Optimization

When the middle code has been generated, we need to perform several optimizations since the code may be ineffective. Some ineffective parts may be introduced by the programmer, while other parts are introduced by the parser. The MiddleCodeOptimization class takes care of the optimization. Since the optimization methods may reveal new optimization opportunities, we need to repeat until we do not detect any more opportunities. The field m\_update is set to true if we find an optimization opportunity.

MiddleCodeOptimizor.cs

using System.Numerics;

using System.Collections.Generic;

The m\_update field is set to true each time an optimization has occurred, and the middle code list is stored in m\_middleCodeList.

namespace CCompiler {

public class MiddleCodeOptimizer {

private bool m\_update;

private List<MiddleCode> m\_middleCodeList;

public MiddleCodeOptimizer(List<MiddleCode> middleCodeList) {

m\_middleCodeList = middleCodeList;

}

First of all, we need to change the addresses of jump instructions, from middle code objects to their index in the code list, by calling ObjectToIntegerAddresses.

public void Optimize() {

ObjectToIntegerAddresses();

Then follows the main loop of the optimization process. We continue to iterate as long as there are optimization possibilities; that is, as long as m\_update is true. The m\_update field is set to false at the beginning, and will be set to true in case of an optimization.

do {

m\_update = false;

ClearGotoNextStatements();

ClearDoubleRelationStatements();

TraceGotoChains();

ClearUnreachableCode();

RemovePushPop();

MergePopPushToTop();

MergeTopPopToPop();

MergeBinary();

//MergeDoubleAssign(); // XXX

SematicOptimization();

OptimizeRelation();

OptimizeCommutative();

OptimizeBinary();

CheckIntegral(); // XXX

CheckFloating(); // XXX

RemoveClearedCode();

} while (m\_update);

}

### Object to Integer Addresses

When we generated the middle code, we use middle code objects as target for jump instructions. The first we need to do is to change the middle code targets to integer targets. The reason for this is that the optimization process of this chapter will result in the removal of middle code instruction, among which some may be jump targets.

public void ObjectToIntegerAddresses() {

We start by defining the address map, which we load with the index of each instruction.

IDictionary<MiddleCode,int> addressMap =

new Dictionary<MiddleCode,int>();

for (int index = 0; index < m\_middleCodeList.Count; ++index) {

addressMap.Add(m\_middleCodeList[index], index);

}

We then iterate through the middle code list and, with the help of the address map, change the targets from middle code instructions to integer values.

for (int index = 0; index < m\_middleCodeList.Count; ++index) {

MiddleCode sourceCode = m\_middleCodeList[index];

if (sourceCode.IsGoto() || sourceCode.IsCarry() ||

sourceCode.IsRelation()) {

Assert.ErrorXXX(sourceCode[0] is MiddleCode);

MiddleCode targetCode = (MiddleCode) sourceCode[0];

Assert.ErrorXXX(addressMap.ContainsKey(targetCode));

sourceCode[0] = addressMap[targetCode];

}

}

}

### Jump Next Instructions

In some cases, there may be a jump instruction that just jumps to the next instruction. Those instructions are meaningless and shall be removed.

1. goto 2

2. ...

private void ClearGotoNextStatements() {

for (int index = 0; index < (m\_middleCodeList.Count - 1); ++index) {

MiddleCode middleCode = m\_middleCodeList[index];

if (middleCode.IsRelationCarryOrGoto()) {

int target = (int) middleCode[0];

When we encounter a jump instruction that jumps to the next instruction, we remove it

if (target == (index + 1)) {

middleCode.Clear();

m\_update = true;

}

}

}

}

### Next-Double Jump Statements

A conditional jump instruction that jumps two steps ahead and is followed by an unconditional jump instruction can be modified in that we reverse the condition and the target.

|  |  |  |
| --- | --- | --- |
| 1. if a < b goto 3  2. goto 10  3. ...  (a) Before optimization | 1. if a >= b goto 10  2. *removed*  3. ...  (b) After optimization |  |

To begin with, we need the inverse map to change the condition. The swap map is used in OptimizeRelation below. The reason we define it at this point in the code is that we cannot have more than one static area in the same class.

public static IDictionary<MiddleOperator, MiddleOperator> m\_inverseMap =

new Dictionary<MiddleOperator, MiddleOperator>() {

{MiddleOperator.Equal, MiddleOperator.NotEqual},

{MiddleOperator.NotEqual, MiddleOperator.Equal},

{MiddleOperator.Carry, MiddleOperator.NotCarry},

{MiddleOperator.NotCarry, MiddleOperator.Carry},

{MiddleOperator.SignedLessThan,

MiddleOperator.SignedGreaterThanEqual},

{MiddleOperator.SignedLessThanEqual,

MiddleOperator.SignedGreaterThan},

{MiddleOperator.SignedGreaterThan,

MiddleOperator.SignedLessThanEqual},

{MiddleOperator.SignedGreaterThanEqual,

MiddleOperator.SignedLessThan},

{MiddleOperator.UnsignedLessThan,

MiddleOperator.UnsignedGreaterThanEqual},

{MiddleOperator.UnsignedLessThanEqual,

MiddleOperator.UnsignedGreaterThan},

{MiddleOperator.UnsignedGreaterThan,

MiddleOperator.UnsignedLessThanEqual},

{MiddleOperator.UnsignedGreaterThanEqual,

MiddleOperator.UnsignedLessThan}

};

We iterate through the middle code list, from the first to the next to last instruction since we inspect the current instruction and the instruction following it.

private void ClearDoubleRelationStatements() {

for (int index = 0; index < (m\_middleCodeList.Count - 1); ++index) {

MiddleCode thisCode = m\_middleCodeList[index],

nextCode = m\_middleCodeList[index + 1];

if ((thisCode.IsRelation() || thisCode.IsCarry()) &&

nextCode.IsGoto()) {

int target1 = (int) thisCode[0],

target2 = (int) nextCode[0];

If the instruction jumps over the next instruction, we use the inverse map to change the current instruction, and we clear the next instruction.

if (target1 == (index + 2)) {

MiddleOperator operator1 = thisCode.Operator;

thisCode.Operator = m\_inverseMap[operator1];

thisCode[0] = target2;

nextCode.Clear();

m\_update = true;

}

}

}

}

### Jump Chains

A jump chain is a sequence of unconditional jump instructions where each instruction in the chain, except the last one, jumps to another unconditional jump instruction. It would be more effective if they all jump to the same target as the last jump instruction.

|  |  |  |
| --- | --- | --- |
| 1. goto 3  2. ..  3. goto 5  4. ...  5. goto 7  6. ...  7. a = 1  (a) Before optimization | 1. goto 7  2. ..  3. goto 7  4. ...  5. goto 7  6. ...  7. a = 1  (b) After optimization |  |

First, we trace all the unconditional jump instructions by calling TraceGotoChains, that in turn calls TraceGoto. Then we change all the jump targets.

private void TraceGotoChains() {

for (int index = 1; index < m\_middleCodeList.Count; ++index) {

MiddleCode middleCode = m\_middleCodeList[index];

if (middleCode.IsRelationCarryOrGoto()) {

ISet<int> sourceSet = new HashSet<int>();

sourceSet.Add(index);

For each jump instruction we trace the goto chain and if the first and last jump instructions differs, we replace all the jumps in the chain with the last jump.

int firstTarget = (int) middleCode[0];

int finalTarget = TraceGoto(firstTarget, sourceSet);

if (firstTarget != finalTarget) {

foreach (int source in sourceSet) {

MiddleCode sourceCode = m\_middleCodeList[source];

sourceCode[0] = finalTarget;

}

m\_update = true;

}

}

}

}

The TraceGoto method follows the jump instructions as long as they continue to jump to new targets.

private int TraceGoto(int target, ISet<int> sourceSet) {

MiddleCode objectCode = m\_middleCodeList[target];

if (!sourceSet.Contains(target) && objectCode.IsGoto()) {

sourceSet.Add(target);

int nextTarget = (int) objectCode[0];

return TraceGoto(nextTarget, sourceSet);

}

else {

return target;

}

}

### Remove Unreachable Code

Code that is not reachable from the first instruction of the function shall be removed. We call SearchReachableCode that follows all conditional and unconditional jumps instructions and save their line numbers in the visited set. We also report an error if we reach the last instruction of a function that do not return void.

|  |  |  |
| --- | --- | --- |
| if (1)  a = 1;  else  b = 2;  (a) C code | 1. goto 1  2. a = 1  3. goto 5  4. b = 2  5. ...  (b) Before optimization | 1. *removed*  2. a = 1  3. *removed*  4. *removed*  5. ...  (b) After optimization |

private void ClearUnreachableCode() {

ISet<MiddleCode> visitedSet = new HashSet<MiddleCode>();

SearchReachableCode(0, visitedSet);

for (int index = 0; index < (m\_middleCodeList.Count - 1); ++index) {

MiddleCode middleCode = m\_middleCodeList[index];

if (!visitedSet.Contains(middleCode)) {

m\_middleCodeList[index].Clear();

m\_update = true;

}

}

}

private void SearchReachableCode(int index, ISet<MiddleCode> visitedSet) {

for (; index < m\_middleCodeList.Count; ++index) {

MiddleCode middleCode = m\_middleCodeList[index];

if (visitedSet.Contains(middleCode)) {

return;

}

visitedSet.Add(middleCode);

if (middleCode.IsRelation() || middleCode.IsCarry()) {

int target = (int) middleCode[0];

SearchReachableCode(target, visitedSet);

}

else if (middleCode.IsGoto()) {

int target = (int) middleCode[0];

SearchReachableCode(target, visitedSet);

return;

}

else if (middleCode.Operator == MiddleOperator.Return) {

if (m\_middleCodeList[index + 1].Operator == MiddleOperator.Exit) {

visitedSet.Add(m\_middleCodeList[index + 1]);

}

return;

}

else if (middleCode.Operator == MiddleOperator.FunctionEnd) {

Symbol funcSymbol = (Symbol) middleCode[0];

Assert.Error(funcSymbol.Type.ReturnType.IsVoid(),

funcSymbol.Name,

Message.Reached\_the\_end\_of\_a\_non\_\_void\_function);

return;

}

}

}

### Remove Push-Pop Chains

Sometimes there may be a push followed by a pop of the same symbol, or no symbol at all, which in meaningless and shall be removed.

|  |  |  |
| --- | --- | --- |
| 1. push x  2. pop x  1. push x  2. pop | 1. *removed*  2. *removed*  1. *removed*  2. *removed* |  |

However, we do not remove code where different symbols are pushed and popped.

|  |  |  |
| --- | --- | --- |
| 1. push x  2. pop y |  |  |

public void RemovePushPop() {

for (int index = 0; index < (m\_middleCodeList.Count - 1); ++index) {

MiddleCode thisCode = m\_middleCodeList[index],

nextCode = m\_middleCodeList[index + 1];

if ((thisCode.Operator == MiddleOperator.PushFloat) &&

(nextCode.Operator == MiddleOperator.PopFloat) &&

((thisCode[0] == nextCode[0]) || (nextCode[0] == null))) {

thisCode.Clear();

nextCode.Clear();

m\_update = true;

}

}

}

### Merge Pop-Push Chains

Sometimes there may be a pop followed by a push of the same symbol, which shall be replaced by a top.

|  |  |  |
| --- | --- | --- |
| 1. pop x  2. push x | 1. top x  2. *removed* |  |

public void MergePopPushToTop() {

for (int index = 0; index < (m\_middleCodeList.Count - 1); ++index) {

MiddleCode thisCode = m\_middleCodeList[index],

nextCode = m\_middleCodeList[index + 1];

if ((thisCode.Operator == MiddleOperator.PopFloat) &&

(nextCode.Operator == MiddleOperator.PushFloat) &&

(thisCode[0] == nextCode[0])) {

thisCode.Operator = MiddleOperator.TopFloat;

nextCode.Clear();

m\_update = true;

}

}

}

### Change Top-Pop to Pop

Sometimes there may be a push followed by a pop of the same symbol, or no symbol at all, which in meaningless and shall be removed.

|  |  |  |
| --- | --- | --- |
| 1. top x  2. pop | 1. pop x  2. *empty* |  |

public void MergeTopPopToPop() {

for (int index = 0; index < (m\_middleCodeList.Count - 1); ++index) {

MiddleCode thisCode = m\_middleCodeList[index],

nextCode = m\_middleCodeList[index + 1];

if ((thisCode.Operator == MiddleOperator.TopFloat) &&

(nextCode.Operator == MiddleOperator.PopFloat) &&

(nextCode[0] == null)) {

thisCode.Operator = MiddleOperator.PopFloat;

nextCode.Clear();

m\_update = true;

}

}

}

### Merge Binary

When a variable I assigned a binary expression, the resulting middle code is two instructions where the first instruction assign a temporary variable to the expression, and then assign the temporary variable to the resulting variable. We restore the expression to its original form, which allows for generation of more effective assembly code in Chapter 12.

|  |  |  |
| --- | --- | --- |
| a = b + c;  (a) C code | 1. t0 = b + c  2. a = t0  (b) Before optimization | 1. a = b + c  2. *removed*  (b) After optimization |

private void MergeBinary() {

for (int index = 0; index < (m\_middleCodeList.Count - 1); ++index) {

MiddleCode thisCode = m\_middleCodeList[index],

nextCode = m\_middleCodeList[index + 1];

if ((/\*thisCode.IsUnary() ||\*/ thisCode.IsBinary()) &&

(nextCode.Operator == MiddleOperator.Assign) &&

((Symbol) thisCode[0]).IsTemporary() &&

(thisCode[0] == nextCode[1])) {

thisCode[0] = nextCode[0];

nextCode.Clear();

m\_update = true;

}

}

}

### Semantic Optimization

The semantic optimization simplifies expression such add addition of zero or multiplication of one, it handles a number of situations:

|  |  |  |
| --- | --- | --- |
| **Category** | **Example** | **Result** |
| Constant Expression | 2 \* 3 | 6 |
| Binary Addtition | 0 + i  i + 0 | i  i |
|  |  |  |
| Binary Subtraction | 0 - i  i - 0 | -i  i |
| Binary Multiplication | 0 \* i  1 \* i  i \* 0  i \* 1 | 0  i  0  i |
| Binary Division and Modulo | i / 1  i % 1 | i  i |

private void SematicOptimization() {

for (int index = 0; index < m\_middleCodeList.Count; ++index) {

MiddleCode thisCode = m\_middleCodeList[index];

if (thisCode.IsBinary()) {

Symbol resultSymbol = (Symbol) thisCode[0],

leftSymbol = (Symbol) thisCode[1],

rightSymbol = (Symbol) thisCode[2],

newSymbol = null;

Even though the ConstantExpression class evaluates all constant expressions caused by the programmer, the pointer arithmetic during the parsing may have cause more constant expressions, expressions where both operands are constant values. In that case we call ArithmeticIntegral in ConstantExpression to evaluate the value of the expression.

if ((leftSymbol.Value is BigInteger) && // t0 = 2 \* 3

(rightSymbol.Value is BigInteger)) {

newSymbol =

ConstantExpression.ArithmeticIntegral(thisCode.Operator,

leftSymbol, rightSymbol);

}

In case of additions where the left operand is zero, 0 + i, we keep the right operand.

// t0 = 0 + i

else if ((thisCode.Operator == MiddleOperator.BinaryAdd) &&

(leftSymbol.Value is BigInteger) &&

(leftSymbol.Value.Equals(BigInteger.Zero))) {

newSymbol = rightSymbol;

}

// t0 = 0 - i; t0 = -i

else if ((thisCode.Operator == MiddleOperator.BinarySubtract) &&

(leftSymbol.Value is BigInteger) &&

(leftSymbol.Value.Equals(BigInteger.Zero))) {

thisCode.Operator = MiddleOperator.UnarySubtract;

thisCode[0] = thisCode[1];

thisCode[1] = null;

}

In case of additions or subtractions where the right operand is zero, i + 0 or i - 0, we keep the left operand.

// t0 = i + 0

// t0 = i - 0

else if (((thisCode.Operator == MiddleOperator.BinaryAdd) ||

(thisCode.Operator == MiddleOperator.BinarySubtract)) &&

(rightSymbol.Value is BigInteger) &&

(rightSymbol.Value.Equals(BigInteger.Zero))) {

newSymbol = leftSymbol;

}

In case of multiplications where the left operand is zero, the result is zero.

// t0 = 0 \* i

else if (((thisCode.Operator == MiddleOperator.SignedMultiply) ||

(thisCode.Operator == MiddleOperator.UnsignedMultiply)) &&

(leftSymbol.Value is BigInteger) &&

(leftSymbol.Value.Equals(BigInteger.Zero))) {

newSymbol = new Symbol(resultSymbol.Type, BigInteger.Zero);

}

In case of multiplications where the left operand is one, we keep the right operand.

// t0 = 1 \* i

else if (((thisCode.Operator == MiddleOperator.SignedMultiply) ||

(thisCode.Operator == MiddleOperator.UnsignedMultiply)) &&

(leftSymbol.Value is BigInteger) &&

(leftSymbol.Value.Equals(BigInteger.One))) {

newSymbol = rightSymbol;

}

In case of multiplications where the right operand is zero, the result is zero.

// t0 = i \* 0

else if (((thisCode.Operator == MiddleOperator.SignedMultiply) ||

(thisCode.Operator == MiddleOperator.UnsignedMultiply)) &&

(rightSymbol.Value is BigInteger) &&

(rightSymbol.Value.Equals(BigInteger.Zero))) {

newSymbol = new Symbol(resultSymbol.Type, BigInteger.Zero);

}

In case of multiplications or division where the right operand is one, we keep the left operand.

// t0 = i \* 1

// t0 = i / 1

else if (((thisCode.Operator == MiddleOperator.SignedMultiply) ||

(thisCode.Operator == MiddleOperator.UnsignedMultiply) ||

(thisCode.Operator == MiddleOperator.SignedDivide) ||

(thisCode.Operator == MiddleOperator.UnsignedDivide) ||

(thisCode.Operator == MiddleOperator.SignedModulo) ||

(thisCode.Operator == MiddleOperator.UnsignedModulo)) &&

(rightSymbol.Value is BigInteger) &&

(rightSymbol.Value.Equals(BigInteger.One))) {

newSymbol = leftSymbol;

}

If the new symbol is not null, we replace the expression with the new symbol. If the result symbol is a temporary symbol. We iterate until we find the place where the result symbol is accessed and replace the result symbol with the new symbol at that place.

|  |  |
| --- | --- |
| t1 = i + 0  j = t1  k = t1  (a) Before optimization | j = i  k = i  (b) After optimization |

if (newSymbol != null) {

if (resultSymbol.IsTemporary()) {

thisCode.Operator = MiddleOperator.Empty;

if (newSymbol.IsTemporary()) {

int index2;

for (index2 = (index - 1); index2 >= 0; --index2) {

MiddleCode previousCode = m\_middleCodeList[index2];

if (previousCode[0] == resultSymbol) {

previousCode[0] = newSymbol;

break;

}

}

}

{ int index2;

for (index2 = index + 1; index2 < m\_middleCodeList.Count;

++index2) {

MiddleCode nextCode = m\_middleCodeList[index2];

if (nextCode[0] == resultSymbol) {

nextCode[0] = newSymbol;

}

if (nextCode[1] == resultSymbol) {

nextCode[1] = newSymbol;

}

if (nextCode[2] == resultSymbol) {

nextCode[2] = newSymbol;

}

}

}

}

If the result symbol is not temporary symbol, we change the instruction into an assignment, where the result symbol is assigned to the new symbol. For instance, we change i = 0 + j to i = j.

else {

thisCode.Operator = MiddleOperator.Assign; // i = 0 + j;

thisCode[1] = newSymbol; // i = j;

thisCode[2] = null;

}

m\_update = true;

}

}

}

}

### Optimize Relation Expression

In order to optimize the final assembly code generation, we shall swap the operator in expression where the left operand is a value. This operation does not decrease the number of middle code instructions. Instead, it will decrease the number of assembly code instructions, since an integer value cannot be the left-most operands.

|  |  |
| --- | --- |
| 1. if 10 < i goto 20  (a) Before optimization | 1. if i > 10 goto 20  (b) After optimization |

When generating the final assembly code, we cannot have an integer value as the left expression in a relational expression. Therefore, we swap the operands if the left operand holds an integer value. The expression cannot hold two integer values, in that case the ConstantExpression class of Chapter 8 would have reduced the expression to its resulting value.

Moreover, if the left expression is an array, function, or string, we want to use its address rather than its value. Similar to the integer value case, we cannot use the address directly in the assembly code. Therefore, if the left expression is an array, function or string, and the right expression is not an array, function, or string, or an integer value, we also swap the expressions.

We use the m\_swapMap map to swap the operator. Note that that map is not the same map as the m\_inverseMap we used in the Next-Jump-Double case above. When inversing an expression, we change its meaning, when we swap the expression it still has the same meaning, we just arrange the operands so that we can generate more efficient assembly code.

public static IDictionary<MiddleOperator, MiddleOperator> m\_swapMap =

new Dictionary<MiddleOperator, MiddleOperator>() {

{MiddleOperator.Equal, MiddleOperator.Equal},

{MiddleOperator.NotEqual, MiddleOperator.NotEqual},

{MiddleOperator.SignedLessThan, MiddleOperator.SignedGreaterThan},

{MiddleOperator.SignedGreaterThan, MiddleOperator.SignedLessThan},

{MiddleOperator.SignedLessThanEqual,

MiddleOperator.SignedGreaterThanEqual},

{MiddleOperator.SignedGreaterThanEqual,

MiddleOperator.SignedLessThanEqual},

{MiddleOperator.UnsignedLessThan, MiddleOperator.UnsignedGreaterThan},

{MiddleOperator.UnsignedGreaterThan, MiddleOperator.UnsignedLessThan},

{MiddleOperator.UnsignedLessThanEqual,

MiddleOperator.UnsignedGreaterThanEqual},

{MiddleOperator.UnsignedGreaterThanEqual,

MiddleOperator.UnsignedLessThanEqual}

};

private void SwapRelation() {

foreach (MiddleCode middleCode in m\_middleCodeList) {

if (middleCode.IsRelation()) {

Symbol leftSymbol = (Symbol) middleCode[1],

rightSymbol = (Symbol) middleCode[2];

if ((leftSymbol.Value is BigInteger) ||

(leftSymbol.Type.IsArrayFunctionOrString()) &&

!rightSymbol.Type.IsArrayFunctionOrString() &&

!(rightSymbol.Value is BigInteger)) {

middleCode.Operator = m\_swapMap[middleCode.Operator];

middleCode[1] = rightSymbol;

middleCode[2] = leftSymbol;

}

}

}

}

### Optimize Communicative Expression

private void OptimizeCommutative() {

foreach (MiddleCode middleCode in m\_middleCodeList) {

if (middleCode.IsCommutative()) {

Symbol leftSymbol = (Symbol) middleCode[1],

rightSymbol = (Symbol) middleCode[2];

// not 1 - i

if (leftSymbol.Type.IsIntegralPointerArrayOrFunction() &&

(leftSymbol.Value is BigInteger)) {

middleCode[1] = rightSymbol;

middleCode[2] = leftSymbol;

}

}

}

}

### Remove Trivial Assignment

Trivial assignment, where a symbol is assigned the value of the same symbol, such as x = x;, shall be removed.

private void RemoveTrivialAssign() {

foreach (MiddleCode middleCode in m\_middleCodeList) {

MiddleOperator middleOperator = middleCode.Operator;

if (middleOperator == MiddleOperator.Assign) {

Symbol resultSymbol = (Symbol) middleCode[0],

assignSymbol = (Symbol) middleCode[1];

if (resultSymbol == assignSymbol) {

middleCode.Operator = MiddleOperator.Empty;

m\_update = true;

}

}

}

}

### Remove Cleared Code

Technically, it is not strictly necessary to remove cleared middle code instructions since they do not generate assembly code. However, it simplifies the optimization methods of this section.

public void RemoveClearedCode() {

We iterate backwards through the middle code list for performance reasons.

for (int index = (m\_middleCodeList.Count - 1); index >= 0;--index){

if (m\_middleCodeList[index].Operator == MiddleOperator.Empty) {

For each empty instruction we must go through all other instruction and decrease all target by one that is greater than the index of the instruction to be removed.

foreach (MiddleCode middleCode in m\_middleCodeList) {

if (middleCode.IsRelationCarryOrGoto()) {

int target = (int) middleCode[0];

if (target > index) {

middleCode[0] = target - 1;

}

}

}

m\_middleCodeList.RemoveAt(index);

}

}

}

# Assembly Code Generation

When we have generated and optimized the middle code, we continue to generate the final assembly code. The first step is to generate the assembly code with tracks. A track is a place holder for a register and follows a (yet unknown) register through the code. The track is then replaced by a proper register by the register allocator. Finally, the assembly code instructions are written in plain text.

## Runtime Management

When it comes to runtime management, we use the classic style where we allocate an activation record for each function call, beginning with the initial call to main. Each activation record holds the data for the functions’ parameters and local variables and constants. We use the regular frame pointer to hold the address of the current function call. The activation record also holds the frame pointer of the previous activation record and the return address; that is, the address to jump to when the execution of the current function returns to the calling function. More specifically: the address of the instruction following the call in the calling function.

What makes it a little bit more complicated is that a function may be variadic; that is, it has a variable number of parameters, like printf and scanf. We introduce a second frame pointer: the variadic frame pointer for variadic functions. It holds the address of the activation record plus the size of the extra parameters in the call to a variadic function.

To sum it up: the activation record holds the return address, regular frame pointer, and variadic frame pointer of the calling function as well as the parameters, potential extra parameter in case of a variadic function, and auto and register variables and constants of the current function. The variadic frame pointer is undefined in case of a regular calling function. However, we must always have the variadic frame pointer in the activation record since the function can be called by both variadic and regular functions.

When a function calls another function (or itself recursively) the activation record of the called function is located at the address above the activation record of the calling function. The return address, regular and variadic frame pointers of the calling function are stored at the beginning of the activation record of the called function. When a function returns to the calling function, the regular and variadic frame pointers are restored to the values of the calling function, and a jump back to the return address occurs.

The return value is not part of the activation record. It is stored in a register for integral and pointer types and at the floating-point stack for floating types. In case of a struct or union, its address is returned in a register.

The return address of the activation record for the initial main call is initialized to zero. When main return and the return address is zero, an exit of the execution occurs instead of a return. But only for the first call, subsequent recursive calls to main have their own regular return addresses since it is quite possible to make recursive calls to main. For instance, the following program writes the number one to ten.

void main() {

static count = 1;

if (count <= 10) {

printf("%d ", count++);

main();

}

}

Let us look at the following example.

void main() {

int a = 1, b = 2;

f(11, 12);

}

void f(int i, int j) {

int c = 3, d = 4;

g(13, 14);

}

void g(int k, int l) {

int e = 5, f = 6;

// point

}

The program above has the call stack as follows when the execution of the program reached the point. The regular parameters refers to parameters defined in the parameter list. In the next example, we look into variadic parameter: extra parameters on calls to variadic functions.

Variable f: 6

Variable e: 5

Regular Parameter l: 14

Regular Parameter k: 13

Variadic Frame Pointer

Regular Frame Pointer

Return Address: point 2

g

Variable d: 4

Variable c: 3

Regular Parameter j: 12

Regular Parameter i: 11

Variadic Frame Pointer

Regular Frame Pointer

Return Address: point 1

f

Variable b: 2

Variable a:1

Variadic Frame Pointer

Regular Frame Pointer

Return Address: 0

main

In the example above, the variadic frame pointer is ignored. In the following example, however, we let f and g be variadic functions. The variadic status is marked by three dots (‘…’).

void main() {

int a = 1, b = 2;

f(11, 12, 21);

}

void f(int i, int j, ...) {

int c = 3, d = 4;

g(13, 14, 22, 23, 24);

}

void g(int k, int l, ...) {

int e = 5, f = 6;

// point

}

Variable f: 6

Variable e: 5

Variadic parameter: 24

Variadic parameter: 23

Variadic parameter: 22

Regular Parameter l: 14

Regular Parameter k: 13

Variadic Frame Pointer

Regular Frame Pointer

Return Address: point 2

g

Variable d: 4

Variable c: 3

Variadic parameter: 21

Regular Parameter j: 12

Regular Parameter i: 11

Variadic Frame Pointer

Regular Frame Pointer

Return Address: point 1

f

Variable b: 2

Variable a: 1

Variadic Frame Pointer

Regular Frame Pointer

Return Address: 0

main

The variadic parameters are stored after the regular parameters in the activation record. Similar to the previous example, the regular frame pointers point at the beginning of the activation record of the calling function. However, the variadic frame pointers points at the beginning of the activation record of the calling function, plus the size of the variadic parameters. For the sake of argument, let us assume that an integer value allocates four bytes. Then the variadic frame pointer of f is the regular frame pointer plus four bytes (one integer value). The variadic frame pointer of g the regular frame pointer plus twelve bytes (three integer values). When refereeing to parameters in an ecliptic function, the regular frame pointer is used, and when refereeing to variables, the variadic frame pointer is used. The variadic parameters are accessible by using the address of the last regular parameter (it has to be at least one) and reading the values by increase the address, which is the task of the macros va\_list, va\_start, and va\_arg in the stdarg standard library.

## Assembly Operator

Like the middle code operators, we also have the assembly code operators. Several operators come in several varieties. For instance, the add instruction comes in the base form, which is used when a register is involved, like add ax, 123, sub ax, [bp + 2] or xor [count], bx. The other varieties are used when there is no register involved, and we therefore have to specify the size of the value to be assigned, like add word [bp + 2], 123.

AssemblyOperator.cs

namespace CCompiler {

public enum AssemblyOperator {

add, add\_byte, add\_dword, add\_qword, add\_word, return\_address,

and, and\_byte, and\_dword, and\_qword, and\_word, call,

cmp, cmp\_byte, cmp\_dword, cmp\_qword, cmp\_word, comment,

dec, dec\_byte, dec\_dword, dec\_qword, dec\_word,

define\_address, define\_value, define\_zero\_sequence,

div, div\_byte, div\_dword, div\_qword, div\_word, empty,

fabs, fadd, faddp, fchs, fcompp, fdiv, fdivp, fdivr,

fdivrp, fild\_dword, fild\_qword, fild\_word,

fist\_dword, fist\_qword, fist\_word,

fistp\_dword, fistp\_qword, fistp\_word,

fld1, fld\_dword, fld\_qword, fldcw, fldz,

fmul, fmulp, fst\_dword, fst\_qword, fstcw,

fstp\_dword, fstp\_qword, fstsw,

fsub, fsubp, fsubr, fsubrp, ftst,

idiv, idiv\_byte, idiv\_dword, idiv\_qword, idiv\_word,

imul, imul\_byte, imul\_dword, imul\_qword, imul\_word,

inc, inc\_byte, inc\_dword, inc\_qword, inc\_word, interrupt,

ja, jae, jb, jbe, jc, je, jg, jge, jl, jle, jmp, jnc, jne, jnz, jz,

label, lahf, mov, mov\_byte, mov\_dword, mov\_qword, mov\_word,

mul, mul\_byte, mul\_dword, mul\_qword, mul\_word,

neg, neg\_byte, neg\_dword, neg\_qword, neg\_word, new\_middle\_code,

nop, not, not\_byte, not\_dword, not\_qword, not\_word,

or, or\_byte, or\_dword, or\_qword, or\_word, pop, ret, sahf, set\_track\_size,

shl, shl\_byte, shl\_dword, shl\_qword, shl\_word,

shr, shr\_byte, shr\_dword, shr\_qword, shr\_word,

sub, sub\_byte, sub\_dword, sub\_qword, sub\_word, syscall,

xor, xor\_byte, xor\_dword, xor\_qword, xor\_word

};

};

## Assembly Code

The AssemblyCode class handles one assembly code instruction. It holds methods for initialization and optimization of a single instruction, a set of methods for testing and conversion of values and register, and the ToString method that writes the instruction in plain text.

AssemblyCode.cs

using System;

using System.Text;

using System.Numerics;

using System.Collections.Generic;

First, there is a set of special registers:

* RegularFrameRegister. The regular frame pointer: the address of the beginning of the current activation record.
* VariadicFrameRegister. The variadic frame pointer: the address of the beginning of the current activation record plus the size (in bytes) of the variadic parameters, in order to correctly access the local variables of the function, which are located after the variadic parameters. Ignored in regular functions.
* ReturnValueRegister. Holds the function return value in case of integral or pointer type, or its address in case of struct or union type. Return values of floating types are not stored in a register or on the activation record, but rather at the floating-point stack. In C, it is not allowed to return values or array of function types.
* ReturnPointerRegister. Holds the address of the function return value in case of struct or union type.
* ShiftRegister. Holds the value of the right operand in a shift operation. Due to conditions of the architecture, it must always be cl.

namespace CCompiler {

public class AssemblyCode {

public static Register RegularFrameRegister;

public static Register VariadicFrameRegister;

public static Register ReturnValueRegister;

public static Register ReturnPointerRegister;

public const Register ShiftRegister = Register.cl;

We chose bp, di, and bx as the registers for the regular frame pointer, variadic frame pointer, and return value, since they (together with si) can be used as addresses in assembly operations.

static AssemblyCode() {

RegularFrameRegister =

RegisterToSize(Register.bp, TypeSize.PointerSize);

VariadicFrameRegister =

RegisterToSize(Register.di, TypeSize.PointerSize);

ReturnValueRegister = RegisterToSize(Register.bx, TypeSize.PointerSize);

ReturnPointerRegister =RegisterToSize(Register.bx,TypeSize.PointerSize);

}

An assembly instruction is made up by an operator and at most three operands. The operands are defined as objects, and they can be Register, int, string and BigInteger. All offsets are int, and all integer values are BigInteger.

private AssemblyOperator m\_operator;

private object[] m\_operandArray = new object[3];

The constructor initializes the instruction, and calls FromAdditionToIncrement, which changes addition and subtraction with one to increment and decrement, and CheckSize, which changes the operators in accordance with the given size, unless it is zero.

public AssemblyCode(AssemblyOperator objectOp, object operand0,

object operand1, object operand2 = null,

int size = 0) {

m\_operator = objectOp;

m\_operandArray[0] = operand0;

m\_operandArray[1] = operand1;

m\_operandArray[2] = operand2;

FromAdditionToIncrement();

CheckSize(size);

}

public AssemblyOperator Operator {

get { return m\_operator; }

set { m\_operator = value; }

}

The index property handles the operand array.

public object this[int index] {

get { return m\_operandArray[index]; }

set { m\_operandArray[index] = value; }

}

### Assembly Code Optimization

Similar to the middle code optimization in Chapter 11, there is also some optimization of the assembly code. The FromAdditionToIncrement method changes the addition of one or subtraction of minus one to increment, and subtraction of one and addition of minus on to decrement. The first condition is that operator indeed is addition or subtraction and that the first operand is a track, a register, or a string.

private void FromAdditionToIncrement() {

if (((Operator == AssemblyOperator.add) ||

(Operator == AssemblyOperator.sub)) &&

((m\_operandArray[0] is Track) || (m\_operandArray[0] is Register) ||

(m\_operandArray[0] is string))) {

Then we check whether the second operand is a BigInteger and the third operand is null. In that case the value to be inspected has index one and we call CheckIncrement for further processing.

if ((m\_operandArray[1] is BigInteger) && (m\_operandArray[2] == null)){

CheckIncrement(1);

}

If the second operand is an integer and operand is a BigInteger, the value to be inspected has index two.

else if ((m\_operandArray[1] is int) &&

(m\_operandArray[2] is BigInteger)) {

CheckIncrement(2);

}

}

}

The CheckIncrement method inspects the value to be added or subtracted. If the operation is addition of one or subtraction of minus one, we replace the operator with increment.

private void CheckIncrement(int valueIndex) {

int value = (int) ((BigInteger) m\_operandArray[valueIndex]);

if (((Operator == AssemblyOperator.add) && (value == 1)) ||

((Operator == AssemblyOperator.sub) && (value == -1))) {

m\_operator = AssemblyOperator.inc;

m\_operandArray[valueIndex] = null;

}

In the same way, if the operation is addition of minus one subtraction of one, we replace the operator with decrement.

else if (((Operator == AssemblyOperator.add) && (value == -1)) ||

((Operator == AssemblyOperator.sub) && (value == 1))) {

m\_operator = AssemblyOperator.dec;

m\_operandArray[valueIndex] = null;

}

}

The CheckSize method changes the operator to a size operator. For instance, add [bp + 2], 3 shall be changed to add\_word [bp + 2], 3 if the address represents a 2-byte value (size is two). Note that this method does not perform optimization, it just set the correct operator.

private void CheckSize(int size) {

if ((size != 0) && ((m\_operandArray[0] is Register) ||

(m\_operandArray[0] is Track) || (m\_operandArray[0] is String)) &&

(m\_operandArray[1] is int) &&

(IsUnary() || ((m\_operandArray[2] is BigInteger) ||

(m\_operandArray[2] is String)))) {

m\_operator = OperatorToSize(m\_operator, size);

}

}

### Operator Test Methods

There is a set of methods that test whether the operator holds certain properties. The IsUnary method return true is the operator is unary; that is, it takes one operand. Note that we regard multiplication and division as unary operators, even though it takes two operands. That is because the left operand is stored in a register, only the second operand is explicitly stated. We also regard loading and storing of floating-point values as unary operators.

public bool IsUnary() {

string operatorName = Enum.GetName(typeof(AssemblyOperator), Operator);

return operatorName.StartsWith("neg") ||

operatorName.StartsWith("not") ||

operatorName.StartsWith("inc") ||

operatorName.StartsWith("dec") ||

operatorName.StartsWith("mul") ||

operatorName.StartsWith("imul") ||

operatorName.StartsWith("div") ||

operatorName.StartsWith("idiv") ||

operatorName.StartsWith("fst") ||

operatorName.StartsWith("fld") ||

operatorName.StartsWith("fist") ||

operatorName.StartsWith("fild");

}

The jump and call test methods tests whether the first operand is a register. The jump instruction usually jumps to a specifically stated address. However, when returning from a function call, the address is stored in a register.

public bool IsJumpRegister() {

return (Operator == AssemblyOperator.jmp) &&

(m\_operandArray[0] is Register);

}

public bool IsJumpNotRegister() {

return (Operator == AssemblyOperator.jmp) &&

!(m\_operandArray[0] is Register);

}

A regular function call is to a specifically stated address. However, it is also possible to call a function whose address is stored in a function pointer.

public bool IsCallRegister() {

return (Operator == AssemblyOperator.call) &&

(m\_operandArray[0] is Register);

}

public bool IsCallNotRegister() {

return (Operator == AssemblyOperator.call) &&

(m\_operandArray[0] is string);

}

The IsRelationNotRegister method returns true if the operator is a relational jump operator. Note that there is no IsRelationRegister, since there are only jumps to specific addresses, not to addresses stored in registers.

public bool IsRelationNotRegister() {

switch (Operator) {

case AssemblyOperator.je:

case AssemblyOperator.jne:

case AssemblyOperator.jl:

case AssemblyOperator.jle:

case AssemblyOperator.jg:

case AssemblyOperator.jge:

case AssemblyOperator.jb:

case AssemblyOperator.jbe:

case AssemblyOperator.ja:

case AssemblyOperator.jae:

case AssemblyOperator.jc:

case AssemblyOperator.jnc:

return true;

default:

return false;

}

}

### Register Overlapping

When allocating registers, we have to make sure the registers do not overlap. To begin with, the m\_registerOverlapSet set holds sets of all overlapping registers. Note that the al and ah registers are stored in separate sets, since they do not overlap.

private static ISet<ISet<Register>> m\_registerOverlapSet =

new HashSet<ISet<Register>>() {

new HashSet<Register>() {Register.al, Register.ax,

Register.eax, Register.rax},

new HashSet<Register>() {Register.ah, Register.ax,

Register.eax, Register.rax},

new HashSet<Register>() {Register.bl, Register.bx,

Register.ebx, Register.rbx},

new HashSet<Register>() {Register.bh, Register.bx,

Register.ebx, Register.rbx},

new HashSet<Register>() {Register.cl, Register.cx,

Register.ecx, Register.rcx},

new HashSet<Register>() {Register.ch, Register.cx,

Register.ecx, Register.rcx},

new HashSet<Register>() {Register.dl, Register.dx,

Register.edx, Register.rdx},

new HashSet<Register>() {Register.dh, Register.dx,

Register.edx, Register.rdx},

new HashSet<Register>() {Register.si, Register.esi, Register.rsi},

new HashSet<Register>() {Register.di, Register.edi, Register.rdi},

new HashSet<Register>() {Register.bp, Register.ebp, Register.rbp},

new HashSet<Register>() {Register.sp, Register.esp, Register.rsp}

};

The RegisterOverlaps methods returns true if the two registers overlap. If at least one register is null, they do not overlap, and we return false.

public static bool RegisterOverlap(Register? register1,

Register? register2) {

if ((register1 == null) || (register2 == null)) {

return false;

}

Otherwise, we iterate through the m\_registerOverlapSet. If one of the register sets holds both the registers, they overlap, and we return true.

foreach (ISet<Register> registerSet in m\_registerOverlapSet) {

if (registerSet.Contains(register1.Value) &&

registerSet.Contains(register2.Value)) {

return true;

}

}

If we have iterated through the m\_registerOverlapSet without finding a set holding both the registers, the registers do not overlap, and we return false.

return false;

}

### Register Size

The SizeOfRegister method uses the m\_registerSizeMap map to look up and return the size of the register.

private static IDictionary<Register,int> m\_registerSizeMap =

new Dictionary<Register,int>() {

{Register.al, 1}, {Register.bl, 1}, {Register.cl, 1}, {Register.dl, 1},

{Register.ah, 1}, {Register.bh, 1}, {Register.ch, 1}, {Register.dh, 1},

{Register.ax, 2}, {Register.bx, 2}, {Register.cx, 2}, {Register.dx, 2},

{Register.eax,4}, {Register.ebx,4}, {Register.ecx,4}, {Register.edx,4},

{Register.rax,8}, {Register.rbx,8}, {Register.rcx,8}, {Register.rdx,8},

{Register.si, 2}, {Register.di, 2}, {Register.sp, 2}, {Register.bp, 2},

{Register.esi,4}, {Register.edi,4}, {Register.esp,4}, {Register.ebp,4},

{Register.rsi,8}, {Register.rdi,8}, {Register.rsp,8}, {Register.rbp, 8}

};

public static int SizeOfRegister(Register register) {

return m\_registerSizeMap[register];

}

The RegisterToSize method converts a register to the given size. The m\_registerSet set holds lists of registers of increasing sizes. Note that there are lists rather than sets, because the order of the registers is significant.

private static ISet<IList<Register>> m\_registerSet =

new HashSet<IList<Register>>() {

new Register[] {Register.al, Register.ax, Register.eax, Register.rax},

new Register[] {Register.bl, Register.bx, Register.ebx, Register.rbx},

new Register[] {Register.cl, Register.cx, Register.ecx, Register.rcx},

new Register[] {Register.dl, Register.dx, Register.edx, Register.rdx},

new Register[] {default(Register), Register.si,

Register.esi, Register.rsi},

new Register[] {default(Register), Register.di,

Register.edi, Register.rdi},

new Register[] {default(Register), Register.bp,

Register.ebp, Register.rbp},

new Register[] {default(Register), Register.sp,

Register.esp, Register.rsp}

};

The m\_sizeToIndexMap map maps the size of the registers to the index in the register lists of the m\_registerSet set above.

private static IDictionary<int,int> m\_sizeToIndexMap =

new Dictionary<int, int>() {{1, 0}, {2, 1}, {4, 2}, {8, 3}};

The RegisterToSize method returns the register with the given size. If the register already holds the given size, we just return the register.

public static Register RegisterToSize(Register register, int size) {

if (m\_registerSizeMap[register] == size) {

return register;

}

Otherwise, we iterate through the m\_registerSet set, and check if any of the register sets contains the register. If it contains the register, we return the register on the index in the list corresponding to the size of the register.

foreach (IList<Register> registerList in m\_registerSet) {

if (registerList.Contains(register)) {

int index = m\_sizeToIndexMap[size];

Assert.ErrorXXX((index >= 0) && (index < registerList.Count));

return registerList[index];

}

}

Assert.ErrorXXX(false);

return default(Register);

}

The SizeOfOperator method returns the size of the operator. We obtain the name of the operator by calling the GetName method in the Enum class, and test whether is holds the suffix “byte”, “word”, “dword”, or “qword”, in which cases the operator size is one, two, four, or eight bytes.

public static int SizeOfOperator(AssemblyOperator objectOp) {

string name = Enum.GetName(typeof(AssemblyOperator), objectOp);

string suffix = name.Substring(name.IndexOf("\_") + 1);

switch (suffix) {

case "byte":

return 1;

case "word":

return 2;

case "dword":

return 4;

default: // "qword":

return 8;

}

}

The OperatorToSize method changes a plain operator to a size operator. If the operator size is one, two, four, or eight, its name shall be suffixed by “\_byte”, “\_word”, “\_dword”, or “\_qword”.

public static AssemblyOperator OperatorToSize

(AssemblyOperator objectOp, int size) {

string name = Enum.GetName(typeof(AssemblyOperator), objectOp);

Assert.ErrorXXX(objectOp != AssemblyOperator.interrupt);

switch (size) {

case 1:

name = name + "\_byte";

break;

case 2:

name = name + "\_word";

break;

case 4:

name = name + "\_dword";

break;

case 8:

name = name + "\_qword";

break;

}

We use the Parse method of the Enum class to obtain the resulting operator from the name.

Assert.ErrorXXX(name.Contains("\_"));

return ((AssemblyOperator) Enum.Parse(typeof(AssemblyOperator), name));

}

The SizeOfValue method return a size of a value, with two exceptions. In case of a mov operator, the size is given by the operator, and in case of a cmp operator with value zero, the size is one. Otherwise, we just call the second SizeOfValue to get the value size.

public static int SizeOfValue(BigInteger value, AssemblyOperator op) {

string name = Enum.GetName(typeof(AssemblyOperator), op);

if (name.StartsWith("mov")) {

return SizeOfOperator(op);

}

else if (name.StartsWith("cmp") && (value == 0)) {

return 1;

}

else {

return SizeOfValue(value);

}

}

The second version of SizeOfValue return the size of a value. The minimum and maximum value is and , where is the number of bits of the type.

public static int SizeOfValue(BigInteger value) {

if (value == 0) {

return 0;

}

else if ((-128 <= value) && (value <= 127)) {

return 1;

}

else if ((-32768 <= value) && (value <= 32767)) {

return 2;

}

else if ((-2147483648 <= value) && (value <= 2147483647)) {

return 4;

}

else {

return 8;

}

}

### ToString

The ToString method returns the instruction in plain text. We have six categories of operations:

public override string ToString() {

object operand0 = m\_operandArray[0],

operand1 = m\_operandArray[1],

operand2 = m\_operandArray[2];

string operatorName = Enum.GetName(typeof(AssemblyOperator),

Operator).Replace("\_", " ");

Instructions that perform computations are nullary (without operands), unary (with one operand), or binary (with two operands). In binary operations, the left operand must be a register or an address; a static name or an integer value is not allowed. The operations can be further divided into six categories:

|  |  |  |  |
| --- | --- | --- | --- |
| **Type** | **Category** | **Description** | **Examples** |
| Nullary | 1 | operator | fadd |
| Unary | 2 | operator register | neg ax |
| 3 | operator [register + offset]  operator [static name + offset] | inc [bp + 2]  dec [stdin + 4] |
| Binary | 5 | operator register, register  operator register, static name  operator register, integer value | mov ax, bx  add ax, stdin  sub ax, 123 |
| 6 | operator [register + offset], register  operator [register + offset], static name  operator [register + offset], integer value  operator [static name + offset], register  operator [static name + offset], static name  operator [static name + offset], integer value | mov [bp + 2], bx  add [bp + 2], stdin  sub [bp + 2], 123  mov [stdin + 4], bx  add [stdin + 4], stdin  sub [stdin + 4], 123 |
| 6 | operator register, [register + offset]  operator register, [static name + offset] | mov ax, [bp + 2]  add ax, [stdin +4] |

In case of a nullary operator, we simply return its name.

if (IsNullary()) {

return "\t" + operatorName;

}

Unary operations of category two and three, where the operand is an address or a register. In case of an address, the first operand is a register or a string (static name), the second operand is an integer (offset), and the third operand is null. In case of a register, the second and third operands are null.

else if (IsUnary()) {

if (((operand0 is Register) || (operand0 is string)) &&

(operand1 is int) && (operand2 == null)) {

return "\t" + operatorName + " [" + operand0 +

WithSign(operand1) + "]";

}

else if ((operand0 is Register) && (operand1 == null) &&

(operand2 == null)) {

return "\t" + operatorName + " " + operand0;

}

Next, we handle binary operations. In category four, the first operand is a register, and the second operand is a register, a string (static name), or a BigInteger (integer value).

else if (IsBinary()) {

if ((operand0 is Register) && ((operand1 is Register) ||

(operand1 is string) || (operand1 is BigInteger)) &&

(operand2 == null)) {

Assert.ErrorXXX(!(operand0 is string));

return "\t" + operatorName + " " + operand0 + ", " + operand1;

}

In category five, the first operand is a register or a string (static name), the second operand is an integer (offset), and the third operand is register, a string (static name), or a BigInteger (integer value).

else if (((operand0 is Register) || (operand0 is string)) &&

(operand1 is int) && ((operand2 is Register) ||

(operand2 is string) || (operand2 is BigInteger))) {

return "\t" + operatorName +

" [" + operand0 + WithSign(operand1) + "], " + operand2;

}

In category six, the first operand is a register, the second operand is register or string (static name), and the third operand is an integer (offset).

else if ((operand0 is Register) && ((operand1 is Register) ||

(operand1 is string)) && (operand2 is int)) {

return "\t" + operatorName + " " + operand0 +

", [" + operand1 + WithSign(operand2) + "]";

}

}

In case of a label or comment, we simple return them. Note that we add a colon (‘:’) to the label and insert a semicolon (‘;’).

else if (Operator == AssemblyOperator.label) {

return "\n " + operand0 + ":";

}

else if (Operator == AssemblyOperator.comment) {

return "\t; " + operand0;

}

When defining a value, we call the ToVisibleString method in case of a string.

else if (Operator == AssemblyOperator.define\_value) {

Sort sort = (Sort) operand0;

object value = operand1;

if (sort == Sort.String) {

return "\tdb " + ToVisibleString((string) operand1);

}

If the value is not a string, it is an integral or floating value. In case of floating value, we add a dot (‘.’), unless it already has a dot.

else {

string text = operand1.ToString();

if (((sort == Sort.Float) || (sort == Sort.Double) ||

(sort == Sort.LongDouble)) && !text.Contains(".")) {

text += ".0";

}

When we define the value, we use different directives depending on the size of the value: db (define byte) for one byte, dw (define word) for two bytes, dd (define double-word) for four bytes, and dq (define quarto) for eight bytes.

switch (TypeSize.Size(sort)) {

case 1:

return "\tdb " + text;

case 2:

return "\tdw " + text;

case 4:

return "\tdd " + text;

case 8:

return "\tdq " + text;

}

}

}

When defining an address, we use the dq directive, since the size of addresses is eight bytes. The WithSign method write a positive or negative offset, unless the offset is zero.

else if (Operator == AssemblyOperator.define\_address) {

string name = (string) operand0;

int offset = (int) operand1;

return "\tdq " + name + WithSign(offset);

}

When defining a sequence of zeros, we use the times and db directive, which repeats the following definition, which is the zero value of one-byte size.

else if (Operator == AssemblyOperator.define\_zero\_sequence) {

int size = (int) operand0;

return "\ttimes " + size + " db 0";

}

For a jump or function call, we use the jmp instruction.

else if (IsJumpRegister() || IsCallRegister() ||

IsCallNotRegister()) {

return "\tjmp " + operand0;

}

For address return, we load the target into the address.

else if (Operator == AssemblyOperator.return\_address) {

string target = SymbolTable.CurrentFunction.UniqueName +

Symbol.SeparatorId + operand2;

return "\tmov qword [" + operand0 + WithSign(operand1) + "], " +

target;

}

In case of conditional and unconditional jump instructions, we have three cases. If the third operand (operand2) is an integer, we jump to the address of a middle code instruction. More specific, we jump to a label, which is made up by the name of the current function, and the middle code index given by the third operand.

else if (IsRelationNotRegister() || IsJumpNotRegister()) {

If the third operand is string, we have a special case: the jump instruction is part of the handling of the command line arguments. We simply jump to the address given by the string.

Otherwise, we have jump in a memory copy of a struct or union.

Assert.ErrorXXX(operand2 is int);

string label = SymbolTable.CurrentFunction.UniqueName +

Symbol.SeparatorId + operand2;

return "\t" + operatorName + " " + label;

}

}

The ToVisibleString method returns the text with non-graphical character written with their ascii numbers. For instance, the text “Hello\nWorld\r”, will be translated to “Hello”, 10, “World”, 13.

private static string ToVisibleString(string text) {

StringBuilder buffer = new StringBuilder();

The insideString variable is true as long as the current character is located inside a string.

bool insideString = false;

We iterate through the text and, for each character, check if it is not a graphical character. In that case we end the string, if necessary, and add the ascii value of the character.

foreach (char c in text) {

if (Char.IsControl(c) || (c == '\"') || (c == '\'')) {

if (insideString) {

buffer.Append("\", " + ((int) c).ToString() + ", ");

insideString = false;

}

else {

buffer.Append(((int) c).ToString() + ", ");

}

}

If the character is a graphical character, we begin the string, if necessary, and add the character to the string.

else {

if (insideString) {

buffer.Append(c);

}

else {

buffer.Append("\"" + c.ToString());

insideString = true;

}

}

}

We end the string by adding the terminating zero character.

if (insideString) {

buffer.Append("\", 0");

}

else {

buffer.Append("0");

}

return buffer.ToString();

}

The MakeMemory method return a label for the memory copy method.

public static string MakeLabel(int labelIndex) {

return "label" + Symbol.SeparatorId + labelIndex;

}

The WithSign method returns the offset of an address preceded by a plus or minus sign. In case of a zero value, an empty string is returned.

private string WithSign(object value) {

int offset = (int) value;

if (offset > 0) {

return " + " + offset;

}

else if (offset < 0) {

return " - " + (-offset);

}

else {

return "";

}

}

}

}

## Tracks

A track is a place holder for a register through the assembly code. Let us look at the following example.

|  |  |
| --- | --- |
| return a + b & c - d;  (a) C code | 1. temp0 = a + b  2. temp1 = c - d  3. temp2 = temp0 & temp1  4. return temp2  (b) Middle code |

The following middle code is generated.

Let us assume that a, b, c, d, and e are placed on the current activation record of a regular function with offsets 6, 8, 10, and 12, and the bp and di registers are the regular frame pointer and the variadic frame pointer.

The following assembly code is generated, with the tracks track0, track1, and track2 as place holders for the registers. Each track represents a register, but we do not yet know which one.

mov track0,[bp + 6]

add track0,[bp + 8]

mov track1,[bp + 10]

sub track1,[bp + 12]

and track0, track1

In the return sequence, we reset the variadic and regular frame pointers to the values of the calling function, which is placed at offset 4 and 2 on the activation record, and then jump to the return address, which is placed at the beginning of the activation record. Note the order between the two middle lines. We cannot swap them, since bp is set on the third line.

mov track2, [bp]

mov di, [bp + 4]

mov bp, [bp + 2]

jmp track2

The return value is stored in track0, and shall thereby be assigned the return value register. I have chosen bx for the return value register, since it can be used both as an operand in arithmetic operations and as an address register. So track0 is assigned bx when the return instruction is added to the code.

The register allocator assigns the remaining tracks track1 and track2 to suitable registers. Since track0 and track2 do not overlap, it is quite possible to assign them the same register, such as ax.

mov bx,[bp + 6]

add bx,[bp + 8]

mov ax,[bp + 10]

sub ax,[bp + 12]

and bx, ax

mov ax, [bp]

mov di, [bp + 4]

mov bp, [bp + 2]

jmp ax

The Track class holds a track representing a known or unknown register.

Track.cs

using System;

using System.Collections.Generic;

namespace CCompiler {

public class Track {

The first constructor takes a symbol and a potential register as parameters, and initializes the current and maximum sizes.

public Track(Symbol symbol, Register? register = null) {

Register = register;

Assert.ErrorXXX(symbol != null);

Assert.ErrorXXX(!symbol.Type.IsStructOrUnion());

CurrentSize = m\_maxSize = symbol.Type.ReturnSize();

}

The second constructor takes a type and initializes the current and maximum sizes.

public Track(Type type) {

Assert.ErrorXXX(type != null);

Assert.ErrorXXX(!type.IsStructOrUnion());

Assert.ErrorXXX(!type.IsArrayFunctionOrString());

CurrentSize = m\_maxSize = type.Size();

}

A track has a current size and a maximal size. The current size may vary between assembly code instructions. We need the maximal size for the register allocator. It the size more than one byte, a reduced set of register is available.

public int CurrentSize {get;set;}

private int m\_maxSize;

public int MaxSize {

get { return m\_maxSize; }

set { m\_maxSize = Math.Max(m\_maxSize, value); }

}

The min and max index hold he indexes of the first and last assembly code instruction where the track occurs. We need them to decide whether tracks overlap.

private int m\_minIndex = -1, m\_maxIndex = -1;

public int Index {

set {

m\_minIndex = (m\_minIndex != -1) ? Math.Min(m\_minIndex, value) : value;

m\_maxIndex = Math.Max(m\_maxIndex, value);

}

}

Although the idea is that a track does not hold a register but is rather assigned a register by the register allocator, there are some cases where the track is assigned a register from the beginning. A track can hold a specific register. In multiplication and division, the left operands must be stored in a specific register. In left or right shift, the right operands must be stored in the cl register. When returning a value, it must be stored in a specific register.

public Register? Register {get; set;}

If the track is used as a pointer, a reduced set of registers are available.

public bool Pointer {get; set;}

The Overlaps method returns true if the tracks overlaps. Two tracks overlap if the min and max line numbers overlap.

public static bool Overlaps(Track track1, Track track2) {

Assert.ErrorXXX((track1.m\_minIndex != -1) && (track1.m\_maxIndex != -1));

Assert.ErrorXXX((track2.m\_minIndex != -1) && (track2.m\_maxIndex != -1));

return !(((track1.m\_maxIndex < track2.m\_minIndex) ||

(track2.m\_maxIndex < track1.m\_minIndex)));

}

}

}

## Register Allocation

The architecture holds a set of registers. On particular feature of the registers is that they to some extent overlaps. See Appendix C for a more detailed description.

Register.cs

namespace CCompiler {

public enum Register {al, ah, ax, eax, rax, bl, bh, bx, ebx, rbx,

cl, ch, cx, ecx, rcx, dl, dh, dx, edx, rdx,

si, esi, rsi, di, edi, rdi,

sp, esp, rsp, bp, ebp, rbp};

}

The RegisterAllocator method take the total track set and total assembly list; that is, the track set for the whole function, and the assembly list for the whole function. The first task is the find out which tracks overlaps each other. We construct the graph in the following way: the tracks are the vertices of the graph and two vertices have an edge if their tracks overlaps. The register allocation is then performed as a graph coloring, two vertices connected by an edge cannot be given the same color. In our case: two tracks that overlaps cannot be assigned the same register.

Graph coloring is a NP-complete problem, which means that there is no known algorithm that works on polynomial time. To be sure that we have found the best solution, we have to exam every possible solution, which would require an unrealistic amount of time. Instead, in this book we perform a deep search, where we sort the vertices into a list and for each vertex try to find a register not already allocated by any of its neighbors. If we cannot find such register, we backtrack and try another combination. If we finally find a solution where each vertex is mapped to a register not mapped by any of its neighbors, we have found an optimal solution. If we do not find a solution, we report an error: shortage of registers.

RegisterAllocator.cs

using System.Collections.Generic;

namespace CCompiler {

public class RegisterAllocator {

public RegisterAllocator(ISet<Track> totalTrackSet,

List<AssemblyCode> assemblyCodeList) {

Graph<Track> totalTrackGraph = new Graph<Track>(totalTrackSet);

First, we build the total track graph, with all track as vertices and an edge between two vertices if their tracks overlaps. Two tracks that overlap shall not be assigned the same register, or two overlapping registers.

foreach (Track track1 in totalTrackSet) {

foreach (Track track2 in totalTrackSet) {

if (!track1.Equals(track2) && Track.Overlaps(track1, track2)) {

totalTrackGraph.AddEdge(track1, track2);

}

}

}

Then we split the total graph into independent subgraphs, of performance reasons. It takes less time to evaluate a set of smaller graphs then one large graph.

ISet<Graph<Track>> split = totalTrackGraph.Split();

When the graph has been split, we iterate through the subgraphs and perform a deep-first search on each of them.

foreach (Graph<Track> trackGraph in split) {

List<Track> trackList = new List<Track>(trackGraph.VertexSet);

Assert.Error(DeepFirstSearch(trackList, 0, trackGraph),

Message.Out\_of\_registers);

If the DeepFirstSearch method returns true, we have a graph coloring; that is, the tracks have been assigned registers that do not overlap. If DeepFirstSearch return false, we have not found a graph coloring, and we report an error, that we are out of registers.

ISet<Graph<Track>> split = totalTrackGraph.Split();

foreach (Graph<Track> trackGraph in split) {

List<Track> trackList = new List<Track>(trackGraph.VertexSet);

Assert.Error(DeepFirstSearch(trackList, 0, trackGraph),

Message.Out\_of\_registers);

}

If we have iterated through the track graphs without encountering shortage of register, each track has been assigned a register. We call SetRegistersInCodeList that iterates through the code and replace each track with its assigned register.

SetRegistersInCodeList(assemblyCodeList);

}

private static void SetRegistersInCodeList(List<AssemblyCode>

assemblyCodeList) {

foreach (AssemblyCode assemblyCode in assemblyCodeList) {

if (assemblyCode.Operator == AssemblyOperator.set\_track\_size) {

Track track = (Track) assemblyCode[0];

object operand1 = assemblyCode[1];

When iterating through the code list, we also need to keep track of the current size of the tracks, in order to replace the track with a register of correct size.

if (operand1 is int) {

track.CurrentSize = (int) operand1;

}

else {

track.CurrentSize = ((Track) operand1).CurrentSize;

}

assemblyCode.Operator = AssemblyOperator.empty;

}

else {

Check(assemblyCode, 0);

Check(assemblyCode, 1);

Check(assemblyCode, 2);

}

}

}

The Check method checks for each position in the assembly code is a reference a track is replaced by a register of the correct size.

private static void Check(AssemblyCode assemblyCode, int position) {

if (assemblyCode[position] is Track) {

Track track = (Track) assemblyCode[position];

Assert.ErrorXXX(track.Register != null);

assemblyCode[position] =

AssemblyCode.RegisterToSize(track.Register.Value, track.CurrentSize);

}

}

The DeepFirstSearch method searches the graph in a deep-first manner. It takes the track list and the current index in that list as well as the track graph.

private bool DeepFirstSearch(List<Track> trackList, int listIndex,

Graph<Track> trackGraph) {

If the index equals the size of the track list, we return true because we have iterated through the list and found a match; that is, each track has been assigned a register and no overlapping tracks have the same register.

if (listIndex == trackList.Count) {

return true;

}

If the current track has already been assigned a register, we just call DeepFirstSearch with the next index.

Track track = trackList[listIndex];

if (track.Register != null) {

return DeepFirstSearch(trackList, listIndex + 1, trackGraph);

}

If the current track has not been assigned a register, we look up the set of possible register and the set of neighbor vertices; that is, the set of overlapping tracks.

ISet<Register> possibleSet = GetPossibleSet(track);

ISet<Track> neighbourSet = trackGraph.GetNeighbourSet(track);

We iterate through the set of possible register and, for each register that does not cause an overlapping, we assign the track the register and call DeepFirstSearch recursively with the next index. If the call returns true, we have found a total mapping of registers to the track, and we just return true. However, if the call does return false, we just try with another of the possible registers. If none of the registers causes a match, we clear the register of the track and return false.

foreach (Register possibleRegister in possibleSet) {

if (!OverlapNeighbourSet(possibleRegister, neighbourSet)) {

track.Register = possibleRegister;

if (DeepFirstSearch(trackList, listIndex + 1, trackGraph)) {

return true;

}

track.Register = null;

}

}

track.Register = null;

return false;

}

The OverlapNeighbourSet method return true if the register overlaps any of its neighbors. The RegisterOverlap method in the AssemblyCode class test whether two register overlaps.

private bool OverlapNeighbourSet(Register register,

ISet<Track> neighbourSet) {

foreach (Track neighbourTrack in neighbourSet) {

if (AssemblyCode.RegisterOverlap(register, neighbourTrack.Register)) {

return true;

}

}

return false;

}

The VariadicFunctionPointerRegisterSet set holds the possible pointer registers of a variadic function while RegularFunctionPointerRegisterSet holds the possible pointer registers of a regular function. The Byte1RegisterSet set holds all registers of one byte while Byte2RegisterSet holds all registers of two bytes.

public static ISet<Register>

VariadicFunctionPointerRegisterSet = new HashSet<Register>() {

AssemblyCode.RegisterToSize(Register.bp, TypeSize.PointerSize),

AssemblyCode.RegisterToSize(Register.si, TypeSize.PointerSize),

AssemblyCode.RegisterToSize(Register.di, TypeSize.PointerSize),

AssemblyCode.RegisterToSize(Register.bx, TypeSize.PointerSize)

},

RegularFunctionPointerRegisterSet =

new HashSet<Register>(VariadicFunctionPointerRegisterSet),

Byte1RegisterSet = new HashSet<Register>() {

Register.al,Register.ah, Register.bl, Register.bh,

Register.cl, Register.ch, Register.dl, Register.dh

},

Byte2RegisterSet = new HashSet<Register>() {

Register.ax, Register.bx, Register.cx, Register.dx

};

We remove the frame register from RegularFunctionPointerRegisterSet and VariadicFunctionPointer-RegisterSet, since we need it to point at the current activation record. We also remove the variadic register from VariadicFunctionPointerRegisterSet, since we need it to point at the activation record in a variadic function.

static RegisterAllocator() {

VariadicFunctionPointerRegisterSet.

Remove(AssemblyCode.RegularFrameRegister);

RegularFunctionPointerRegisterSet.

Remove(AssemblyCode.RegularFrameRegister);

RegularFunctionPointerRegisterSet.

Remove(AssemblyCode.VariadicFrameRegister);

}

The GetPossibleSet method returns the possible set a track, depending on whether the track holds a pointer, or the size of the track.

private static ISet<Register> GetPossibleSet(Track track) {

if (track.Pointer) {

if (SymbolTable.CurrentFunction.Type.IsVariadic()) {

return RegularFunctionPointerRegisterSet;

}

else {

return VariadicFunctionPointerRegisterSet;

}

}

If the track does not hold a pointer wee look into its size. If the size is one, we have a larger set to choose from. There are eight non-pointer registers of size one while there are four registers of the other sizes.

else if (track.MaxSize == 1) {

return Byte1RegisterSet;

}

We return the set of registers of size two, even if the size is actually larger than two. In that case, the RegisterToSize method in the AssemblyCode class will convert the register to the correct size.

else {

return Byte2RegisterSet;

}

}

}

}

## Assembly Code Generation

The AssemblyCodeGenerator class holds methods for generating the final assembly code. When a value is stored in a register, to be used by a later instruction, its symbol and track is added to the m\_trackMap map. The generated assembly code instructions are stored in the m\_assemblyCodeList list.

AssemblyCodeGenerator.cs

using System.Linq;

using System.Numerics;

using System.Collections.Generic;

namespace CCompiler {

public class AssemblyCodeGenerator {

public IDictionary<Symbol,Track> m\_trackMap =

new Dictionary<Symbol,Track>();

public List<AssemblyCode> m\_assemblyCodeList;

The m\_floatStackSize field keeps track of the size of the current floating-point stack. It can hold at most seven values.

private int m\_floatStackSize = 0;

public const int FloatingStackMaxSize = 7;

The MainName field holds the name of the main function.

public static string MainName = "main";

public static string InitializerName = Symbol.SeparatorId + "initializer";

public static string ArgsName = Symbol.SeparatorId + "args";

public static string PathName = Symbol.SeparatorId + "PathName";

The constructor takes the assembly code list.

public AssemblyCodeGenerator(List<AssemblyCode> assemblyCodeList) {

m\_assemblyCodeList = assemblyCodeList;

}

The RegisterAllocation method performs the register allocation by creating an object of the RegisterAllocator class. It may seem strange, but it actually works, since the constructor performs the register allocation and adds the register to the assembly code list.

private void RegisterAllocation(ISet<Track> trackSet) {

new RegisterAllocator(trackSet, m\_assemblyCodeList);

}

The static GenerateAssembly method generates the assembly code by creating an object of the AssemblyCodeGenerator class that generates the assembly code with tracks, and replacing the tracks with proper registers by performing register allocation.

public static void GenerateAssembly(List<MiddleCode> middleCodeList,

List<AssemblyCode> assemblyCodeList) {

AssemblyCodeGenerator objectCodeGenerator =

new AssemblyCodeGenerator(assemblyCodeList);

objectCodeGenerator.AssemblyCodeList(middleCodeList);

ISet<Track> trackSet = objectCodeGenerator.TrackSet();

objectCodeGenerator.RegisterAllocation(trackSet);

}

The first AddAssemblyCode method adds a new assembly instruction to the assembly code list m\_assemblyCodeList.

public AssemblyCode AddAssemblyCode(AssemblyOperator objectOp,

object operand0 = null, object operand1 = null,

object operand2 = null, int size = 0) {

AssemblyCode assemblyCode =

new AssemblyCode(objectOp, operand0, operand1, operand2, size);

m\_assemblyCodeList.Add(assemblyCode);

return assemblyCode;

}

The second AddAssemblyCode method adds a new assembly instruction to the given assembly code list. It is static and is called when generating special assembly code, such as initialization code and code for command line arguments.

public static AssemblyCode AddAssemblyCode(List<AssemblyCode> list,

AssemblyOperator objectOp, object operand0 = null,

object operand1 = null, object operand2 = null,

int size = 0) {

AssemblyCode assemblyCode =

new AssemblyCode(objectOp, operand0, operand1, operand2, size);

list.Add(assemblyCode);

return assemblyCode;

}

### The Long Switch Statement

The constructor iterates through the middle code list and calls an appropriate method for each kind of middle code instruction.

public void AssemblyCodeList(List<MiddleCode> middleCodeList){

for (int middleIndex = 0; middleIndex < middleCodeList.Count;

++middleIndex) {

MiddleCode middleCode = middleCodeList[middleIndex];

AddAssemblyCode(AssemblyOperator.new\_middle\_code, middleIndex);

If the current function is not null (if the code is located in function scope rather global scope), we add a label in order to make the assembly code more legible. If the middle code instruction is an initializer, we just add a label with the text of the middle code instruction.

if (SymbolTable.CurrentFunction != null) {

if ((middleCode.Operator == MiddleOperator.Initializer) ||

(middleCode.Operator == MiddleOperator.InitializerZero)) {

AddAssemblyCode(AssemblyOperator.label, null,

middleCode.ToString());

}

If the middle code instruction is not an initializer, we add a label with the unique name of the function. If the middle code index is greater the zero, we also add it to the label.

else {

string label = SymbolTable.CurrentFunction.UniqueName;

if (middleIndex > 0) {

label += Symbol.SeparatorId + middleIndex;

}

AddAssemblyCode(AssemblyOperator.label, label,

middleCode.ToString());

}

}

Now follows the long switch statement, where each middle code instruction is matched to a function call that generates the corresponding assembly code.

switch (middleCode.Operator) {

case MiddleOperator.CallHeader:

FunctionPreCall(middleCode);

break;

case MiddleOperator.Call:

FunctionCall(middleCode, middleIndex);

break;

case MiddleOperator.PostCall:

FunctionPostCall(middleCode);

break;

case MiddleOperator.Return:

Return(middleCode, middleIndex);

break;

case MiddleOperator.Exit:

Exit(middleCode);

break;

case MiddleOperator.Jump:

Jump(middleCode);

break;

case MiddleOperator.AssignRegister:

LoadToRegister(middleCode);

break;

case MiddleOperator.InspectRegister:

InspectRegister(middleCode);

break;

case MiddleOperator.JumpRegister:

JumpToRegister(middleCode);

break;

case MiddleOperator.Interrupt:

Interrupt(middleCode);

break;

case MiddleOperator.SysCall:

SystemCall(middleCode);

break;

case MiddleOperator.Initializer:

Initializer(middleCode);

break;

case MiddleOperator.InitializerZero:

InitializerZero(middleCode);

break;

In case of assignment, we inspect the type of the symbol to be assigned. If it is a struct or union, we call StructUnionAssign. Otherwise, the type must be integral, and we call IntegralAssign. Note that the assignment middle code instruction is not used for assignment of floating values. In that case, we instead top or pop the floating value stack.

case MiddleOperator.AssignInit:

StructUnionAssignInit(middleCode);

break;

case MiddleOperator.Assign: {

Symbol symbol = (Symbol) middleCode[0];

if (symbol.Type.IsStructOrUnion()) {

StructUnionAssign(middleCode, middleIndex);

}

else {

IntegralAssign(middleCode);

}

}

break;

case MiddleOperator.BitwiseAnd:

case MiddleOperator.BitwiseOr:

case MiddleOperator.BitwiseXOr:

case MiddleOperator.ShiftLeft:

case MiddleOperator.ShiftRight:

IntegralBinary(middleCode);

break;

In case of addition or subtract, we check the type of the left operand. If it is a floating value, we call FloatingBinary. Otherwise, the type is integral, and we call IntegralBinary.

case MiddleOperator.BinaryAdd:

case MiddleOperator.BinarySubtract: {

Symbol resultSymbol = (Symbol) middleCode[1];

if (resultSymbol.Type.IsFloating()) {

FloatingBinary(middleCode);

}

else {

IntegralBinary(middleCode);

}

}

break;

We have a similar case when performing multiplication or division. If case of floating values, we call FloatingBinary (the same method as in the floating addition and subtracting case). Otherwise, the type is integral, and we call IntegralMultiply.

case MiddleOperator.SignedMultiply:

case MiddleOperator.SignedDivide:

case MiddleOperator.SignedModulo:

case MiddleOperator.UnsignedMultiply:

case MiddleOperator.UnsignedDivide:

case MiddleOperator.UnsignedModulo: {

Symbol resultSymbol = (Symbol) middleCode[0];

if (resultSymbol.Type.IsFloating()) {

FloatingBinary(middleCode);

}

else {

IntegralMultiply(middleCode);

}

}

break;

case MiddleOperator.Carry:

case MiddleOperator.NotCarry:

CarryExpression(middleCode);

break;

In case of equality and relational operators, we call FloatingRelation in case of floating values and IntegralBinary otherwise.

case MiddleOperator.Equal:

case MiddleOperator.NotEqual:

case MiddleOperator.SignedLessThan:

case MiddleOperator.SignedLessThanEqual:

case MiddleOperator.SignedGreaterThan:

case MiddleOperator.SignedGreaterThanEqual:

case MiddleOperator.UnsignedLessThan:

case MiddleOperator.UnsignedLessThanEqual:

case MiddleOperator.UnsignedGreaterThan:

case MiddleOperator.UnsignedGreaterThanEqual: {

Symbol leftSymbol = (Symbol) middleCode[1];

if (leftSymbol.Type.IsFloating()) {

FloatingRelation(middleCode);

}

else {

IntegralBinary(middleCode);

}

}

break;

case MiddleOperator.Case:

Case(middleCode);

break;

case MiddleOperator.CaseEnd:

CaseEnd(middleCode);

break;

The same goes for the unary operations. In case of floating type we call FloatingUnary, and in case of integral type we call InategralUnary.

case MiddleOperator.UnaryAdd:

case MiddleOperator.UnarySubtract:

case MiddleOperator.BitwiseNot: {

Symbol resultSymbol = (Symbol) middleCode[0];

if (resultSymbol.Type.IsFloating()) {

FloatingUnary(middleCode);

}

else {

IntegralUnary(middleCode);

}

}

break;

case MiddleOperator.Address:

Address(middleCode);

break;

case MiddleOperator.Dereference: {

Symbol symbol = (Symbol) middleCode[1];

if (symbol.Type.IsFloating()) {

FloatingDereference(middleCode);

}

else {

IntegralDereference(middleCode);

}

}

break;

case MiddleOperator.DecreaseStack:

Assert.ErrorXXX((--m\_floatStackSize) >= 0);

break;

case MiddleOperator.CheckTrackMapFloatStack:

Assert.ErrorXXX((m\_trackMap.Count == 0) &&

(m\_floatStackSize == 0));

break;

case MiddleOperator.PushZero:

PushSymbol(new Symbol(Type.DoubleType, (decimal) 0));

break;

case MiddleOperator.PushOne:

PushSymbol(new Symbol(Type.DoubleType, (decimal) 1));

break;

case MiddleOperator.PushFloat:

PushSymbol((Symbol) middleCode[0]);

break;

case MiddleOperator.TopFloat:

TopPopSymbol((Symbol) middleCode[0], TopOrPop.Top);

break;

case MiddleOperator.PopFloat:

TopPopSymbol((Symbol) middleCode[0], TopOrPop.Pop);

break;

case MiddleOperator.PopEmpty:

PopEmpty();

break;

case MiddleOperator.IntegralToIntegral:

IntegralToIntegral(middleCode, middleIndex);

break;

case MiddleOperator.IntegralToFloating:

IntegralToFloating(middleCode);

break;

case MiddleOperator.FloatingToIntegral:

FloatingToIntegral(middleCode);

break;

case MiddleOperator.ParameterInitSize:

StructUnionParameterInit(middleCode);

break;

case MiddleOperator.Parameter: {

Symbol paramSymbol = (Symbol) middleCode[2];

if (paramSymbol.Type.IsFloating()) {

FloatingParameter(middleCode);

}

else if (paramSymbol.Type.IsStructOrUnion()) {

StructUnionParameter(middleCode, middleIndex);

}

else {

IntegralParameter(middleCode);

}

}

break;

case MiddleOperator.GetReturnValue: {

Symbol returnSymbol = (Symbol) middleCode[0];

if (returnSymbol.Type.IsStructOrUnion()) {

StructUnionGetReturnValue(middleCode);

}

else if (returnSymbol.Type.IsFloating()) {

Assert.Error((++m\_floatStackSize) <= FloatingStackMaxSize,

null, Message.Floating\_stack\_overflow);

}

else {

IntegralGetReturnValue(middleCode);

}

}

break;

}

}

}

### Track Set Generation

When the assembly code has been generated, a lot of tracks has been added to the assembly code. The task of the TrackSet method is to pick up the tracks from the assembly code and generate the set of tracks. Each track holds a list of entries, specifying the positions of the track in the code.

private ISet<Track> TrackSet() {

ISet<Track> trackSet = new HashSet<Track>();

for (int index = 0; index < m\_assemblyCodeList.Count; ++index) {

AssemblyCode assemblyCode = m\_assemblyCodeList[index];

In case of the set-track-size instruction, we set the current size of its track and replace the instruction with an empty instruction. The current size is then used when adding entries to the track so that the registers in the end are given the correct size.

if (assemblyCode.Operator == AssemblyOperator.set\_track\_size) {

Track track = (Track)assemblyCode[0];

if (assemblyCode[1] is int) {

track.MaxSize = (int) assemblyCode[1];

}

else {

track.MaxSize = ((Track) assemblyCode[1]).MaxSize;

}

}

In all other cases, we exam the operators of the instruction by calling CheckTrack.

else {

CheckTrack(trackSet, assemblyCode, 0, index);

CheckTrack(trackSet, assemblyCode, 1, index);

CheckTrack(trackSet, assemblyCode, 2, index);

}

}

return trackSet;

}

The CheckTrack method takes the track set, one of the operands in an assembly code instruction, the position of the operand (zero, one, or two), and the index of the assembly code instruction in the assembly code list. We need to set the index of the assembly code list in the track in order to later decide whether tracks overlap.

private void CheckTrack(ISet<Track> trackSet, AssemblyCode assemblyCode,

int position, int index) {

if (assemblyCode[position] is Track) {

Track track = (Track) assemblyCode[position];

trackSet.Add(track);

track.Index = index;

}

}

### Function Calls

The BaseRegister method returns the base register of a symbol representing an auto or register variable or a parameter. However, the symbol can also be null, in case of an unnamed parameter. The method returns the regular frame pointer or the variadic frame pointer. Note that we always use the regular frame pointer in regular functions.

public Register BaseRegister(Symbol symbol) {

Assert.ErrorXXX((symbol == null) || symbol.IsAutoOrRegister());

We return the variadic frame pointer if the function is variadic and the symbol is not a parameter. If the symbol is null, it represents an unnamed parameter in which case the variadic frame pointer is returned.

if (SymbolTable.CurrentFunction.Type.IsVariadic() &&

(symbol != null) && !symbol.IsParameter()) {

return AssemblyCode.VariadicFrameRegister;

}

If the function is not variadic, or if the symbol is a parameter, we return the regular frame pointer. The activation record is organized so that the parameters are located before the variables, with potential extra parameter between the regular parameters and the variables. Therefore, we use the regular frame pointer for parameters in both variadic and regular functions, and the variadic frame pointer for variables in variadic functions since they are located after the extra parameters.

else {

return AssemblyCode.RegularFrameRegister;

}

}

The FunctionPreCall methods is called before the actual function call, and before the parameter list of the function call. The idea is that each expression in the parameters shall have access to all the registers and the whole floating-point value stack, that no parameter values are stored in the registers or at the floating-point stack. Therefore, FunctionPreCall storer the values currently stored in registers and on the floating-point value stack at the activation record. The FunctionPostCall method below restores the values after the function call.

The m\_totalExtraSize field keeps track of the current record size, which increases with nested calls. The m\_recordSizeStack field holds the record sizes of nested calls. Techically, we could manage with only the stack, and sum the record sizes to find the total record size. However, we include m\_totalExtraSize for efficiency reasons.

private int m\_totalExtraSize = 0;

private Stack<int> m\_recordSizeStack = new Stack<int>();

As mentioned above, each parameter expression shall have access to all registers. Therefore, we need to save the track map of each nested call in the m\_trackMapStack field. We need also keep track of the position of each saved register on the activation record in m\_registerMapStack. Both stacks are popped in the PostFunctionCall method below.

private Stack<IDictionary<Symbol,Track>> m\_trackMapStack =

new Stack<IDictionary<Symbol,Track>>();

private Stack<IDictionary<Track,int>> m\_registerMapStack =

new Stack<IDictionary<Track,int>>();

In FunctionPreCall, we start by obtaining the base register, which is the regular or variadic frame pointer, depending on whether the function is variadic. The recordSize variable is the original size of the activation record while the extraSize field holds the size of the extra values in the registers and on the floating-point value stack.

public void FunctionPreCall(MiddleCode middleCode) {

Register baseRegister = BaseRegister(null);

int recordSize = (int) middleCode[0], extraSize = 0;

We load the post map with the tracks of the register with their locations on the activation record. For each register to be stored on the activation record, we add a mov instruction. The extraSize field is increased for each register.

int totalSize = 0;

int doubleTypeSize = Type.DoubleType.Size();

foreach (int size in m\_topStack) {

totalSize += size;

}

extraSize += totalSize \* doubleTypeSize;

IDictionary<Track,int> registerMap = new Dictionary<Track,int>();

foreach (KeyValuePair<Symbol, Track> pair in m\_trackMap) {

Track track = pair.Value;

AddAssemblyCode(AssemblyOperator.mov, baseRegister,

recordSize + extraSize, track);

registerMap.Add(track, recordSize + extraSize);

Symbol symbol = pair.Key;

extraSize += symbol.Type.Size();

}

After the registers have been stored on the activation record, we store the values of the floating-point value stack on the activation record. Again, the extraSize field is being increased with the size of the values.

int topSize = m\_floatStackSize - totalSize;

m\_topStack.Push(topSize);

for (int count = 0; count < topSize; ++count) {

AddAssemblyCode(AssemblyOperator.fstp\_qword, baseRegister,

recordSize + extraSize);

extraSize += doubleTypeSize;

}

When the values have been stored on the activation record, we need to store the register map for the registers to be restored after the function call in the PostFunctionCall method below.

m\_registerMapStack.Push(registerMap);

m\_recordSizeStack.Push(extraSize);

We also use the m\_totalExtraSize field to store the total extra size for the in case of nested calls. Technically, we do not need this field since we could obtain the same value by summarizing the values pushed on the m\_recordSizeStack stack. However, it has been included for efficiency reasons.

m\_totalExtraSize += extraSize;

The track map is pushed on the stack, and a new object is instantiated as the new track map. The track map will be restored after the function call.

m\_trackMapStack.Push(m\_trackMap);

m\_trackMap = new Dictionary<Symbol, Track>();

}

The m\_returnFloating is set to true if the function returns a floating value. In case that the value is stored on the floating-point value stack and need to be considered by the PostCallFunction method below.

private bool m\_returnFloating = false;

When calling a function there are several cases to consider: the function symbol holding the function to be called may be a proper function or a pointer to a function, and the caller or callee function (or both) may be variadic.

public void FunctionCall(MiddleCode middleCode, int index) {

int recordSize = ((int) middleCode[0]) + m\_totalExtraSize;

Symbol calleeSymbol = (Symbol) middleCode[1];

int extraSize = (int) middleCode[2];

The type is the function or, in case of a pointer the pointer type, the function the pointer points at.

Type calleeType = calleeSymbol.Type.IsFunction()

? calleeSymbol.Type : calleeSymbol.Type.PointerType;

Both the caller and callee function may be variadic.

bool callerVariadic = SymbolTable.CurrentFunction.Type.IsVariadic(),

calleeVariadic = calleeType.IsVariadic();

The frame register is the variadic frame pointer register if the caller function is variadic. If it not, it is the regular frame pointer register.

Register frameRegister = callerVariadic

? AssemblyCode.VariadicFrameRegister

: AssemblyCode.RegularFrameRegister;

We start by adding the assembly code instruction for the return address. To begin with, we set it to the index of the next middle code instruction. The address will later be changed by the WindowsJumpInfo method to a proper assembly code address.

AddAssemblyCode(AssemblyOperator.return\_address, frameRegister,

recordSize + SymbolTable.ReturnAddressOffset,

(BigInteger) (index + 1));

We set the current regular frame pointer and, in case of a variadic caller function, the variadic pointer to their offset in the activations record of the callee function. In this way, we can reset their values in accordance with the caller function returning from the function call.

AddAssemblyCode(AssemblyOperator.mov, frameRegister,

recordSize + SymbolTable.RegularFrameOffset,

AssemblyCode.RegularFrameRegister);

if (callerVariadic) {

AddAssemblyCode(AssemblyOperator.mov, frameRegister,

recordSize + SymbolTable.VariadicFrameOffset,

AssemblyCode.VariadicFrameRegister);

}

If the callee function is a pointer to a function, we load its value into the jumpTrack track. We must load the value into the track before we increase the frame register below.

Track jumpTrack = null;

if (!calleeSymbol.Type.IsFunction()) {

jumpTrack = LoadValueToRegister(calleeSymbol);

}

We add the size of the caller function’s activation record to the frame pointer, so that it pointes at the activation record of the callee function.

AddAssemblyCode(AssemblyOperator.add, frameRegister, // add di, 10

(BigInteger) recordSize);

If the caller function is variadic, the frame pointer that we just added is the variadic frame pointer, and we also need to set the regular frame pointer to the same value.

if (callerVariadic) { // mov bp, di

AddAssemblyCode(AssemblyOperator.mov,

AssemblyCode.RegularFrameRegister,

AssemblyCode.VariadicFrameRegister);

}

If the callee function, but not the caller function, is variadic, the frame pointer we just added is the variadic frame pointer, and we also need to set the regular frame pointer to the same value.

else if (calleeVariadic) {

AddAssemblyCode(AssemblyOperator.mov,

AssemblyCode.VariadicFrameRegister,

AssemblyCode.RegularFrameRegister);

}

If the callee function is variadic, and the extra size is more than zero (if there are extra arguments in the function calls that are not matched to the declared parameters), we add the size to the variadic frame pointer. The callee function will have both a regular frame pointer and a variadic frame pointer. However, the variadic pointer will point to a higher address to give space to the extra arguments.

if (calleeVariadic && (extraSize > 0)) {

AddAssemblyCode(AssemblyOperator.add,

AssemblyCode.VariadicFrameRegister,

(BigInteger) extraSize);

}

Finally. if the callee function is a function (not a pointer to a function) we a middle code call instruction. The m\_returnFloating field is set to true if the callee function returns a floating value. It will later be inspected by the FunctionPostCall method.

if (calleeSymbol.Type.IsFunction()) {

AddAssemblyCode(AssemblyOperator.call, calleeSymbol.UniqueName);

m\_returnFloating = calleeSymbol.Type.ReturnType.IsFloating();

}

If the callee function is a pointer to a function and jump to the address stored in jumpTrack.

else {

AddAssemblyCode(AssemblyOperator.jmp, jumpTrack);

m\_returnFloating =

calleeSymbol.Type.PointerType.ReturnType.IsFloating();

}

}

The FunctionPostCall method is called after a function call. Its task is to restore the actions performed by FunctionPreCall above.

public void FunctionPostCall(MiddleCode middleCode) {

Register baseRegister = BaseRegister(null);

m\_trackMap = m\_trackMapStack.Pop();

IDictionary<Track,int> registerMap = m\_registerMapStack.Pop();

First, we iterate through the post map defined in FunctionPreCall before the function call. We load the values stored in the map into the registers.

foreach (KeyValuePair<Track,int> pair in registerMap) {

Track track = pair.Key;

int offset = pair.Value;

AddAssemblyCode(AssemblyOperator.mov, track, baseRegister,offset);

}

Then we investigate the floating value stack. If floating stack value before the function call was non-empty, we need to restore it.

Assert.ErrorXXX(m\_topStack.Count > 0);

int topSize = m\_topStack.Pop();

if (topSize > 0) {

int recordOffset = (int) middleCode[0];

int doubleTypeSize = Type.DoubleType.Size();

int recordSize = m\_recordSizeStack.Pop();

If the return value of the previous function call is a floating value, we need to temporary restore it when we restore the floating value stack, since the return value shall be placed at the top of the floating stack.

if (m\_returnFloating) {

AddAssemblyCode(AssemblyOperator.fstp\_qword, baseRegister,

recordOffset + recordSize);

}

We iterate through the floating values, as they are located above the activation record. We begin by pushing the top-most value and iterate to the bottom-most value.

int currentOffset = recordOffset + recordSize;

for (int count = 0; count < topSize; ++count) {

currentOffset -= doubleTypeSize;

AddAssemblyCode(AssemblyOperator.fld\_qword, baseRegister,

currentOffset);

}

If the function returned a floating value, we push it from its temporary location to the top of the floating value stack.

if (m\_returnFloating) {

AddAssemblyCode(AssemblyOperator.fld\_qword, baseRegister,

recordOffset + recordSize);

}

m\_totalExtraSize -= recordSize;

}

else {

m\_totalExtraSize -= m\_recordSizeStack.Pop();

}

}

### Loading Values into Registers

The LoadValueToRegister method loads the value of a symbol into a register. A specific register to be loaded may be given. However, the returned register is represented by a track that will be replaced by a proper register by the register allocator.

public Track LoadValueToRegister(Symbol symbol,

Register? register = null) {

If the register is specified, we must check that no other value is stored in that register. If it is, we must move it to another register.

if (register != null) {

CheckRegister(symbol, register.Value);

}

Then we check if the value is already stored in a register. If it is, we look up it track and compare the specified register with the register of the track.

Track track;

if (m\_trackMap.TryGetValue(symbol, out track)) {

m\_trackMap.Remove(symbol);

If the registers are not null do not overlap, we need to remove the previous register from the track. If we find that the register is associated to a track, we add a mov instruction that moves the value to a new (yet unknown) register. The track holding that register is then returned.

if ((register != null) && (track.Register != null) &&

!AssemblyCode.RegisterOverlap(register, track.Register)) {

Track newTrack = new Track(symbol, register.Value);

AddAssemblyCode(AssemblyOperator.set\_track\_size,

newTrack, track);

AddAssemblyCode(AssemblyOperator.mov, newTrack, track);

return newTrack;

}

If there is no non-overlapping register, we set the register of the track, if it is not null, and return the track.

else {

if (register != null) {

track.Register = register;

}

return track;

}

}

If the track register is null, we assign it the specified register (which may or may not be null).

However, if both the specified register and the track register is non-null, and they do not overlap each other, we need to move the value of the track register to specified register. We create and return a new track for the specified register.

If the symbol does not already have a track, we create a new track for it.

Basically, we have three different cases: the symbol holds an integer value, it holds and address, or a value stored at an address. In case of an integer value, we simply load it int to the register.

if (symbol.Value is BigInteger) {

AddAssemblyCode(AssemblyOperator.mov, track, symbol.Value);

}

If the symbol is an array, function, or string, or if it holds a static address, we load the address of the value into the register, not the value itself. We load the base of the symbol into the register. The base is the symbol’s unique name if it is extern or static, the regular or variadic frame register in case of an auto or register array (a function or string is always static), and the name of a static address. If the offset of the address is non-zero, we add it to the register.

else if (symbol.Type.IsArrayFunctionOrString() ||

(symbol.Value is StaticAddress)) {

AddAssemblyCode(AssemblyOperator.mov, track,

Base(symbol));

int offset = Offset(symbol);

if (offset != 0) {

AddAssemblyCode(AssemblyOperator.add, track,

(BigInteger) offset);

}

}

In all other cases, we load the value of the symbol into the register. The base is the symbol’s unique name if it is extern or static, and the regular or variadic frame register is it is auto or register.

else {

AddAssemblyCode(AssemblyOperator.mov, track,

Base(symbol), Offset(symbol));

}

return track;

}

}

If the symbol is a static address, we load its address into the register. At the moment, we load its name into the register and add its offset, if more than zero. The name will later be replaced by the address by the linker.

If the symbol is an array, we load its base into the register and add its offset, if more than zero. The base may be the unique name of a static symbol, or the regular or variadic frame pointer in case of a parameter or local variable.

If the symbol is not a constant, an array, or a static function, struct, or union, we load the value of the symbol to the register with the help of the symbol’s base and offset. The base may be the unique name of a static symbol, or the regular or variadic frame pointer.

The CheckRegister method checks whether a symbol value is stored in the register

public void CheckRegister(Symbol symbol, Register register) {

foreach (KeyValuePair<Symbol,Track> entry in m\_trackMap) {

Symbol oldSymbol = entry.Key;

Track oldTrack = entry.Value;

We iterate through the track map and check if any track holds a register overlapping the specified register.

if (!oldSymbol.Equals(symbol) &&

AssemblyCode.RegisterOverlap(register, oldTrack.Register)) {

If the find a match, we create a new track and insert instructions for setting the size of the track and moving the value from the previous track to the new track.

Track newTrack = new Track(oldSymbol);

m\_trackMap[oldSymbol] = newTrack;

int lastLine;

for (lastLine = m\_assemblyCodeList.Count - 1; lastLine >= 0;

--lastLine) {

AssemblyCode assemblyCode = m\_assemblyCodeList[lastLine];

if (oldTrack.Equals(assemblyCode[0])) {

break;

}

}

Assert.ErrorXXX(lastLine >= 0);

AssemblyCode setCode =

new AssemblyCode(AssemblyOperator.set\_track\_size,

newTrack, oldTrack);

AssemblyCode movCode =

new AssemblyCode(AssemblyOperator.mov, newTrack, oldTrack);

m\_assemblyCodeList.Insert(lastLine + 1, setCode);

m\_assemblyCodeList.Insert(lastLine + 2, movCode);

break;

}

}

}

The LoadAddressToRegister loads the address of value, rather the value itself, into a register.

public Track LoadAddressToRegister(Symbol symbol,

Register? register = null) {

If the address symbol of the symbol is not null, we simply call LoadValueToRegister with the address symbol, and returns the track. However, we must mark the track as a pointer, which means that the set of possible register re will be restricted.

if (symbol.AddressSymbol != null) {

Track addressTrack = LoadValueToRegister(symbol.AddressSymbol);

addressTrack.Pointer = true;

return addressTrack;

}

If the address symbol is null, we need to obtain the address manually. We start by mov the base of the symbol to the register. The base is the regular or variadic frame pointer in case of an auto or register storage, since the value is located on the activation record, or the name of the symbol in case of extern or static storage. However, the symbol may have a static address its value. In that case, the base is the name of the static address.

else {

Symbol addressSymbol = new Symbol(new Type(symbol.Type));

Track addressTrack = new Track(addressSymbol, register);

Assert.ErrorXXX((addressTrack.Register == null) ||

RegisterAllocator.VariadicFunctionPointerRegisterSet.

Contains(addressTrack.Register.Value));

addressTrack.Pointer = true;

Assert.ErrorXXX(!(symbol.Value is BigInteger));

AddAssemblyCode(AssemblyOperator.mov, addressTrack, Base(symbol));

Then we add the offset of the symbol, which is always non-zero in case of auto or register storage and zero in case of extern or static storage. In case a static address, the offset may be zero or non-zero.

int offset = Offset(symbol);

if (offset != 0) {

AddAssemblyCode(AssemblyOperator.add, addressTrack,

(BigInteger) offset);

}

return addressTrack;

}

}

### Return, Exit, and Jump

The Return method generates the code for returning from a function call. We catch the return address (which we must do before we reset the pointers), reset the regular frame pointer and variadic frame pointer, and jump back to the return address.

public void Return(MiddleCode middleCode, int middleIndex) {

if (SymbolTable.CurrentFunction.UniqueName.Equals

(AssemblyCodeGenerator.MainName)) {

Assert.ErrorXXX(m\_floatStackSize == 0);

AddAssemblyCode(AssemblyOperator.cmp, AssemblyCode.RegularFrameRegister,

SymbolTable.ReturnAddressOffset, BigInteger.Zero,

TypeSize.PointerSize);

AssemblyCode jumpCode =

AddAssemblyCode(AssemblyOperator.je, null, null, middleIndex + 1);

Return();

}

else {

SetReturnValue(middleCode);

Assert.ErrorXXX(m\_floatStackSize == 0);

Return();

}

}

The SetReturnValue method calls StructUnionSetReturnValue or IntegralSetReturnValue in case of a struct or union value, or an integral value. In case of a floating-point value we do nothing, since the return value is already properly placed on the top of the floating-point stack.

private void SetReturnValue(MiddleCode middleCode) {

if (middleCode[1] != null) {

Symbol returnSymbol = (Symbol) middleCode[1];

if (returnSymbol.Type.IsStructOrUnion()) {

StructUnionSetReturnValue(middleCode);

}

else if (returnSymbol.Type.IsFloating()) {

Assert.ErrorXXX((--m\_floatStackSize) == 0);

}

else {

IntegralSetReturnValue(middleCode);

}

}

}

The Return method adds assembly code for return control back to the calling function. We restore the frame and variadic pointers of the calling function and jump back to the calling function. Note that we have to load the return jump address into register before we restore the frame pointer, otherwise we would obtain the return address of the calling function.

private void Return() {

Track track = new Track(Type.VoidPointerType);

AddAssemblyCode(AssemblyOperator.mov, track,

AssemblyCode.RegularFrameRegister,

SymbolTable.ReturnAddressOffset);

AddAssemblyCode(AssemblyOperator.mov,

AssemblyCode.VariadicFrameRegister,

AssemblyCode.RegularFrameRegister,

SymbolTable.VariadicFrameOffset);

AddAssemblyCode(AssemblyOperator.mov, AssemblyCode.RegularFrameRegister,

AssemblyCode.RegularFrameRegister,

SymbolTable.RegularFrameOffset);

AddAssemblyCode(AssemblyOperator.jmp, track);

}

The IntegralGetReturnValue method stores the return register in a track.

public void IntegralGetReturnValue(MiddleCode middleCode) {

Symbol returnSymbol = (Symbol) middleCode[0];

Register returnRegister =

AssemblyCode.RegisterToSize(AssemblyCode.ReturnValueRegister,

returnSymbol.Type.Size());

We call CheckRegister to make sure that if the register already has a value, it is moved to another register.

CheckRegister(returnSymbol, returnRegister);

Track returnTrack = new Track(returnSymbol, returnRegister);

m\_trackMap.Add(returnSymbol, returnTrack);

AddAssemblyCode(AssemblyOperator.empty, returnTrack);

}

The IntegralSetReturnValue loads the symbol of the expression into the return register.

public void IntegralSetReturnValue(MiddleCode middleCode) {

Symbol returnSymbol = (Symbol) middleCode[1];

Register returnRegister =

AssemblyCode.RegisterToSize(AssemblyCode.ReturnValueRegister,

returnSymbol.Type.SizeArray());

LoadValueToRegister(returnSymbol, returnRegister);

m\_trackMap.Remove(returnSymbol);

}

The Exit method generates code for exit the execution of the program. However, the code is different depending on whether we generate Windows or Linux code, and whether there is an exit symbol given.

public void Exit(MiddleCode middleCode) {

Symbol exitSymbol = (Symbol) middleCode[0];

If the exit symbol is not null, we load its value into the rdi register If it is null, we just load zero into the register. This value will be returned to the surrounding operation system.

if (Start.Linux) {

if (exitSymbol != null) {

LoadValueToRegister(exitSymbol, Register.rdi);

}

else {

AddAssemblyCode(AssemblyOperator.mov, Register.rdi,

BigInteger.Zero);

}

The Linux exit code is to load value 60 to register rax and do a system call.

AddAssemblyCode(AssemblyOperator.mov, Register.rax,

(BigInteger) 60); // 0x3C

AddAssemblyCode(AssemblyOperator.syscall);

}

The code for the Windows envisonment is described in Chapter 13.

if (Start.Windows) {

// ...

}

}

The Jump method simply add a jump instruction with the index a middle code instruction as target.

public void Jump(MiddleCode middleCode) {

int jumpTarget = (int) middleCode[0];

AddAssemblyCode(AssemblyOperator.jmp, null, null, jumpTarget);

}

### Load and Inspect Registers

The LoadToRegister method adds assembly code that loads the value of a symbol into a specified register. This method, as well as InspectRegister, CarryExpression, JumpRegister, Interrupt, and SystemCall, is only used by the standard library in internal system calls.

private ISet<Track> m\_syscallSet = new HashSet<Track>();

public void LoadToRegister(MiddleCode middleCode) {

Register register = (Register) middleCode[0];

Symbol symbol = (Symbol) middleCode[1];

Track track = LoadValueToRegister(symbol, register);

m\_syscallSet.Add(track);

}

The InspectRegister method loads the value of a register to a symbol. More specifically, it adds a track holding the symbol and register to the track map.

public void InspectRegister(MiddleCode middleCode) {

Symbol symbol = (Symbol) middleCode[0];

Register register = (Register) middleCode[1];

Track track = new Track(symbol, register);

m\_trackMap.Add(symbol, track);

}

The CarryExpression method adds assembly code that jumps to a middle code target if the carry flag is set. The target will later be changed by the WindowsJumpInfo method to a proper assembly code target.

public void CarryExpression(MiddleCode middleCode) {

AssemblyOperator objectOperator =

m\_middleToIntegralMap[middleCode.Operator];

int jumpTarget = (int) middleCode[0];

AddAssemblyCode(objectOperator, null, null, jumpTarget);

}

The JumpRegister method adds assembly code that jumps to the address stored in a register.

public void JumpToRegister(MiddleCode middleCode) {

Register jumpRegister = (Register) middleCode[0];

AddAssemblyCode(AssemblyOperator.jmp, jumpRegister);

}

The Interrupt method adds code that performs an interrupt call. Before the call, we clear the system call set.

public void Interrupt(MiddleCode middleCode) {

foreach (Track track in m\_syscallSet) {

AddAssemblyCode(AssemblyOperator.empty, track);

}

AddAssemblyCode(AssemblyOperator.interrupt,

(BigInteger) middleCode[0]);

m\_trackMap.Clear();

}

The SystemCall method adds code that performs an interrupt call. Like the interrupt case, we clear the system call set before the call.

public void SystemCall(MiddleCode middleCode) {

foreach (Track track in m\_syscallSet) {

AddAssemblyCode(AssemblyOperator.empty, track);

}

AddAssemblyCode(AssemblyOperator.syscall);

m\_trackMap.Clear();

}

### Initialization

The Initializer method adds code initializing a value. The value becomes static block, or a part of a static block, which

private void Initializer(MiddleCode middleCode) {

Sort sort = (Sort) middleCode[0];

object value = middleCode[1];

If the value is a static address, we add code for defining the address with its name and offset.

if (value is StaticAddress) {

StaticAddress staticAddress = (StaticAddress) value;

string name = staticAddress.UniqueName;

int offset = staticAddress.Offset; // dw name + offset

AddAssemblyCode(AssemblyOperator.define\_address, name, offset);

}

Otherwise, we add code for defining the value. The initialization instruction will later be transformed to a sequence of instructions in the Linux case and memory block in the Windows case.

else {

AddAssemblyCode(AssemblyOperator.define\_value, sort, value);

}

}

The InitializerZero adds an instruction for a sequence of zero values, each holding one byte. Like Initializer above, the initialization instruction will later be transformed to a sequence of instructions in the Linux case and memory block in the Windows case.

private void InitializerZero(MiddleCode middleCode) {

int size = (int) middleCode[0];

Assert.ErrorXXX(size > 0);

AddAssemblyCode(AssemblyOperator.define\_zero\_sequence, size);

}

### Integral Multiplication, Division, and Modulo

The IntegralMultiply method adds assembly code for the integral operators multiplication, division, and module. These operations differ from the previous integral operations in that way that the value of the left operand is always stored in a specific register. However, the right symbol is given in the operation. The value of the right symbol is given in the operations. It can be stored in a register or on a memory address. However, unlike the previous integral operations, we cannot give an integer value directly. Instead, we have to store the value on an address, and give the address to the operation.

Moreover, there are in fact four registers involved in the operation. The specific registers depend on the type size.

The register where we store the value of the left symbol.

The register holding the upper part of the left symbol value. We ignore the value in this register, but we must set it to zero and mark in the track map we have in fact altered its value.

The result of multiplication or division, which is ignored by modulo.

The result of module, which is ignored by multiplication or division.

Even if one of the last two registers is ignored, we still need to mark in the track map that is has in fact been assigned a new value.

The m\_leftRegisterMap map hold the registers where we store the value of the left symbol. The m\_zeroRegisterMap map holds the registers that are to be set to zero before the operation. The m\_productQuintentRegisterMap map holds the registers where the product or quintet is stored after a multiplication or division operations, while m\_remainderRegisterMap holds the registers of the remainder of a modulo operations.

public static IDictionary<int,Register> m\_leftRegisterMap =

new Dictionary<int,Register>() {{1, Register.al}, {2, Register.ax},

{4, Register.eax}, {8, Register.rax}};

public static IDictionary<int,Register> m\_zeroRegisterMap =

new Dictionary<int,Register>() {{1, Register.ah}, {2, Register.dx},

{4, Register.edx}, {8, Register.rdx}};

public static IDictionary<int,Register> m\_productQuintentRegisterMap =

new Dictionary<int,Register>() {{1, Register.al}, {2, Register.ax},

{4, Register.eax}, {8, Register.rax}};

public static IDictionary<int,Register> m\_remainderRegisterMap =

new Dictionary<int,Register>() {{1, Register.ah}, {2, Register.dx},

{4, Register.edx}, {8, Register.rdx}};

The IntegralMultiply method adds code for multiplication operations. To begin with, we load the value of the left symbol into the register given by m\_leftRegisterMap.

public void IntegralMultiply(MiddleCode middleCode) {

Symbol leftSymbol = (Symbol) middleCode[1];

int typeSize = leftSymbol.Type.SizeArray();

Register leftRegister = m\_leftRegisterMap[typeSize];

Track leftTrack = LoadValueToRegister(leftSymbol, leftRegister);

Then we clear the zero register by perform the exclusive operation on itself, which is an effective way to clear a register, in order to provide current input to the operation.

Register zeroRegister = m\_zeroRegisterMap[typeSize];

Track zeroTrack = new Track(leftSymbol, zeroRegister);

AddAssemblyCode(AssemblyOperator.xor, zeroTrack, zeroTrack);

The we call IntegralUnary to load the right symbol into a suitable storage and perform the operation.

Symbol rightSymbol = (Symbol) middleCode[2];

IntegralUnary(middleCode.Operator, rightSymbol, rightSymbol);

When the operation is done, we need to find out which register to keep and which one to discard. We only use one of the registers, depending on the operation. But we also need to mark that the value of the discarded register has been modified, to prevent that another value is stored in the register during the operation.

Register resultRegister, discardRegister;

In case of the modulo operation, the resulting register is picked from m\_remainderRegisterMap, and the register to be discarded m\_productQuintentRegisterMap. In case of multiplication or division, it is the other way around.

if ((middleCode.Operator == MiddleOperator.SignedModulo) ||

(middleCode.Operator == MiddleOperator.UnsignedModulo)) {

resultRegister = m\_remainderRegisterMap[typeSize];

discardRegister = m\_productQuintentRegisterMap[typeSize];

}

else {

resultRegister = m\_productQuintentRegisterMap[typeSize];

discardRegister = m\_remainderRegisterMap[typeSize];

}

We create a new track holding the result symbol and register. If the result symbol is a temporary variable, we store the track in the track map. We also add an empty instruction, to mark that the register has been altered by the operation.

Symbol resultSymbol = (Symbol) middleCode[0];

Track resultTrack = new Track(resultSymbol, resultRegister);

if (resultSymbol.IsTemporary()) {

Assert.ErrorXXX(resultSymbol.AddressSymbol == null);

m\_trackMap.Add(resultSymbol, resultTrack);

AddAssemblyCode(AssemblyOperator.empty, resultTrack);

}

If the result symbol is a regular variable, we store the value in the register at the address of the variable. In this case we do not need to add an empty instruction since the loading of the register marks that is has been in use.

else {

AddAssemblyCode(AssemblyOperator.mov, Base(resultSymbol),

Offset(resultSymbol), resultTrack);

}

However, we need to add an empty instruction for the discarded register, to mark that it has been altered.

Track discaredTrack = new Track(resultSymbol, discardRegister);

AddAssemblyCode(AssemblyOperator.empty, discaredTrack);

}

### Integral Assignment and Parameters

The IntegralAssign method generates code for assignment of an integral value. What makes it a bit complicated is that we must take into consideration of the assignment of a conditional expression. Both the values of the true and false expression shall be assigned to the same register.

public void IntegralAssign(MiddleCode middleCode) {

Symbol resultSymbol = (Symbol) middleCode[0],

assignSymbol = (Symbol) middleCode[1];

IntegralAssign(resultSymbol, assignSymbol);

}

The IntegralParameter method is quite simple. We only need to create the symbol representing the parameter, and then call IntegralAssign.

public void IntegralParameter(MiddleCode middleCode) {

Type toType = (Type) middleCode[1];

Symbol toSymbol = new Symbol(null, false, Storage.Auto, toType);

Symbol fromSymbol = (Symbol) middleCode[2];

int parameterOffset = (int) middleCode[0];

toSymbol.Offset = m\_totalExtraSize + parameterOffset;

IntegralAssign(toSymbol, fromSymbol);

}

The IntegralAssign method is called by IntegralAssign and IntegralParameter.

public void IntegralAssign(Symbol resultSymbol, Symbol assignSymbol) {

Track resultTrack = null, assignTrack = null;

m\_trackMap.TryGetValue(resultSymbol, out resultTrack);

m\_trackMap.TryGetValue(assignSymbol, out assignTrack);

int typeSize = assignSymbol.Type.SizeArray();

If the result symbol is temporary and its address symbol is null, we are facing the assignment of a conditional expression.

if (resultSymbol.IsTemporary()) {

Assert.ErrorXXX(assignTrack == null);

if (resultTrack == null) {

resultTrack = new Track(resultSymbol);

m\_trackMap.Add(resultSymbol, resultTrack);

}

If the assignment symbol is stored in the track map, it means that we assign the true expression of the conditional expression, we start by copying its track to the assignment track. Then we add the assignment track to the track map.

If the assignment symbol is not stored in the track map, we assign the false expression of the conditional expression, we begin by creating a track for the result symbol.

If the assign symbol is an integer value, we move it into the register.

if (assignSymbol.Value is BigInteger) {

AddAssemblyCode(AssemblyOperator.mov, resultTrack,

assignSymbol.Value);

}

If the assign symbol is an array, function, string, or static address, we move its address rather than its value into the register.

else if (assignSymbol.Type.IsArrayFunctionOrString() ||

(assignSymbol.Value is StaticAddress)) {

AddAssemblyCode(AssemblyOperator.mov, resultTrack,

Base(assignSymbol));

int offset = Offset(assignSymbol);

if (offset != 0) {

AddAssemblyCode(AssemblyOperator.add, resultTrack,

(BigInteger) offset);

}

}

Otherwise, we move the value of the symbol into the register.

else {

AddAssemblyCode(AssemblyOperator.mov, resultTrack,

Base(assignSymbol), Offset(assignSymbol));

}

}

else {

if (assignTrack != null) {

AddAssemblyCode(AssemblyOperator.mov, Base(resultSymbol),

Offset(resultSymbol), assignTrack);

m\_trackMap.Remove(assignSymbol);

}

else if (assignSymbol.Value is BigInteger) {

BigInteger bigValue = (BigInteger) assignSymbol.Value;

if (Start.Linux &&

!((-2147483648 <= bigValue) && (bigValue <= 2147483647))) {

assignTrack = new Track(assignSymbol);

AddAssemblyCode(AssemblyOperator.mov, assignTrack,

assignSymbol.Value);

AddAssemblyCode(AssemblyOperator.mov, Base(resultSymbol),

Offset(resultSymbol), assignTrack);

}

else {

AddAssemblyCode(AssemblyOperator.mov, Base(resultSymbol),

Offset(resultSymbol), assignSymbol.Value,

typeSize);

}

}

else if (assignSymbol.Type.IsArrayFunctionOrString() ||

(assignSymbol.Value is StaticAddress)) {

if (assignSymbol.AddressSymbol != null) {

Track addressTrack =

LoadValueToRegister(assignSymbol.AddressSymbol);

addressTrack.Pointer = true;

AddAssemblyCode(AssemblyOperator.mov, Base(resultSymbol),

Offset(resultSymbol), addressTrack);

if (assignSymbol.AddressOffset != 0) {

AddAssemblyCode(AssemblyOperator.add, Base(resultSymbol),

Offset(resultSymbol), (BigInteger)

assignSymbol.AddressOffset, typeSize);

}

}

else {

AddAssemblyCode(AssemblyOperator.mov, Base(resultSymbol),

Offset(resultSymbol), Base(assignSymbol),

TypeSize.PointerSize);

int assignOffset = Offset(assignSymbol);

if (assignOffset != 0) {

AddAssemblyCode(AssemblyOperator.add, Base(resultSymbol),

Offset(resultSymbol), (BigInteger) assignOffset,

typeSize);

}

}

}

else {

assignTrack = LoadValueToRegister(assignSymbol);

AddAssemblyCode(AssemblyOperator.mov, Base(resultSymbol),

Offset(resultSymbol), assignTrack);

}

}

}

If the result symbol is not temporary, we move the value directly into the address of the result symbol.

In case of an integer value, there is a special case due to technical limitations if the size is eight bytes. In that case we have to load the value into a register and load the register into the address, since we cannot load an eight-byte value directly into an address.

If the value size is not eight-byte, we load the value directly into the address.

If the symbol is an array, function, string, or static address, we load the address rather than the value onto the address.

Otherwise, we need to load the value into a register, which we than load onto the address.

### Unary Integral Operations

The Unary method is called to add code for unary subtraction and bitwise not. There is also unary addition, but it generates no code.

First, we need a map from middle code operators to assembly code operators for integral types. We will use the map in the IntegralUnary and IntegralBinary methods.

public static IDictionary<MiddleOperator,AssemblyOperator>

m\_middleToIntegralMap =

new Dictionary<MiddleOperator,AssemblyOperator>() {

{MiddleOperator.BitwiseNot, AssemblyOperator.not},

{MiddleOperator.UnarySubtract, AssemblyOperator.neg},

{MiddleOperator.SignedMultiply, AssemblyOperator.imul},

{MiddleOperator.SignedDivide, AssemblyOperator.idiv},

{MiddleOperator.SignedModulo, AssemblyOperator.idiv},

{MiddleOperator.UnsignedMultiply, AssemblyOperator.mul},

{MiddleOperator.UnsignedDivide, AssemblyOperator.div},

{MiddleOperator.UnsignedModulo, AssemblyOperator.div},

{MiddleOperator.Assign, AssemblyOperator.mov},

{MiddleOperator.BinaryAdd, AssemblyOperator.add},

{MiddleOperator.BinarySubtract, AssemblyOperator.sub},

{MiddleOperator.BitwiseAnd, AssemblyOperator.and},

{MiddleOperator.BitwiseOr, AssemblyOperator.or},

{MiddleOperator.BitwiseXOr, AssemblyOperator.xor},

{MiddleOperator.ShiftLeft, AssemblyOperator.shl},

{MiddleOperator.ShiftRight, AssemblyOperator.shr},

{MiddleOperator.Equal, AssemblyOperator.je},

{MiddleOperator.NotEqual, AssemblyOperator.jne},

{MiddleOperator.Carry, AssemblyOperator.jc},

{MiddleOperator.NotCarry, AssemblyOperator.jnc},

{MiddleOperator.Compare, AssemblyOperator.cmp},

{MiddleOperator.SignedLessThan, AssemblyOperator.jl},

{MiddleOperator.SignedLessThanEqual,AssemblyOperator.jle},

{MiddleOperator.SignedGreaterThan, AssemblyOperator.jg},

{MiddleOperator.SignedGreaterThanEqual, AssemblyOperator.jge},

{MiddleOperator.UnsignedLessThan, AssemblyOperator.jb},

{MiddleOperator.UnsignedLessThanEqual, AssemblyOperator.jbe},

{MiddleOperator.UnsignedGreaterThan, AssemblyOperator.ja},

{MiddleOperator.UnsignedGreaterThanEqual, AssemblyOperator.jae}};

public void IntegralUnary(MiddleCode middleCode) {

Symbol resultSymbol = (Symbol) middleCode[0],

unarySymbol = (Symbol) middleCode[1];

IntegralUnary(middleCode.Operator, resultSymbol, unarySymbol);

}

public void IntegralUnary(MiddleOperator middleOperator,

Symbol resultSymbol, Symbol unarySymbol) {

AssemblyOperator objectOperator =

m\_middleToIntegralMap[middleOperator];

int typeSize = unarySymbol.Type.SizeArray();

Track unaryTrack = null;

if (unarySymbol.Value is BigInteger) {

SymbolTable.StaticSet.Add(ConstantExpression.Value(unarySymbol));

AddAssemblyCode(objectOperator, unarySymbol.UniqueName,

0, null, typeSize);

}

else if (m\_trackMap.TryGetValue(unarySymbol, out unaryTrack)) {

if (middleOperator != MiddleOperator.UnaryAdd) {

AddAssemblyCode(objectOperator, unaryTrack);

}

m\_trackMap.Remove(unarySymbol);

}

We have to cases. If the result symbol equals the unary symbol, we do not need to use a register. We can just perform the operation directly on the address of the symbol. Unless the operator is unary addition, in which case we do nothing.

else if (resultSymbol == unarySymbol) {

Assert.ErrorXXX(unaryTrack == null);

if (middleOperator != MiddleOperator.UnaryAdd) {

AddAssemblyCode(objectOperator, Base(unarySymbol),

Offset(unarySymbol), null, typeSize);

}

}

If the result symbol does not equal the unary symbol, we cannot perform the operation directly. We need to store the value in a register, perform the operation on the register, and store its value in the result symbol. Again, unless the operator is unary addition, in which case we do not perform the operations. However, the value is stored in the same way as the other operators.

else {

unaryTrack = LoadValueToRegister(unarySymbol);

if (middleOperator != MiddleOperator.UnaryAdd) {

AddAssemblyCode(objectOperator, unaryTrack);

}

If the result symbol is a temporary variable, we add the unary track to the result symbol in the track map.

if (resultSymbol.IsTemporary()) {

Assert.ErrorXXX(resultSymbol.AddressSymbol == null);

m\_trackMap.Add(resultSymbol, unaryTrack);

}

If the result symbol is a regular variable, we store the resulting value of the register on its address.

else {

AddAssemblyCode(AssemblyOperator.mov, Base(resultSymbol),

Offset(resultSymbol), unaryTrack);

}

m\_trackMap.Remove(unarySymbol);

}

}

### Integral Binary

The binary integral addition and subtraction as well as the bitwise and, or, and xor operations are rather straightforward. They take two values stored in registers, on memory addresses, or as integer values. However, the left and right shift operation demands that the right operands is stored in a specific register.

The IntegralBinary method calls the second IntegralBinary method with the operator and the operands.

public void IntegralBinary(MiddleCode middleCode) {

Symbol resultSymbol = (Symbol)middleCode[0],

leftSymbol = (Symbol)middleCode[1],

rightSymbol = (Symbol)middleCode[2];

IntegralBinary(middleCode.Operator, resultSymbol,

leftSymbol, rightSymbol);

}

The IntegralRelation method also calls the IntegralBinary with comparation operator and the operands. Note that the result symbol is null in this case. We do not expect a result value of the comparations. Instead, the result is placed in flags that the following jump instruction catches. The assembly code operator corresponding to the middle code operator is looked up in the m\_middleToIntegralMap map.

public void IntegralRelation(MiddleCode middleCode) {

Symbol leftSymbol = (Symbol) middleCode[1],

rightSymbol = (Symbol) middleCode[2];

IntegralBinary(MiddleOperator.Compare, null, leftSymbol, rightSymbol);

AssemblyOperator objectOperator =

m\_middleToIntegralMap[middleCode.Operator];

int target = (int) middleCode[0];

AddAssemblyCode(objectOperator, null, null, target);

}

The IntegralBinary method generates assembly code for binary integral, addition, bitwise, and shift expressions. Basically, we have two cases: we to load the left operand into a register, or we perform the operation on the address of the left operand. In both cases, the right operand may be stored in a register or on an address, it may also be an integer value or and address.

public void IntegralBinary(MiddleOperator middleOperator,

Symbol resultSymbol, Symbol leftSymbol,

Symbol rightSymbol) {

Track leftTrack = null, rightTrack = null;

m\_trackMap.TryGetValue(leftSymbol, out leftTrack);

m\_trackMap.TryGetValue(rightSymbol, out rightTrack);

Generally speaking, we try to keep the values in their current form as long as possible in order to generate as effective code as possible. However, there are some cases where we have to load the value into a register. More specifically, if the result symbol is not null (which it is in comparations) and it does not equal the left symbol (which it does in compound assignment), we load its value into a register.

if ((leftTrack == null) &&

(resultSymbol != null) && (resultSymbol != leftSymbol)) {

leftTrack = LoadValueToRegister(leftSymbol);

}

Another condition is that the left symbol may be an array, a function, or a string, or a static address. If their offset is non-zero, we must load its value into a register.

if ((leftTrack == null) &&

((leftSymbol.Type.IsArrayFunctionOrString() &&

(leftSymbol.Offset != 0)) ||

((leftSymbol.Value is StaticAddress) &&

(((StaticAddress) leftSymbol.Value).Offset != 0)))) {

leftTrack = LoadValueToRegister(leftSymbol);

}

If the same conditions apply on the right symbol, with the exception of the assignment, addition, or subtraction operator. For those operators we can handle the value with or without an offset.

if ((rightTrack == null) &&

(middleOperator != MiddleOperator.Assign) &&

(middleOperator != MiddleOperator.BinaryAdd) &&

(middleOperator != MiddleOperator.BinarySubtract) &&

((rightSymbol.Type.IsArrayFunctionOrString() &&

(rightSymbol.Offset != 0)) ||

((rightSymbol.Value is StaticAddress) &&

(((StaticAddress) rightSymbol.Value).Offset != 0)))) {

rightTrack = LoadValueToRegister(rightSymbol);

}

Another condition we need to check is whether the right symbol holds a large value. Due to technical limitation of the architecture, we must load the value into a register if it exceeds 31 bits. In that case, the value will be loaded to the register with the mov instruction, which is the only allowed instruction for values exceeding 31 bits. Except for the assignment operator, in which case we will use the mov instruction anyway. We do not need to check the left symbol in the same way, since the middle code optimizer has reduced the expression or swapped the operands if the left operand holds an integer value.

if ((rightTrack == null) && (rightSymbol.Value is BigInteger) &&

((middleOperator != MiddleOperator.Assign) ||

(leftTrack == null))) {

BigInteger bigValue = (BigInteger) rightSymbol.Value;

if (Start.Linux &&

!((-2147483648 <= bigValue) && (bigValue <= 2147483647))) {

rightTrack = LoadValueToRegister(rightSymbol);

}

}

In case of the shift operator, we need to load the right symbol into a specific register (which is always cl) due to technical limitations of the architecture.

if (MiddleCode.IsShift(middleOperator)) {

rightTrack =

LoadValueToRegister(rightSymbol, AssemblyCode.ShiftRegister);

}

int typeSize = leftSymbol.Type.Size();

AssemblyOperator objectOperator = m\_middleToIntegralMap[middleOperator];

Assert.ErrorXXX(!(leftSymbol.Value is BigInteger));

In this section, we call the Base and Offset methods. The base is a name in case of an extern or static symbol or a static address, and a register in case of an auto or register symbol. The offset is zero for an extern or static symbol. It may be zero or non-zero for a static address. It is always non-zero for an auto or register symbol.

If the value of left symbol is stored in a register (leftTrack is not null), we have a sequence of cases to consider. If the value of right symbol is also stored in a register, we simply perform the operation on the registers.

if (leftTrack != null) {

if (rightTrack != null) {

AddAssemblyCode(objectOperator, leftTrack, rightTrack);

}

If the right symbol is an integer value, we use that value directly. Note that the integer value does not exceed 31 bits. If it does, the value would have already been loaded int a register.

else if (rightSymbol.Value is BigInteger) {

AddAssemblyCode(objectOperator, leftTrack, rightSymbol.Value);

}

If the right symbol is an array, function, string, or static value, we use its address rather than its value. We start by applying the operator, the call to Base gives the name of the symbol.

else if (rightSymbol.Type.IsArrayFunctionOrString() ||

(rightSymbol.Value is StaticAddress)) {

AddAssemblyCode(objectOperator, leftTrack, Base(rightSymbol));

Then we need to look into the offset. Remember that there can only be addition, subtraction, or assignment if the offset is non-zero. Otherwise, the address would have been loaded into a register. In case of assignment, we add the offset to the address. Otherwise, we just apply the operator (add or sub).

int rightOffset = Offset(rightSymbol);

if (rightOffset != 0) {

if (middleOperator == MiddleOperator.Assign) {

AddAssemblyCode(AssemblyOperator.add, leftTrack,

(BigInteger) rightOffset);

}

else {

AddAssemblyCode(objectOperator, leftTrack,

(BigInteger) rightOffset);

}

}

}

If none of the above cases apply, we use the base and offset of the left symbol.

else {

if (middleOperator == MiddleOperator.Assign) {

leftTrack = new Track(resultSymbol);

}

AddAssemblyCode(objectOperator, leftTrack,

Base(rightSymbol), Offset(rightSymbol));

}

The only remaining issue is what to do with the resulting value, if there is one. If the result is a temporary variable, we add it to the track map to be used by succeeding instructions.

if (resultSymbol != null) {

if (resultSymbol.IsTemporary()) {

Assert.ErrorXXX(resultSymbol.AddressSymbol == null);

m\_trackMap.Add(resultSymbol, leftTrack);

}

If the result is not a temporary variable, we store the value on its address.

else {

AddAssemblyCode(AssemblyOperator.mov, Base(resultSymbol),

Offset(resultSymbol), leftTrack);

}

}

}

If the left symbol is not stored in a register, we have a new sequence of cases to consider. If the right symbol is stored in a register, we simply use it.

else {

if (rightTrack != null) {

AddAssemblyCode(objectOperator, Base(leftSymbol),

Offset(leftSymbol), rightTrack);

}

If the right symbol is an integer value, we simply use it. Again, note that the value does not exceeds 31 bits. If it does, it would have been loaded into a register.

else if (rightSymbol.Value is BigInteger) {

AddAssemblyCode(objectOperator, Base(leftSymbol),

Offset(leftSymbol), rightSymbol.Value, typeSize);

}

If the right symbol is an array, function, string, or static value, we use its address and apply the operators in the same way as above.

else if (rightSymbol.Type.IsArrayFunctionOrString() ||

(rightSymbol.Value is StaticAddress)) {

AddAssemblyCode(objectOperator, Base(leftSymbol),

Offset(leftSymbol), Base(rightSymbol),

TypeSize.PointerSize);

int rightOffset = Offset(rightSymbol);

if (rightOffset != 0) {

if (middleOperator == MiddleOperator.Assign) {

AddAssemblyCode(AssemblyOperator.add, Base(leftSymbol),

Offset(leftSymbol), (BigInteger) rightOffset,

typeSize);

}

else {

AddAssemblyCode(objectOperator, Base(leftSymbol),

Offset(leftSymbol), (BigInteger) rightOffset,

typeSize);

}

}

}

If none of the above cases apply, we load the value of the right symbol into a register and use the base and offset of the left symbol.

else {

rightTrack = LoadValueToRegister(rightSymbol);

AddAssemblyCode(objectOperator, Base(leftSymbol),

Offset(leftSymbol), rightTrack);

}

}

Finally, we need to remove the left and right symbol from the track map, since they shall not be used any more.

m\_trackMap.Remove(leftSymbol);

m\_trackMap.Remove(rightSymbol);

}

### Base and Offset

The Base method returns the base of a symbol, which may be a register, a track, or a name. If the symbol’s value if a static address we return its name, which will later be replaced by a proper address by the linker.

private object Base(Symbol symbol) {

Assert.ErrorXXX(!(symbol.Value is BigInteger));

if (symbol.Value is StaticAddress) {

StaticAddress staticAddress = (StaticAddress) symbol.Value;

return staticAddress.UniqueName;

}

If the symbol’s address symbol is not null, the symbol represents a dereferred symbol in a dereference, index, or arrow expression, and we return the register holding the address. We look up the address track (the track of the address symbol) by calling LoadValueToRegister and mark the track as pointer, which means that the register allocator will choose the register from a smaller set. Since the track is no longer needed, we delete it from the track map before we return it.

else if (symbol.AddressSymbol != null) {

Track addressTrack = LoadValueToRegister(symbol.AddressSymbol);

Assert.ErrorXXX((addressTrack.Register == null) ||

RegisterAllocator.VariadicFunctionPointerRegisterSet.

Contains(addressTrack.Register.Value));

addressTrack.Pointer = true;

m\_trackMap.Remove(symbol.AddressSymbol);

return addressTrack;

}

If the symbol is extern or static, we return its unique name, which will later be replaced by a proper address by the linker.

else if (symbol.IsExternOrStatic()) {

return symbol.UniqueName;

}

Finally, if the symbol is auto or register, we call BaseRegister to obtain the regular or variadic frame pointer.

else { //resultSymbol.IsAutoOrRegister()

return BaseRegister(symbol);

}

}

The Offset method returns the offset of the symbol. If its value is a static address, we return the offset of the static address.

private int Offset(Symbol symbol) {

Assert.ErrorXXX(!(symbol.Value is BigInteger));

if (symbol.Value is StaticAddress) {

StaticAddress staticAddress = (StaticAddress) symbol.Value;

return staticAddress.Offset;

}

If the symbol’s address symbol is not null, we return the offset of the address symbol.

else if (symbol.AddressSymbol != null) {

return symbol.AddressOffset;

}

Finally, in all other case we return the offset of the symbol. The offset is significant in case of auto or register parameters and variables. In case of an extern or static symbol, the offset is always zero and we will not use the value.

else {

return symbol.Offset;

}

}

### Case

The Case is method checks whether the switch symbol equals a value. Normally, we load the value of a symbol into a register, which we add to the track map. After the operation has been performed, we remove the register from the track map. However, when dealing with a sequence of case statements, we load the value of the switch symbol into a register, and then we keep the value in the register during all the case test instructions.

public void Case(MiddleCode middleCode) {

Symbol switchSymbol = (Symbol) middleCode[1];

Track switchTrack = LoadValueToRegister(switchSymbol);

Symbol caseSymbol = (Symbol) middleCode[2];

BigInteger caseValue = (BigInteger) caseSymbol.Value; // cmp ax, 123

AddAssemblyCode(AssemblyOperator.cmp, switchTrack, caseValue);

int target = (int) middleCode[0];

AddAssemblyCode(AssemblyOperator.je, null, null, target);

Note that we do add the symbol to the track map, in order to keep the value in the register thought out all the succeeding case instruction of the switch statement.

m\_trackMap.Add(switchSymbol, switchTrack);

}

The CaseEnd method is called after the last case instruction of the switch statement. We only remove the switch symbol from the track map, since we shall not keep the value in the register after the last case instruction.

public void CaseEnd(MiddleCode middleCode) {

Symbol symbol = (Symbol) middleCode[0];

m\_trackMap.Remove(symbol);

}

### Address

The Address method adds assembly code instructions for the address operator. It is actually quite simple, since we only gave to call LoadAddressToRegister to load the address of the symbol into a register.

public void Address(MiddleCode middleCode) {

Symbol resultSymbol = (Symbol) middleCode[0],

addressSymbol = (Symbol) middleCode[1];

Track track = LoadAddressToRegister(addressSymbol);

m\_trackMap.Add(resultSymbol, track);

m\_trackMap.Remove(addressSymbol);

}

The IntegralDereference method adds assembly code instructions for derefereeing a pointer to an integral value. We only have to call LoadValueToRegister for the address symbol.

public void IntegralDereference(MiddleCode middleCode) {

Symbol resultSymbol = (Symbol) middleCode[0];

Assert.ErrorXXX(resultSymbol.AddressSymbol != null);

Track addressTrack = LoadValueToRegister(resultSymbol.AddressSymbol);

m\_trackMap.Add(resultSymbol.AddressSymbol, addressTrack);

}

The FloatingDereference method adds assembly code instructions for derefereing a pointer to a floating value.

public void FloatingDereference(MiddleCode middleCode) {

Symbol resultSymbol = (Symbol) middleCode[0];

Assert.ErrorXXX(resultSymbol.AddressSymbol != null);

Track addressTrack = LoadValueToRegister(resultSymbol.AddressSymbol);

m\_trackMap.Add(resultSymbol.AddressSymbol, addressTrack);

}

### Floating Binary

Similar to the integral operations, we need the m\_middleToFloatingMap map to translate the middle code operators to their equivalent assembly code operators.

public static IDictionary<MiddleOperator, AssemblyOperator>

m\_middleToFloatingMap =

new Dictionary<MiddleOperator, AssemblyOperator>() {

{MiddleOperator.UnarySubtract, AssemblyOperator.fchs},

{MiddleOperator.BinaryAdd, AssemblyOperator.fadd},

{MiddleOperator.BinarySubtract, AssemblyOperator.fsub},

{MiddleOperator.SignedMultiply, AssemblyOperator.fmul},

{MiddleOperator.SignedDivide, AssemblyOperator.fdiv},

{MiddleOperator.Equal, AssemblyOperator.je},

{MiddleOperator.NotEqual, AssemblyOperator.jne},

{MiddleOperator.SignedLessThan, AssemblyOperator.ja},

{MiddleOperator.SignedLessThanEqual, AssemblyOperator.jae},

{MiddleOperator.SignedGreaterThan, AssemblyOperator.jb},

{MiddleOperator.SignedGreaterThanEqual, AssemblyOperator.jbe}

};

The methods for the floating-point operations are in fact much more simple than the corresponding integral methods. The FloatingUnary method just adds the instruction to the assembly code. The operand of the operator has already been pushed to the floating-point stack, the operator pops that value, and pushes the result to the stack.

public void FloatingUnary(MiddleCode middleCode) {

AddAssemblyCode(m\_middleToFloatingMap[middleCode.Operator]);

}

The FloatingBinary works in the same way, the operator pops the two operands from the floating-point stack and pushes the result on the stack.

public void FloatingBinary(MiddleCode middleCode) {

Assert.ErrorXXX((--m\_floatStackSize) >= 0);

AddAssemblyCode(m\_middleToFloatingMap[middleCode.Operator]);

}

In the FloatingParameter we do not need to actual parameter symbol, since its value has already been pushed on the floating-point stack. We only need to type and offset of the parameter when we pop it on the floating-point stack by calling TopPopSymbol.

public void FloatingParameter(MiddleCode middleCode) {

Type paramType = (Type) middleCode[1];

Symbol paramSymbol = new Symbol(paramType);

int paramOffset = (int) middleCode[0];

paramSymbol.Offset = m\_totalExtraSize + paramOffset;

TopPopSymbol(paramSymbol, TopOrPop.Pop);

}

### Floating Relation

The FloatingRelation method generate code for relation operators for floating types. The fcompp (float compare pop) assembly code instruction performs a comparation on the two values on top of the floating value stack, stores the result in the internal float status word, and pops the values from the stack. The fstsw (float store status word) assembly code instruction stores the floating status word in the ax register. The sahf (store ah into flags) stores the value of register ah (which hold the higher byte of register ax) into the integral flags. Finally, we add a jump instruction that matches the original middle code instruction and is looked up in the middle to floating map.

public void FloatingRelation(MiddleCode middleCode) {

Assert.ErrorXXX((m\_floatStackSize -= 2) >= 0);

int target = (int) middleCode[0];

AddAssemblyCode(AssemblyOperator.fcompp);

AddAssemblyCode(AssemblyOperator.fstsw, Register.ax);

AddAssemblyCode(AssemblyOperator.sahf);

AssemblyOperator objectOperator =

m\_middleToFloatingMap[middleCode.Operator];

AddAssemblyCode(objectOperator, null, null, target);

}

### Floating Push and Pop

When pushing floating values to the floating-point stack, we need the m\_floatPushMap map that maps the whether the type is logical and the type size to assembly code operators.

public static IDictionary<Pair<bool, int>, AssemblyOperator>

m\_floatPushMap = new Dictionary<Pair<bool, int>, AssemblyOperator>() {

{new Pair<bool,int>(false, 2), AssemblyOperator.fild\_word},

{new Pair<bool,int>(false, 4), AssemblyOperator.fild\_dword},

{new Pair<bool,int>(false, 8), AssemblyOperator.fild\_qword},

{new Pair<bool,int>(true, 4), AssemblyOperator.fld\_dword},

{new Pair<bool,int>(true, 8), AssemblyOperator.fld\_qword}

};

The PushSymbol method generates code that pushes the value of a symbol to the floating-point stack.

public void PushSymbol(Symbol symbol) {

Assert.ErrorXXX((++m\_floatStackSize) <= FloatingStackMaxSize);

Track track;

There are two special assembly code instructions for pushing the value zero or one to the floating-point stack: fldz (float load zero) fld1 (float load one). If the value is (integral or floating) zero, we use fldz to load the value zero, and if the value is (integral or floating) one we use fld1 to load the value zero.

if (((symbol.Value is BigInteger) &&

(((BigInteger) symbol.Value).IsZero)) ||

((symbol.Value is decimal) && (((decimal) symbol.Value) == 0))) {

AddAssemblyCode(AssemblyOperator.fldz);

}

else if (((symbol.Value is BigInteger) &&

(((BigInteger) symbol.Value).IsOne)) ||

((symbol.Value is decimal) &&

(((decimal) symbol.Value) == 1))) {

AddAssemblyCode(AssemblyOperator.fld1);

}

For any other integral or floating value, we cannot load it directly. Instead, we need store the value on a memory address and load the value from the address. We add the symbol to the global static set for the linker to be able to find the value, and we add the push operation with the name of symbol, which will be changed to a proper address by the linker.

else {

Pair<bool,int> pair =

new Pair<bool,int>(symbol.Type.IsFloating(), symbol.Type.Size());

AssemblyOperator objectOperator = m\_floatPushMap[pair];

if ((symbol.Value is BigInteger) || (symbol.Value is decimal)) {

SymbolTable.StaticSet.Add(ConstantExpression.Value(symbol));

AddAssemblyCode(objectOperator, symbol.UniqueName, 0);

}

If the symbol does not hold an integral or a floating value, we check if the value is stored in a register. In that case we cannot store the register directly, but like the value case above we need to store the value at the memory address. We use the special integral storage to load the value.

else if (m\_trackMap.TryGetValue(symbol, out track)) {

m\_trackMap.Remove(symbol);

string containerName = AddStaticContainer(symbol.Type);

AddAssemblyCode(AssemblyOperator.mov, containerName, 0, track);

AddAssemblyCode(objectOperator, containerName, 0);

}

If the symbol is an array, function, or string, we also use the integral storage name to store the value. However, in this case will shall load the address of the symbol, rather than its value. We start by loading the base address of the symbol to the integral storage. Then we add its offset, unless it is zero.

else if (symbol.Type.IsArrayFunctionOrString()) {

string containerName = AddStaticContainer(symbol.Type);

AddAssemblyCode(AssemblyOperator.mov, containerName, 0,

Base(symbol), TypeSize.PointerSize);

int offset = Offset(symbol);

if (offset != 0) {

AddAssemblyCode(AssemblyOperator.add, containerName, 0,

(BigInteger) offset, TypeSize.PointerSize);

}

AddAssemblyCode(objectOperator, containerName, 0);

}

If none of the above apply, we perform the operation on the address of the symbol.

else {

AddAssemblyCode(objectOperator, Base(symbol), Offset(symbol));

}

}

}

private static string AddStaticContainer(Type type) {

string containerName = "container" + type.Size() +

"bytes" + Symbol.NumberId;

SymbolTable.StaticSet.Add(ConstantExpression.Value(containerName,

type, null));

return containerName;

}

When popping or topping a value from the floating-point stack, we also need the m\_floatPopMap and m\_floatTopMap maps that map whether the type is logical and the type size to pop and top assembly code instructions.

public static IDictionary<Pair<bool,int>,AssemblyOperator>

m\_floatPopMap = new Dictionary<Pair<bool,int>, AssemblyOperator>() {

{new Pair<bool,int>(false, 2), AssemblyOperator.fistp\_word},

{new Pair<bool,int>(false, 4), AssemblyOperator.fistp\_dword},

{new Pair<bool,int>(false, 8), AssemblyOperator.fistp\_qword},

{new Pair<bool,int>(true, 4), AssemblyOperator.fstp\_dword},

{new Pair<bool,int>(true, 8), AssemblyOperator.fstp\_qword}

};

public static IDictionary<Pair<bool,int>,AssemblyOperator>

m\_floatTopMap = new Dictionary<Pair<bool,int>, AssemblyOperator>() {

{new Pair<bool,int>(false, 2), AssemblyOperator.fist\_word},

{new Pair<bool,int>(false, 4), AssemblyOperator.fist\_dword},

{new Pair<bool,int>(false, 8), AssemblyOperator.fist\_qword},

{new Pair<bool,int>(true, 4), AssemblyOperator.fst\_dword},

{new Pair<bool,int>(true, 8), AssemblyOperator.fst\_qword}

};

public enum TopOrPop {Top, Pop};

The PopEmpty method pops the floating-point stack without storing the value anywhere. However, there is no matching assembly code instruction. The instructions always store the value somewhere. Therefore, we use the integral storage, simply because we have to use some storage.

public void PopEmpty() {

string containerName = AddStaticContainer(Type.LongDoubleType);

AddAssemblyCode(AssemblyOperator.fistp\_word, containerName, 0);

}

The TopPopSymbol method tops or pops the top-most value of the floating-point stack.

public void TopPopSymbol(Symbol symbol, TopOrPop topOrPop) {

Assert.ErrorXXX(symbol != null);

Pair<bool,int> pair =

new Pair<bool,int>(symbol.Type.IsFloating(), symbol.Type.Size());

AssemblyOperator objectOperator;

Depending on the value of the topOrPop parameter, we extract the assembly code instruction from the m\_floatPopMap and m\_floatTopMap maps.

if (topOrPop == TopOrPop.Pop) {

objectOperator = m\_floatPopMap[pair];

Assert.ErrorXXX((--m\_floatStackSize) >= 0);

}

else {

objectOperator = m\_floatTopMap[pair];

}

If the symbol is a temporary variable, we store its value in a register. However, we cannot store the value directly in a register. Therefore, we pop or top the value to the integral storage and loads the value from there into the register.

if (symbol.Type.IsIntegralPointerOrArray() && symbol.IsTemporary()) {

string containerName = AddStaticContainer(symbol.Type);

AddAssemblyCode(objectOperator, containerName, 0);

Track track = new Track(symbol);

AddAssemblyCode(AssemblyOperator.mov, track, containerName, 0);

m\_trackMap.Add(symbol, track);

}

If none of the above cases apply, we pop or top the value to the address of the symbol.

else {

AddAssemblyCode(objectOperator, Base(symbol), Offset(symbol));

}

}

### Type Conversion

The IntegralToIntegral method adds assembly code instructions that converts an integral value from one size to another size.

public void IntegralToIntegral(MiddleCode middleCode, int index) {

Symbol targetSymbol = (Symbol) middleCode[0],

sourceSymbol = (Symbol) middleCode[1];

The sizeArray method returns pointer size for arrays, functions, and strings.

Type targetType = targetSymbol.Type, sourceType = sourceSymbol.Type;

int targetSize = targetType.SizeArray(),

sourceSize = sourceType.SizeArray();

We need target add a set-track-size instruction for

Track sourceTrack = LoadValueToRegister(sourceSymbol);

AddAssemblyCode(AssemblyOperatargetr.set\_track\_size,

sourceTrack, targetSize);

If the source size is smaller than the target size, we add code to mask the upper bits; that is, we set to zero the bits of the target value that do not fit in the source value.

if (sourceSize != targetSize) {

if (sourceSize < targetSize) {

If the target size is eight bytes, we have a special case. We need to load the mask to a register, and use the and instruction to do the actual masking.

if (targetSize == 8) {

Track targetTrack = new Track(targetSymbol);

AddAssemblyCode(AssemblyOperatargetr.mov, targetTrack,

TypeSize.GetMask(sourceType.Sort));

AddAssemblyCode(AssemblyOperatargetr.and,

sourceTrack, targetTrack);

}

If the target size is not eight bytes, we can use the and instruction directly.

else {

AddAssemblyCode(AssemblyOperatargetr.and, sourceTrack,

TypeSize.GetMask(sourceType.Sort));

}

}

If both the source type and target type is signed, we need to preserve the signed status throught the conversion process. If the value is negative, the top-most bit is set to one. Since the source size does not equals the target size, the top-most bits differ between the values. Therefore, if the value is negative, we need to change it to a positive value with the source size and change it back to a negative value with the target size. In this way, the signed status will be preserved. However, if the value is non-negative (zero or positive), we do nothing. We just add to the track map that the values holds a new size.

if (sourceType.IsSigned() && targetType.IsSigned()) {

AddAssemblyCode(AssemblyOperatargetr.set\_track\_size,

sourceTrack, sourceSize);

AddAssemblyCode(AssemblyOperatargetr.cmp, sourceTrack,

BigInteger.Zero);

AddAssemblyCode(AssemblyOperatargetr.jge, null, null, index + 1);

AddAssemblyCode(AssemblyOperatargetr.neg, sourceTrack);

AddAssemblyCode(AssemblyOperatargetr.set\_track\_size,

sourceTrack, targetSize);

AddAssemblyCode(AssemblyOperatargetr.neg, sourceTrack);

}

}

If the target type and source type have the same size, we only add the target symbol with the source track in the track map.

m\_trackMap.Add(targetSymbol, sourceTrack);

}

The IntegralToFloating and FloatingToIntegral are quite simple. In the integral-to-floating-case we push the integral value at the floating-point stack, and in the floating-to-integral case, we pop the value from the stack. Note that there is no floating-to-floating case. In that case, we just keep the value on the stack, the type conversion occurs when the value is pushed, popped, or topped.

public void IntegralToFloating(MiddleCode middleCode) {

Symbol fromSymbol = (Symbol) middleCode[1];

PushSymbol(fromSymbol);

}

public void FloatingToIntegral(MiddleCode middleCode) {

Symbol toSymbol = (Symbol) middleCode[0];

TopPopSymbol(toSymbol, TopOrPop.Pop);

}

### Struct and Union

This section describes the handling of struct and unions, assignment, parameter, and function return value. Note that these methods make no difference between struct and unions, in all cases the task is to move the data of a memory block between two addresses.

The StructUnionAssign method copies the data from the address of the target symbol to address of the source symbol by calling MemoryCopy.

public void StructUnionAssignInit(MiddleCode middleCode) {

Symbol targetSymbol = (Symbol)middleCode[0],

sourceSymbol = (Symbol)middleCode[1];

MemoryCopyInit(targetSymbol, sourceSymbol);

}

public void StructUnionAssign(MiddleCode middleCode, int middleIndex) {

MemoryCopyLoop(middleIndex);

}

The StructUnionParameter copies the data from the address of the parameter symbol to the address on the activation record by calling MemoryCopy.

public void StructUnionParameterInit(MiddleCode middleCode) {

int paramOffset = (int) middleCode[0];

Symbol sourceSymbol = (Symbol) middleCode[2];

Symbol targetSymbol = new Symbol(Type.IntegerPointerType);

targetSymbol.Offset = m\_totalExtraSize + paramOffset;

MemoryCopyInit(targetSymbol, sourceSymbol);

}

public void StructUnionParameter(MiddleCode middleCode, int middleIndex) {

MemoryCopyLoop(middleIndex);

}

The StructUnionGetReturnValue loads the address symbol of the symbol into the return pointer register, which is then added to the track map. In this way, is the value of the address symbol of the struct or union stored in the track and will be used in later operations.

public void StructUnionGetReturnValue(MiddleCode middleCode) {

Symbol targetSymbol = (Symbol) middleCode[0];

CheckRegister(targetSymbol, AssemblyCode.ReturnPointerRegister);

Track targetAddressTrack =

new Track(targetSymbol.AddressSymbol,

AssemblyCode.ReturnPointerRegister);

m\_trackMap.Add(targetSymbol.AddressSymbol, targetAddressTrack);

}

The StructUnionSetReturnValue method

public void StructUnionSetReturnValue(MiddleCode middleCode) {

Symbol returnSymbol = (Symbol) middleCode[1];

LoadAddressToRegister(returnSymbol, AssemblyCode.ReturnPointerRegister);

}

The m\_labelCount field is used by MemoryCopy to generate a new label for each copy process.

private Track m\_targetAddressTrack = null, m\_sourceAddressTrack = null;

private static Track m\_countTrack = null;

The MemoryCopy method adds code to copy the data of a memory block between two memory addresses.

private void MemoryCopyInit(Symbol targetSymbol, Symbol sourceSymbol) {

m\_targetAddressTrack = LoadAddressToRegister(targetSymbol);

m\_sourceAddressTrack = LoadAddressToRegister(sourceSymbol);

We create tracks for the register to count of the data block and the value being copied. If the size of the data block is less than 256, we use the character types (one byte) for the count register, otherwise we use the integer type (four bytes for Linux, two bytes for Windows).

int size = sourceSymbol.Type.Size();

Type countType = (size < 256) ? Type.UnsignedCharType

: Type.UnsignedIntegerType;

We start by loading the size of the data block into the count register, the loop continues until it reaches zero.

m\_countTrack = new Track(countType);

AddAssemblyCode(AssemblyOperator.mov, m\_countTrack, (BigInteger) size);

}

Then the loop begins, we use a label that is unique in the assembly file with the m\_labelCount field.

private void MemoryCopyLoop(int middleIndex) {

Track valueTrack = new Track(Type.UnsignedCharType);

We move a value from the source address to the target address.

AddAssemblyCode(AssemblyOperator.mov, valueTrack,

m\_sourceAddressTrack, 0);

AddAssemblyCode(AssemblyOperator.mov, m\_targetAddressTrack,

0, valueTrack);

We increment the source and target address.

AddAssemblyCode(AssemblyOperator.inc, m\_sourceAddressTrack);

AddAssemblyCode(AssemblyOperator.inc, m\_targetAddressTrack);

We decrement the count register.

AddAssemblyCode(AssemblyOperator.dec, m\_countTrack);

If the count register does not equal zero, we jump to the label. If it does equal zero, the copy process is done.

AddAssemblyCode(AssemblyOperator.cmp, m\_countTrack, BigInteger.Zero);

AddAssemblyCode(AssemblyOperator.jne, null, null, middleIndex);

}

### Initialization Code

The initialization of data blocks differs between the Linux and the Windows environment. In this section we look into the Linux environment, the Windows environment is described in Chapter 13.

AssemblyCodeGenerator.cs

public static void InitializationCodeList() {

List<AssemblyCode> assemblyCodeList = new List<AssemblyCode>();

AddAssemblyCode(assemblyCodeList, AssemblyOperator.label, "\_start");

AddAssemblyCode(assemblyCodeList, AssemblyOperator.comment,

"Initializerialize Stack Pointer");

We begin by initialize the frame pointer. It set to the address of the stack, which is the address directly after the code and static data. We do not need to set the variadic pointer, since the main function is not variadic.

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

AssemblyCode.RegularFrameRegister, Linker.CallStackStart);

We initialize the heap to 64 kilobytes (65536 bytes) and set the heap pointer at address 65534 to zero, since the heap is empty at the beginning of the execution.

if (Start.Linux) {

AddAssemblyCode(assemblyCodeList, AssemblyOperator.comment,

"Initializerialize Heap Pointer");

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov\_dword,

Linker.CallStackStart, 65534,

Linker.CallStackStart);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.add\_dword,

Linker.CallStackStart, 65534,

(BigInteger) 65534);

We also need to initialize the floating-point unit control word so that floating point operations truncates.

AddAssemblyCode(assemblyCodeList, AssemblyOperator.comment,

"Initializerialize FPU Control Word, truncate mode " +

"=> set bit 10 and 11.");

AddAssemblyCode(assemblyCodeList, AssemblyOperator.fstcw,

AssemblyCode.RegularFrameRegister, 0);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.or\_word,

AssemblyCode.RegularFrameRegister, 0, (BigInteger) 0x0C00);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.fldcw,

AssemblyCode.RegularFrameRegister, 0);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Linker.CallStackStart, 0, BigInteger.Zero,

TypeSize.PointerSize);

List<string> textList = new List<string>();

textList.Add("section .text");

ISet<string> externSet = new HashSet<string>();

AssemblyCodeGenerator.LinuxTextList(assemblyCodeList, textList,

externSet);

SymbolTable.InitSymbol =

new StaticSymbolLinux(AssemblyCodeGenerator.InitializerName,

textList, externSet);

}

if (Start.Windows) {

// ...

}

}

### Command Line Arguments

Assuming that the main function is declared as void main(int argc, char\* argv[]); and that we start the execution with the command line main Hello World the following structure becomes generated:

main

’m’

’a’

’i’

’n’

’\0’

’H’

’e’

’l’

’l’

’o’

’\0’

’W’

’o’

’r’

’l’

’d’

’\0’

Return address: 0

Frame pointer of

calling function

Ellipse pointer of

calling function

argc: 3

argv

null

The initialization code looks as follows. We begin by popping the system stack to rbx to obtain the number of arguments, which is three in the example above. We also copy the value to rax. We will use rbx to count down the number or arguments, and rax to store the number on the activation record. Moreover, we use rbp as the frame pointer and copy its value in rdx. We will put the pointer to the argument directly above the code and data part of the code, and directly below the activation record for the first main call.

pop rbx

mov rax, rbx

mov rdx, rbp

We count down the number of arguments in rbx. When it has reached zero, we quit the loop. Each pointer to e new argument is popped from the system stack, added to memory with rbp. The rbp is increased by eight bytes, and rbx is decreased by one.

$args$loop:

cmp rbx, 0

je $args$exit

pop rsi

mov [rbp], rsi

add rbp, 8

dec rbx

jmp $args$loop

When all argument pointers have been added, we add the final null pointer and increase rbp. By then the argument list has been properly added to the memory.

$args$exit:

mov qword [rbp], 0

add rbp, 8

The rbp register has by now been given its correct value, it now points at the beginning of the activation record. We set the return address (at offset zero) to zero, indicating the returning from the function shall result in an exit from the execution. We also set the argc parameter (at offset 24) to the value eax, holding the number of arguments, and the argv parameter (at offset 28) the value of rdx, holding the value of the first argument pointer.

mov qword [rbp], 0

mov [rbp + 24], eax

mov [rbp + 28], rdx

The code for obtaining the command line arguments differs between the Linux and the Windows environment.

public static void ArgumentCodeList() {

List<AssemblyCode> assemblyCodeList = new List<AssemblyCode>();

When to execution starts, the system stack is loaded with values corresponding to the command line arguments. We pop the stack to the rbx register to obtain the number of arguments.

if (Start.Linux) {

/\* pop rbx

mov rax, rbx

mov rdx, rbp \*/

AddAssemblyCode(assemblyCodeList, AssemblyOperator.comment,

"Initialize Command Line Arguments");

AddAssemblyCode(assemblyCodeList, AssemblyOperator.pop, Register.rbx);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Register.rax, Register.rbx);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Register.rdx, Register.rbp);

/\* $args$loop:

cmp rbx, 0

je $args$exit

pop rsi

mov [rbp], rsi

add rbp, 8

dec rbx

jmp $args$loop \*/

AddAssemblyCode(assemblyCodeList, AssemblyOperator.label,

"$args$loop");

AddAssemblyCode(assemblyCodeList, AssemblyOperator.cmp,

Register.rbx, BigInteger.Zero);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.je,

null, null, "$args$exit");

AddAssemblyCode(assemblyCodeList, AssemblyOperator.pop, Register.rsi);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Register.rbp, 0, Register.rsi);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.add, Register.rbp,

(BigInteger) TypeSize.PointerSize);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.dec, Register.rbx);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.jmp,

null, null, "$args$loop");

AddAssemblyCode(assemblyCodeList, AssemblyOperator.label,

"$args$exit");

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov\_qword,

Register.rbp, 0, BigInteger.Zero);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.add, Register.rbp,

(BigInteger) TypeSize.PointerSize);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Register.rbp,0, BigInteger.Zero,TypeSize.PointerSize);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov, Register.rbp,

SymbolTable.FunctionHeaderSize, Register.eax);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov, Register.rbp,

SymbolTable.FunctionHeaderSize +

TypeSize.SignedIntegerSize, Register.rdx);

/\* $args$exit:

mov qword [rbp], 0

add rbp, 8 \*/

List<string> textList = new List<string>();

ISet<string> externSet = new HashSet<string>();

AssemblyCodeGenerator.LinuxTextList(assemblyCodeList, textList,

externSet);

SymbolTable.ArgsSymbol =

new StaticSymbolLinux(AssemblyCodeGenerator.ArgsName,

textList, externSet);

}

if (Start.Windows) {

// ...

}

}

### Text List

When the assembly code has been generated, we need to be written as a text file, which is the task of LinuxTextList. It iterates through the assembly code list and writes the assembly code instruction at one line each by calling ToString in AssemblyCode. It also adds names add the externSet parameter for each name in the instructions.

public static void LinuxTextList(IList<AssemblyCode> assemblyCodeList,

IList<string> textList,

ISet<string> externSet) {

foreach (AssemblyCode assemblyCode in assemblyCodeList) {

AssemblyOperator assemblyOperator = assemblyCode.Operator;

object operand0 = assemblyCode[0],

operand1 = assemblyCode[1],

operand2 = assemblyCode[2];

if (assemblyOperator == AssemblyOperator.define\_value) {

Sort sort = (Sort) operand0;

if ((sort != Sort.String) && (operand1 is string)) {

string name1 = (string) operand1;

if (!name1.Contains(Symbol.SeparatorId)) {

externSet.Add(name1);

}

}

}

else if ((assemblyOperator != AssemblyOperator.label) &&

(assemblyOperator != AssemblyOperator.comment)) {

if (operand0 is string) {

string name0 = (string) operand0;

if (!name0.Contains(Symbol.SeparatorId)) {

externSet.Add(name0);

}

}

if (operand1 is string) {

string name1 = (string) operand1;

if (!name1.Contains(Symbol.SeparatorId)) {

externSet.Add(name1);

}

}

if (operand2 is string) {

string name2 = (string) operand2;

if (!name2.Contains(Symbol.SeparatorId)) {

externSet.Add(name2);

}

}

}

string text = assemblyCode.ToString();

if (text != null) {

textList.Add(text);

}

}

}

# Executable Code Generation

So far in this book, the compiler has generated assembly code files and a make file. In this chapter, the compiler instead generates an executable file. I have chosen the 16-bit .com format, which is a simple file format. Unfortunately, the format is no longer supported by Windows. Therefore, we need to use a simulator the supports the format. I have chosen DosBox, which is a simple simulator capable of executing files in the .com format.

## The Windows environment

In the book so far, we have regarded the Start.Linux condition on several occasions, indicating code specific for the Linux target machine. In this chapter we go the outer way around and regard the Start.Windows condition, which indicates code specific for the Windows target machine.

The difference between the Windows environment, compared to the Linux environment, is:

* The types hold different sizes.
* A static value is stored as a byte list rather than a text list, with an access map that keeps track of the accesses of other static values.
* A function is stored as a byte list, with an access map. A function also holds a call map to keep track of functions calls and a return set to keep track of the return addresses of the calls.
* To exit the execution, we perform an interrupt call rather than a system call.
* The initialization code is different.
* The command line arguments code is also different. The command line is stored in the first 256 bytes of the segment.
* The total code, including code, static values, call stack and heap, is stored in 64 kilobytes.
* In the standard library, we use interrupts instead of system calls to access the surrounding operating system.

The accesses and calls are stored with their unique names, and replaced by proper address by the linker in the next step. The return addresses are stored as the address relative the beginning of the function and is replaced by the corresponding global address by the linker. The reason we keep the accesses and calls in separate maps is that accesses are replaced by absolute global addresses while calls are replaced by addresses relative to the call.

### Main

Main.cs

using System;

using System.IO;

using System.Text;

using System.Globalization;

using System.Collections.Generic;

namespace CCompiler {

public class Start {

// ...

public static void Main(string[] args){

// ...

if (Start.Windows) {

ObjectCodeTable.Initializer();

}

// ...

try {

// ...

if (Start.Linux) {

// ...

}

We iterate through the command line arguments in the same way as in Chapter 12. However, we add the doLink variable that is set to true if at least one source file becomes recompiled and we thereby shall link the files into an executable file. We also state the file suffix as “.obj” rather than “.asm”.

if (Start.Windows) {

bool doLink = false;

foreach (string arg in argList) {

FileInfo file = new FileInfo(SourcePath + arg);

if (rebuild || !IsGeneratedFileFresh(file, ".obj")) {

if (print) {

Console.Out.WriteLine("Compiling \"" +

file.FullName + ".c\".");

}

CompileSourceFile(file);

doLink = true;

}

}

if (doLink) {

FileInfo targetFile =

new FileInfo(TargetPath + argList[0] + ".com");

Linker linker = new Linker();

CCompiler\_Main.Scanner.Path = null;

foreach (string arg in argList) {

FileInfo file = new FileInfo(SourcePath + arg);

if (print) {

Console.Out.WriteLine("Loading \"" + file.FullName +

".obj\".");

}

ReadObjectFile(file, linker);

}

linker.Generate(targetFile);

}

else if (print) {

Console.Out.WriteLine(SourcePath + argList[0] +

".com is up-to-date.");

}

}

// ...

}

}

// ...

public static void CompileSourceFile(FileInfo file) {

// ...

if (Start.Linux) {

// ...

}

// ...

if (Start.Windows) {

FileInfo objectFile = new FileInfo(file.FullName + ".obj");

BinaryWriter binaryWriter =

new BinaryWriter(File.Open(objectFile.FullName, FileMode.Create));

binaryWriter.Write(SymbolTable.StaticSet.Count);

foreach (StaticSymbol staticSymbol in SymbolTable.GlobalStaticSet) {

staticSymbol.Write(binaryWriter);

}

binaryWriter.Close();

}

}

}

}

The ReadObjectFile method is called for reading an object file when it is not necessary to recompile the source file. An object file is made up by static object, and we read the object and add them to the linker.

public static void ReadObjectFile(FileInfo file, Linker linker) {

FileInfo objectFile = new FileInfo(file.FullName + ".obj");

try {

BinaryReader dataInputStream =

new BinaryReader(File.OpenRead(objectFile.FullName));

The first value of the file is the number of static objects. For each object we call the Load method of the StaticSymbolWindows class and the Add method of the Linker class.

int linkerSetSize = dataInputStream.ReadInt32();

for (int count = 0; count < linkerSetSize; ++count) {

StaticSymbolWindows staticSymbol = new StaticSymbolWindows();

staticSymbol.Read(dataInputStream);

linker.Add(staticSymbol);

}

dataInputStream.Close();

}

catch (Exception exception) {

Console.Out.WriteLine(exception.StackTrace);

Assert.Error(exception.Message);

}

}

### Type Size

To begin with, we need new type sizes for the Windows environment. A character is, as always, one byte. A short is one byte, an integer or a pointer is two bytes, and a long is four bytes. Note that an integer and a pointer have equal sizes in the Windows environment, as opposed to the Linux environment.

TypeSize.cs

static TypeSize() {

// ...

if (Start.Linux) {

// ...

}

if (Start.Windows) {

PointerSize = 2;

SignedIntegerSize = 2;

m\_sizeMap.Add(Sort.Void, 0);

m\_sizeMap.Add(Sort.Function, 0);

m\_sizeMap.Add(Sort.Logical, 1);

m\_sizeMap.Add(Sort.Array, 2);

m\_sizeMap.Add(Sort.Pointer, 2);

m\_sizeMap.Add(Sort.String, 2);

m\_sizeMap.Add(Sort.SignedChar, 1);

m\_sizeMap.Add(Sort.UnsignedChar, 1);

m\_sizeMap.Add(Sort.SignedShortInt, 1);

m\_sizeMap.Add(Sort.UnsignedShortInt, 1);

m\_sizeMap.Add(Sort.SignedInt, 2);

m\_sizeMap.Add(Sort.UnsignedInt, 2);

m\_sizeMap.Add(Sort.SignedLongInt, 4);

m\_sizeMap.Add(Sort.UnsignedLongInt, 4);

m\_sizeMap.Add(Sort.Float, 4);

m\_sizeMap.Add(Sort.Double, 8);

m\_sizeMap.Add(Sort.LongDouble, 8);

m\_signedMap.Add(1, Sort.SignedChar);

m\_signedMap.Add(2, Sort.SignedInt);

m\_signedMap.Add(4, Sort.SignedLongInt);

m\_unsignedMap.Add(1, Sort.UnsignedChar);

m\_unsignedMap.Add(2, Sort.UnsignedInt);

m\_unsignedMap.Add(4, Sort.UnsignedLongInt);

m\_minValueMap.Add(Sort.Logical, 0);

m\_minValueMap.Add(Sort.SignedChar, -128);

m\_minValueMap.Add(Sort.UnsignedChar, 0);

m\_minValueMap.Add(Sort.SignedShortInt, -128);

m\_minValueMap.Add(Sort.UnsignedShortInt, 0);

m\_minValueMap.Add(Sort.SignedInt, -32768);

m\_minValueMap.Add(Sort.UnsignedInt, 0);

m\_minValueMap.Add(Sort.Array, 0);

m\_minValueMap.Add(Sort.Pointer, 0);

m\_minValueMap.Add(Sort.SignedLongInt, -2147483648);

m\_minValueMap.Add(Sort.UnsignedLongInt, 0);

m\_maxValueMap.Add(Sort.Logical, 1);

m\_maxValueMap.Add(Sort.SignedChar, 127);

m\_maxValueMap.Add(Sort.UnsignedChar, 255);

m\_maxValueMap.Add(Sort.SignedShortInt, 127);

m\_maxValueMap.Add(Sort.UnsignedShortInt, 255);

m\_maxValueMap.Add(Sort.SignedInt, 32767);

m\_maxValueMap.Add(Sort.UnsignedInt, 65535);

m\_maxValueMap.Add(Sort.Array, 65535);

m\_maxValueMap.Add(Sort.Pointer, 65535);

m\_maxValueMap.Add(Sort.SignedLongInt, 2147483647);

m\_maxValueMap.Add(Sort.UnsignedLongInt, 4294967295);

}

}

### Static Symbol

The StaticSymbolWindows class is a sub class of StaticSymbol in Section 6.2.

StaticSymbolWindows.cs

using System;

using System.IO;

using System.Text;

using System.Globalization;

using System.Collections.Generic;

namespace CCompiler {

public class StaticSymbolWindows : StaticSymbol {

The resulting code of each function definition and static variable and constant are stored in m\_byteList.

private List<byte> m\_byteList;

A static variable or constant may access other static variables or constant, we store the addresses of the accesses in m\_accessMap. Moreover, a function may call other functions, we store the addresses of the calls in m\_callMap. The addresses are relative the beginning of the function or static object. They are later replaced by proper addresses by the linker. The difference between the addresses of the final accesses and calls is that the access addresses are absolute while the call addresses are relative the address of the call.

private IDictionary<int,string> m\_accessMap, m\_callMap;

When a called function returns to the calling is, we store the return address in m\_returnSet. The addresses are relative the address of the function call in the calling function. It is later replaced by a proper address by the linker.

private ISet<int> m\_returnSet;

We need the default constructor when loading object of the class from a file.

public StaticSymbolWindows() {

// Empty.

}

The second constructor takes a unique name, a byte list, and an access map. This constructor is intended for static variables and constants.

public StaticSymbolWindows(string uniqueName, List<byte> byteList = null,

IDictionary<int,string> accessMap = null)

:base(uniqueName) {

m\_byteList = (byteList != null) ? byteList : (new List<byte>());

m\_accessMap = (accessMap != null) ? accessMap

: (new Dictionary<int,string>());

m\_callMap = new Dictionary<int,string>();

m\_returnSet = new HashSet<int>();

}

The third constructor takes a unique name, a byte list, and an access map, a call map, and a return set. This constructor is intended for function definitions.

public StaticSymbolWindows(string uniqueName, List<byte> byteList,

IDictionary<int,string> accessMap,

IDictionary<int,string> callMap, ISet<int> returnSet)

:base(uniqueName) {

m\_byteList = byteList;

m\_accessMap = accessMap;

m\_callMap = callMap;

m\_returnSet = returnSet;

}

public List<byte> ByteList {

get { return m\_byteList; }

}

public IDictionary<int,string> AccessMap {

get { return m\_accessMap; }

}

public IDictionary<int,string> CallMap {

get { return m\_callMap; }

}

public ISet<int> ReturnSet {

get { return m\_returnSet; }

}

The Write method writes the byte list, access map, call map, and return set to the file. We call base to write the unique name of the object.

public override void Write(BinaryWriter outStream) {

base.Write(outStream);

We first write the size (number of bytes) of the byte list, then we write its bytes.

outStream.Write(m\_byteList.Count);

foreach (sbyte b in m\_byteList) {

outStream.Write(b);

}

In the same way, we write the size of the access map, and the then its key-value pairs.

outStream.Write(m\_accessMap.Count);

foreach (KeyValuePair<int,string> entry in m\_accessMap) {

outStream.Write(entry.Key);

outStream.Write(entry.Value);

}

outStream.Write(m\_callMap.Count);

foreach (KeyValuePair<int,string> entry in m\_callMap) {

outStream.Write(entry.Key);

outStream.Write(entry.Value);

}

outStream.Write(m\_returnSet.Count);

foreach (int address in m\_returnSet) {

outStream.Write(address);

}

}

The Read method mirrors the Write method above. We read the unique name, byte list, access map, call map, and return set of the object.

public override void Read(BinaryReader inStream) {

base.Read(inStream);

First, we read the size of the byte list, and then its bytes.

{ m\_byteList = new List<byte>();

int byteListSize = inStream.ReadInt32();

for (int index = 0; index < byteListSize; ++index) {

byte b = inStream.ReadByte();

m\_byteList.Add(b);

}

}

In the same way, we read the size of the access map, and then its key-value pairs.

{ m\_accessMap = new Dictionary<int,string>();

int accessMapSize = inStream.ReadInt32();

for (int index = 0; index < accessMapSize; ++index) {

int address = inStream.ReadInt32();

string name = inStream.ReadString();

m\_accessMap.Add(address, name);

}

}

{ m\_callMap = new Dictionary<int,string>();

int callMapSize = inStream.ReadInt32();

for (int index = 0; index < callMapSize; ++index) {

int address = inStream.ReadInt32();

string name = inStream.ReadString();

m\_callMap.Add(address, name);

}

}

{ m\_returnSet = new HashSet<int>();

int returnSetSize = inStream.ReadInt32();

for (int index = 0; index < returnSetSize; ++index) {

int address = inStream.ReadInt32();

m\_returnSet.Add(address);

}

}

}

}

}

### Static Value

When creating a static value in the Windows environment, the executable code is stored in a list of bytes. Moreover, we also need an access map to keep track of the accesses of the value. For instance, a pointer may point to another static value.

ConstantExpression.cs

public static StaticSymbol Value(string uniqueName, Type type,

object value) {

// ...

if (Start.Linux) {

// ...

}

if (Start.Windows) {

List<byte> byteList = new List<byte>();

IDictionary<int,string> accessMap = new Dictionary<int,string>();

AssemblyCodeGenerator.

GenerateTargetWindows(assemblyCodeList, byteList,

accessMap, null, null);

return (new StaticSymbolWindows(uniqueName, byteList, accessMap));

}

// ...

}

}

}

### Function End

When creating a function in the Windows environment, we need the byte list and access map as in the static value case above. However, we also need a call map and a return set to keep track of the function calls and the return addresses of those calls.

MiddleCodeGenerator.cs

public static void FunctionEnd(Statement statement) {

// ...

if (Start.Linux) {

// ...

}

if (Start.Windows) {

List<byte> byteList = new List<byte>();

IDictionary<int,string> accessMap = new Dictionary<int,string>();

IDictionary<int,string> callMap = new Dictionary<int,string>();

ISet<int> returnSet = new HashSet<int>();

AssemblyCodeGenerator.GenerateTargetWindows

(assemblyCodeList, byteList, accessMap, callMap, returnSet);

StaticSymbol staticSymbol =

new StaticSymbolWindows(SymbolTable.CurrentFunction.UniqueName,

byteList, accessMap, callMap, returnSet);

SymbolTable.StaticSet.Add(staticSymbol);

}

// ...

}

### Target Code Generation

public static void GenerateTargetWindows

(List<AssemblyCode> assemblyCodeList, List<byte> byteList,

IDictionary<int,string> accessMap, IDictionary<int,string> callMap,

ISet<int> returnSet) {

AssemblyCodeGenerator objectCodeGenerator =

new AssemblyCodeGenerator(assemblyCodeList);

objectCodeGenerator.WindowsJumpInfo();

objectCodeGenerator.WindowsByteList

(byteList, accessMap, callMap, returnSet);

}

### Exit

To exit the execution and return control to the surrounding operating system in the Windows environment, we perform an interrupt 33 system call, with the value 76 in the ah register. We store the return value in the al register. If there is no return value, we store zero in al.

AssemblyCodeGenerator.cs

public void Exit(MiddleCode middleCode) {

Symbol exitSymbol = (Symbol) middleCode[0];

if (Start.Linux) {

// ...

}

if (Start.Windows) {

if (exitSymbol != null) {

LoadValueToRegister(exitSymbol, Register.al);

}

else {

AddAssemblyCode(AssemblyOperator.mov, Register.al,

BigInteger.Zero);

}

AddAssemblyCode(AssemblyOperator.mov, Register.ah,

(BigInteger) 76); // 0x4C

AddAssemblyCode(AssemblyOperator.interrupt, (BigInteger) 33); // 0x21

}

}

### Initialization Code

The initialization of data blocks differs between the Linux and the Windows environment.

AssemblyCodeGenerator.cs

public static void InitializationCodeList() {

List<AssemblyCode> assemblyCodeList = new List<AssemblyCode>();

AddAssemblyCode(assemblyCodeList, AssemblyOperator.comment,

"Initializerialize Stack Pointer");

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

AssemblyCode.RegularFrameRegister, Linker.CallStackStart);

if (Start.Linux) {

// ...

}

if (Start.Windows) {

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov\_word,

null, 65534, (BigInteger)65534);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.fstcw,

AssemblyCode.RegularFrameRegister, 0);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.or\_word,

AssemblyCode.RegularFrameRegister, 0, (BigInteger) 0x0C00);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.fldcw,

AssemblyCode.RegularFrameRegister, 0);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Linker.CallStackStart, 0, BigInteger.Zero,

TypeSize.PointerSize);

List<byte> byteList = new List<byte>();

IDictionary<int, string> accessMap = new Dictionary<int, string>();

IDictionary<int, string> callMap = new Dictionary<int, string>();

ISet<int> returnSet = new HashSet<int>();

AssemblyCodeGenerator.GenerateTargetWindows(assemblyCodeList,

byteList, accessMap, callMap, returnSet);

StaticSymbol staticSymbol =

new StaticSymbolWindows(AssemblyCodeGenerator.InitializerName,

byteList, accessMap, callMap, returnSet);

SymbolTable.StaticSet.Add(staticSymbol);

}

}

### Command Line Arguments

The code for obtaining the command line arguments differs between the Linux code in Chapter XXX. The main difference is the arguments (excluding the file name) is stored at address 129, each word may begin with one or several spaces, and that the list is finished by a return character. The file name is unfortunately not stored, why we instead let the linker stored the name of the final executable file with the $Path name

main

’a’

’i’

’n’

’\0’

’e’

’l’

’l’

’o’

’ ’

’W’

’o’

’r’

’l’

’d’

’ ’

Return address: 0

Frame pointer of

calling function

Ellipse pointer of

calling function

argc: 3

argv

null

’\r’

129

’m’

’H’

$Path

’ ’

’ ’

’e’

’l’

’l’

’o’

’ ’

’W’

’o’

’r’

’l’

’d’

’ ’

’\0’

129

’H’

’\0’

’ ’

Before

After

We start by saving the current value of bp in si, we will need it when we set the main parameter list at the end. We then move to $Path address to the current address of bp and increase it value by two bytes (a pointer in the Windows environment is bytes long). By now we have the name of the execution file stored at the beginning of the array.

mov si, bp

mov word [bp], $Path

add bp, 2

We let ax hold the number of arguments, which is one to begin with since we have already included the execution file name. We let bx hold the current address of the characters in the command line, which starts at address 129. We begin by checking the special case where we have no arguments at all. In that case, address 129 hold the return character (with ASCII value 13) and we jump to ListDone.

mov ax, 1

mov bx, 129

cmp byte [bx], 13

je ListDone

If the argument list is not empty, we start by iterating through the looping through the potential initial space characters. They shall not be included in the argument word. We continue to increase bx and jump back to SpaceLoop an long as we encounter space characters. Note that we cannot encounter a return character since they only occur after the last argument word. The case without any arguments words, where the list only holds one single return character, we have dealt with above.

SpaceLoop:

cmp byte [bx], 32

jne WordStart

inc bx

jmp SpaceLoop

When we encounter a word different from space, we have found the beginning a new argument word. We increment ax, holding the number of arguments, with one. We also store the address of the first character in the word and the array pointed at by bp and increase bp with two bytes.

WordStart:

inc ax

mov word [bp], bx

add bp, 2

Then we iterate the argument word, until we encounter a space (ASCII value 32) of a return (ASCII value 13). In case of a space character, the current word is done and we jump to WordDone to address the next word. In case of a return character, the list is done and we jump to ListDone. If none of those cases apply, we increase bx and jump back to WordLoop to handle the next character of the word.

WordLoop:

cmp byte [bx], 32

je WordDone

cmp byte [bx], 13

je ListDone

inc bx

jmp WordLoop

When we are finish with a word, we replace its finishing space with a zero character. We also increase bx to let it point at the character following the word, and jump back to SpaceLoop to find the next word in the list.

WordDone:

mov byte [bx], 0

inc bx

jmp SpaceLoop

When we have iterated through the whole list, there are som finishing up to do. Just as when we encountered a space character at the end of a word, we need to change the last character to a zero character also when we encounter a return character. We also add a zero to the array, since the last entry in the argument array shall be a null pointer. Finally, we increase bx with two.

ListDone:

mov byte [bx], 0

mov word [bp], 0

add bp, 2

By now, bx points at the beginning of the first activation record, for the initial call to main. We need to set its return address (offset zero) to zero, for the return statement to exit the execution, and the argc and argv parameters. We mov the number of arguments (ax) to argc (offset six), and the beginning of the argument list (si) to argv (offset eight).

mov word [bp], 0

mov [bp + 6], ax

mov [bp + 8], si

Below follows the code that generates the argument list code.

AssemblyCodeGenerator.cs

public static void ArgumentCodeList() {

List<AssemblyCode> assemblyCodeList = new List<AssemblyCode>();

if (Start.Linux) {

// ...

}

if (Start.Windows) {

/\* mov si, bp

mov word [bp], $Path

add bp, 2

mov ax, 1

mov bx, 129

cmp byte [bx], 13

je ListDone \*/

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Register.si, Register.bp);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov\_word,

Register.bp, 0, AssemblyCodeGenerator.PathName);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.add,

Register.bp, (BigInteger) 2);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Register.ax, BigInteger.One);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Register.bx, (BigInteger) 129);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.cmp\_byte,

Register.bx, 0, (BigInteger) 13);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.je,

null, assemblyCodeList.Count + 17);

/\* SpaceLoop:

cmp byte [bx], 32

jne WordStart

inc bx

jmp SpaceLoop \*/

AddAssemblyCode(assemblyCodeList, AssemblyOperator.cmp\_byte,

Register.bx, 0, (BigInteger) 32);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.jne,

null, assemblyCodeList.Count + 3);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.inc, Register.bx);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.jmp,

null, assemblyCodeList.Count - 3);

/\* WordStart:

inc ax

mov word [bp], bx

add bp, 2 \*/

AddAssemblyCode(assemblyCodeList, AssemblyOperator.inc, Register.ax);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Register.bp, 0, Register.bx);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.add,

Register.bp, (BigInteger) 2);

/\* WordLoop:

cmp byte [bx], 32

je WordDone

cmp byte [bx], 13

je ListDone

inc bx

jmp WordLoop \*/

AddAssemblyCode(assemblyCodeList, AssemblyOperator.cmp\_byte,

Register.bx, 0, (BigInteger) 32);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.je,

null, assemblyCodeList.Count + 5);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.cmp\_byte,

Register.bx, 0, (BigInteger) 13);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.je,

null, assemblyCodeList.Count + 6);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.inc, Register.bx);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.jmp,

null, assemblyCodeList.Count - 5);

/\* WordDone:

mov byte [bx], 0

inc bx

jmp SpaceLoop \*/

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov\_byte,

Register.bx, 0, BigInteger.Zero);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.inc, Register.bx);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.jmp,

null, assemblyCodeList.Count - 15);

/\* ListDone:

mov byte [bx], 0

mov word [bp], 0

add bp, 2

mov word [bp], 0

mov [bp + 6], ax

mov [bp + 8], si \*/

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov\_byte,

Register.bx, 0, BigInteger.Zero);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov\_word,

Register.bp, 0, BigInteger.Zero);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.add,

Register.bp, (BigInteger) 2);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov\_word,

Register.bp, 0, BigInteger.Zero);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Register.bp, 6, Register.ax);

AddAssemblyCode(assemblyCodeList, AssemblyOperator.mov,

Register.bp, 8, Register.si);

List<byte> byteList = new List<byte>();

IDictionary<int, string> accessMap = new Dictionary<int, string>();

IDictionary<int, string> callMap = new Dictionary<int, string>();

ISet<int> returnSet = new HashSet<int>();

AssemblyCodeGenerator.

GenerateTargetWindows(assemblyCodeList, byteList,

accessMap, callMap, returnSet);

StaticSymbol staticSymbol =

new StaticSymbolWindows(AssemblyCodeGenerator.ArgsName, byteList,

accessMap, callMap, returnSet);

SymbolTable.StaticSet.Add(staticSymbol);

}

}

### Windows Jump Info

The WindowsJumpInfo method first changes the jump addresses from middle code line numbers to assembly code line numbers, and then to relative byte size addresses. To begin with we need the middleToAssemblyMap map to keep track of the assembly code addresses.

AssemblyCodeGenerator.cs

private void WindowsJumpInfo() {

IDictionary<int,int> middleToAssemblyMap = new Dictionary<int,int>();

We iterate through the assembly code list and add the assembly code line number each time we encounter a new-middle-code instruction. We really need just the line numbers of first of the assembly line that corresponds to a middle code instruction, since all jumps are to those addresses.

for (int assemblyIndex = 0;

assemblyIndex < m\_assemblyCodeList.Count; ++assemblyIndex) {

AssemblyCode assemblyCode = m\_assemblyCodeList[assemblyIndex];

if (assemblyCode.Operator == AssemblyOperator.new\_middle\_code) {

int middleIndex = (int) assemblyCode[0];

middleToAssemblyMap.Add(middleIndex, assemblyIndex);

assemblyCode.Operator = AssemblyOperator.empty;

}

}

Then we iterate through the assembly code list and change all jump targets from middle code line numbers to assembly code line numbers. Note that the middle code line number was stored in the third position (index 2) of the instruction, and that we set the assembly line number to the second position (index 1).

for (int line = 0; line < m\_assemblyCodeList.Count; ++line) {

AssemblyCode assemblyCode = m\_assemblyCodeList[line];

if (/\*(assemblyCode[0] == null) &&\*/ !(assemblyCode[1] is int) &&

(assemblyCode.IsRelationNotRegister() ||

assemblyCode.IsJumpNotRegister())) {

int middleTarget = (int) assemblyCode[2];

assemblyCode[1] = middleToAssemblyMap[middleTarget];

}

}

When we have changed the jump targets from middle code line numbers to assembly code line numbers, the second step is to change the targets into the number of bytes between the next assembly instruction (the instruction following the current instruction) and the target instruction, which is a negative value in case of backward jumps.

First, we need the assemblyToByteMap map to keep track of the address of each assembly code instruction, counted in bytes from the beginning of the code. Then we iterate through the assembly code list and add the byte address of each instruction. At the moment, all jump instructions are equally large (in bytes), but that may change later on.

IDictionary<int,int> assemblyToByteMap = new Dictionary<int,int>();

{ int byteSize = 0, line = 0;

foreach (AssemblyCode assemblyCode in m\_assemblyCodeList) {

assemblyToByteMap.Add(line++, byteSize);

if (!(assemblyCode.IsRelationNotRegister() ||

assemblyCode.IsJumpNotRegister())) {

byteSize += assemblyCode.ByteList().Count;

}

}

assemblyToByteMap.Add(m\_assemblyCodeList.Count, byteSize);

}

When we change the assembly code targets to relative byte targets, we may have to iterate several times because short jumps (between -128 and 127, inclusive) instructions are smaller than long jumps instructions. At the beginning, all jumps are long jumps, but some of them may be changed to short jump during the iteration.

while (true) {

for (int line = 0; line < (m\_assemblyCodeList.Count - 1); ++line) {

AssemblyCode thisCode = m\_assemblyCodeList[line],

nextCode = m\_assemblyCodeList[line + 1];

if (/\*(thisCode[0] == null) && \*/

(thisCode.IsRelationNotRegister() ||

thisCode.IsJumpNotRegister())) {

int assemblyTarget = (int) thisCode[1];

The byteSource variable is the byte address of the target assembly instruction following the current instruction, byteTarget is the byte address of the target instruction, and byteDistance is difference between byteTarget and byteSource, which becomes the final target address.

int byteSource = assemblyToByteMap[line + 1],

byteTarget = assemblyToByteMap[assemblyTarget];

int byteDistance = byteTarget - byteSource;

Assert.ErrorXXX(byteDistance != 0);

thisCode[0] = byteDistance;

}

}

When the jump instruction has been given proper target address, we iterate through the assembly list and update the assemblyToByteMap. Some byte address may have been changed because some long jump instructions have been changed to short jump instructions. If any address has been changed, update is set to true and the overall loop will perform another iteration.

bool update = false;

{ int byteSize = 0, line = 0;

foreach (AssemblyCode objectCode in m\_assemblyCodeList) {

if (assemblyToByteMap[line] != byteSize) {

assemblyToByteMap[line] = byteSize;

update = true;

}

byteSize += objectCode.ByteList().Count;

++line;

}

}

If nothing has been changed from the previous iteration, every assembly jump instruction holds the correct target address, and we break the loop.

if (!update) {

break;

}

}

Finally, we need to iterate through the assembly code list, and change the return addresses in the same way as the byte target addresses above.

for (int line = 0; line < m\_assemblyCodeList.Count; ++line) {

AssemblyCode assemblyCode = m\_assemblyCodeList[line];

if (assemblyCode.Operator == AssemblyOperator.return\_address) {

int middleAddress = (int) ((BigInteger) assemblyCode[2]);

int assemblyAddress = middleToAssemblyMap[middleAddress];

int byteAddress = assemblyToByteMap[assemblyAddress];

assemblyCode[2] = (BigInteger) byteAddress;

}

}

}

### Windows Byte List

The WindowsByteList method iterates through the assembly code list and generates the byte list for each instruction by calling ByteList in AssemblyCode. It also adds names to the accessMap, callMap, and returnSet parameters.

AssemblyCodeGenerator.cs

private void WindowsByteList(List<byte> byteList,

IDictionary<int,string> accessMap,

IDictionary<int,string> callMap,

ISet<int> returnSet) {

foreach (AssemblyCode assemblyCode in m\_assemblyCodeList) {

We add the byte list of the instruction to the byte list parameter.

byteList.AddRange(assemblyCode.ByteList());

Unless the instruction is a label, comment, or zero sequence, we check whether there are any names in the code that shall be added to the access map or the call map, or the return set.

if ((assemblyCode.Operator != AssemblyOperator.label) &&

(assemblyCode.Operator != AssemblyOperator.comment) &&

(assemblyCode.Operator != AssemblyOperator.define\_zero\_sequence)){

If the instruction is an address definition, we add its name to the access map.

if (assemblyCode.Operator == AssemblyOperator.define\_address) {

string name = (string) assemblyCode[0];

accessMap.Add(byteList.Count - TypeSize.PointerSize, name);

}

If the instruction is a value definition, we need to look into its sort. If it is a pointer or a static address, we add its name.

else if (assemblyCode.Operator == AssemblyOperator.define\_value) {

Sort sort = (Sort) assemblyCode[0];

object value = assemblyCode[1];

if (sort == Sort.Pointer) {

if (value is string) {

accessMap.Add(byteList.Count - TypeSize.PointerSize,

(string) value);

}

else if (value is StaticAddress) {

StaticAddress staticAddress = (StaticAddress) value;

accessMap.Add(byteList.Count - TypeSize.PointerSize,

staticAddress.UniqueName);

}

}

}

In case of a function call, add the name of the function to be called to the call map.

else if ((assemblyCode.Operator == AssemblyOperator.call) &&

(assemblyCode[0] is string)) {

string calleeName = (string) assemblyCode[0];

int address = byteList.Count - TypeSize.PointerSize;

callMap.Add(address, calleeName);

}

In case of a function return, we add the address the return set.

else if (assemblyCode.Operator == AssemblyOperator.return\_address) {

int address = byteList.Count - TypeSize.PointerSize;

returnSet.Add(address);

}

If the first operand is a string, we have two cases: if the third operand is a value of BigInteger, we need decrease the address with its size and with the size of a pointer. Otherwise, we just decrease the address with the pointer size.

else if (assemblyCode[0] is string) { // Add [g + 1], 2

if (assemblyCode[2] is BigInteger) {

int size = AssemblyCode.SizeOfValue((BigInteger)assemblyCode[2],

assemblyCode.Operator);

int address = byteList.Count - TypeSize.PointerSize - size;

accessMap.Add(address, (string) assemblyCode[0]);

}

else {

int address = byteList.Count - TypeSize.PointerSize;

accessMap.Add(address, (string) assemblyCode[0]);

}

}

If the second or third operand is a string, we add its name the address, which is the size of the instruction minus the pointer size.

else if (assemblyCode[1] is string) { // mov ax, [g + 1]; mov ax, g

int address = byteList.Count - TypeSize.PointerSize;

accessMap.Add(address, (string) assemblyCode[1]);

}

else if (assemblyCode[2] is string) { // Add [bp + 2], g

int address = byteList.Count - TypeSize.PointerSize;

accessMap.Add(address, (string) assemblyCode[2]);

}

}

}

}

}

}

### Byte List

The ByteList method return the assembly code instruction converted to a list of bytes.

AssemblyCode.cs

public List<byte> ByteList() {

object operand0 = m\_operandArray[0],

operand1 = m\_operandArray[1],

operand2 = m\_operandArray[2];

For empty instruction, or for a label or comment instruction we just return an empty list.

if ((Operator == AssemblyOperator.empty) ||

(Operator == AssemblyOperator.label) ||

(Operator == AssemblyOperator.comment)) {

return (new List<byte>());

}

For the definition of an address, we add the offset to the byte list.

else if (Operator == AssemblyOperator.define\_address) {

int offset = (int) operand1;

List<byte> byteList = new List<byte>(new byte[TypeSize.PointerSize]);

LoadByteList(byteList, 0, TypeSize.PointerSize, (BigInteger) offset);

return byteList;

}

For a zero value, we just return a list of zeros.

else if (Operator == AssemblyOperator.define\_zero\_sequence) {

int size = (int) operand0;

return (new List<byte>(new byte[size]));

}

A value can be a floating value, a string, a static address, a pointer, or an integral value.

else if (Operator == AssemblyOperator.define\_value) {

Sort sort = (Sort) operand0;

object value = operand1;

List<byte> byteList;

In case of float, double, or long double, we call GetBytes in the standard library BitConverter class.

if (sort == Sort.Float) {

float floatValue = (float) ((decimal) operand0);

byteList = new List<byte>(BitConverter.GetBytes(floatValue));

}

else if ((sort == Sort.Double) || (sort == Sort.LongDouble)) {

double doubleValue = (double) ((decimal) value);

byteList = new List<byte>(BitConverter.GetBytes(doubleValue));

}

In case of string we fill the byte list with the characters of the string and add the terminating zero character.

else if (sort == Sort.String) {

string text = (string) value;

byteList = new List<byte>();

foreach (char c in text) {

byteList.Add((byte) c);

}

byteList.Add((byte) 0);

}

Otherwise, we have an integral of pointer value. If the value is a static address, we load its offset into the byte list. The address itself will be added by the linker later on. If the value is not a static address, we just load it into the byte list.

else {

int size = TypeSize.Size(sort);

byteList = new List<byte>(new byte[size]);

if (value is StaticAddress) {

StaticAddress staticAddress = (StaticAddress) value;

LoadByteList(byteList, 0, size,

(BigInteger) staticAddress.Offset);

}

else {

LoadByteList(byteList, 0, size, (BigInteger) value);

}

}

return byteList;

}

Next, we have a jump or a call to a register; that is, a jump of call to an address stored in a register.

else if (IsJumpRegister() || IsCallRegister()) {

Register register = (Register) operand0;

return LookupByteArray(AssemblyOperator.jmp, register);

}

A regular function call is equivalent with a long jump. The actual address of the called function will be set by the linker later on.

else if (IsCallNotRegister()) {

return LookupByteArray(AssemblyOperator.jmp, TypeSize.PointerSize);

}

else if (Operator == AssemblyOperator.return\_address) {

Register register = (Register) operand0;

int offset = (int) operand1;

int size = SizeOfValue(offset);

int address = (int)((BigInteger)operand2);

List<byte> byteList =

LookupByteArray(AssemblyOperator.mov\_word, register,

size, TypeSize.PointerSize);

LoadByteList(byteList, byteList.Count - (size + TypeSize.PointerSize),

size, offset);

LoadByteList(byteList, byteList.Count - TypeSize.PointerSize,

TypeSize.PointerSize, address);

return byteList;

}

The size of the address is either one (short jump) or two (long jump), which we load into the byte list.

else if (IsRelationNotRegister() || IsJumpNotRegister()) {

int address = (int) operand0;

int size = SizeOfValue(address);

if (address == 127) { // XXX

size = 2;

}

List<byte> byteList = LookupByteArray(Operator, size);

LoadByteList(byteList, byteList.Count - size, size, address);

return byteList;

}

Let us continue with the operations that perform computations and let us start with the nullary operators; that is, operators that take no operands. For instance: lahf. We call LookupByteArray with the operator as argument to obtain and return the assay of bytes corresponding to the assembly instruction.

else if ((operand0 == null) && (operand1 == null) &&

(operand2 == null)) {

return LookupByteArray(Operator);

}

Then we continue with the unary operators; that is, operators that take one operand. We have three different cases:

|  |  |  |
| --- | --- | --- |
| **Category** | **Description** | **Examples** |
| 1 | unary operator register | inc ax |
| 2 | unary operator integer value | int 33 |
| 3 | unary operator [register + offset]  unary operator [static name + offset] | neg word [bp + 2]  not word [stdin + 4] |

The operators of the first category take a register as operand. We call LookupByteArray with the operator and register.

// inc ax

else if ((operand0 is Register) && (operand1 == null) &&

(operand2 == null)) {

Register register = (Register) operand0;

return LookupByteArray(Operator, register);

}

The operators of the second category take an integer value as operand. For instance: int 33. We call LookupByteArray just like in the previous cases. But we must also load the array with the integer value of the operation, which we do by calling LoadByteList with the value and its size. We determine the size of the value by calling SizeOfValue. Note that we have to subtract the size of the value from the size of the byte list to obtain the location to insert the value in the byte list. Integer values are always BigInteger objects.

// int 33

else if ((operand0 is BigInteger) && (operand1 == null) &&

(operand2 == null)) {

BigInteger value = (BigInteger) operand0;

int size = SizeOfValue(value);

List<byte> byteList = LookupByteArray(Operator, size);

LoadByteList(byteList, byteList.Count - size, size, value);

return byteList;

}

In the third category of the unary operations the operand is an address, made up by a register and an offset or a static variable and an offset. For instance: inc [bp + 2] or inc [global + 4]. Note that offsets are always int values. If the address includes a register, the size of the offset is determined by calling SizeOfValue. If the address instead includes a static variable, the offset size is always the size of a pointer.

// inc [bp + 2]; inc [global + 4]

else if (((operand0 is Register) || (operand0 is string)) &&

(operand1 is int) && (operand2 == null)) {

int offset = (int) operand1;

int size = (operand0 is Register) ? SizeOfValue(offset)

: TypeSize.PointerSize;

List<byte> byteList = LookupByteArray(Operator, operand0, size);

LoadByteList(byteList, byteList.Count - size, size, offset);

return byteList;

}

For binary operators, we have seven categories. Note that we in category six have a special case, where we load a value into an address given only by an offset. This case is used in initialization at the beginning of the code.

|  |  |  |
| --- | --- | --- |
| **Category** | **Description** | **Examples** |
| 1 | binary operator register, register | mov ax, bx |
| 2 | binary operator register, static name | add ax, stdin |
| 3 | binary operator register, integer value | sub ax, 123 |
| 4 | binary operator [register + offset], register  binary operator [static name + offset], register | mov [bp + 2], bx  mov [stdin + 4], bx |
| 5 | binary operator [register + offset], static name  binary operator [static name + offset], static name | add word [bp + 2], stdin  add word [stdin + 4], stdin |
| 6 | binary operator [register + offset], integer value  binary operator [static name + offset], integer value  binary operator [offset], integer value | sub word [bp + 2], 123  sub word [stdin + 4], 123  mov word [65534], 65534 |
| 7 | binary operator register, [register + offset]  binary operator register, [static name + offset] | mov ax, [bp + 2]  add ax, [stdin +4] |

In category one, we assign a register the value of another register.

// 1: mov ax, bx

else if ((operand0 is Register) && (operand1 is Register) &&

(operand2 == null)) {

Register toRegister = (Register) operand0,

fromRegister = (Register) operand1;

return LookupByteArray(Operator, toRegister, fromRegister);

}

In category two, we assign a register the address a static variable. We have to load the zero value into the byte list, to mark that there is not offset. The linker will add the actual address of the variable later on.

// 2: mov ax, global

else if ((operand0 is Register) && (operand1 is string) &&

(operand2 == null)) {

Register register = (Register) operand0;

List<byte> byteList =

LookupByteArray(Operator, register, TypeSize.PointerSize);

LoadByteList(byteList, byteList.Count - TypeSize.PointerSize,

TypeSize.PointerSize, 0);

return byteList;

}

In category three, we load an integral value into a register. We must check the operator, in case of mov or and, the size of the value is given by the register. Otherwise, the size is given by the value.

// 3: mov ax, 123

else if ((operand0 is Register) && (operand1 is BigInteger) &&

(operand2 == null)) {

Register register = (Register) operand0;

BigInteger value = (BigInteger) operand1;

int size = ((Operator == AssemblyOperator.mov) ||

(Operator == AssemblyOperator.and))

? SizeOfRegister(register) : SizeOfValue(value);

List<byte> byteList = LookupByteArray(Operator, register, size);

LoadByteList(byteList, byteList.Count - size, size, value);

return byteList;

}

In category four, we assign the value of a register to an address, defined by a register and an offset or a static variable and an offset. Also in this case, we need to obtain the size of the offset. Its is given by SizeOfValue in case of a register address. In case of a static variable address, the size is always the size of a pointer.

// 4: mov [bp + 2], ax; mov [global + 4], ax

else if (((operand0 is Register) || (operand0 is string)) &&

(operand1 is int) && (operand2 is Register)) {

Register register = (Register) operand2;

int offset = (int) operand1;

int size = (operand0 is Register) ? SizeOfValue(offset)

: TypeSize.PointerSize;

List<byte> byteList =

LookupByteArray(Operator, operand0, size, register);

LoadByteList(byteList, byteList.Count - size, size, offset);

return byteList;

}

In category five, we assign the address of a static variable to an address, given by a register and an offset or a static variable and an offset. We load the offset into the byte list. However, we also need to load the zero value to the byte list, to mark that the address is zero at the moment. The linker will add the actual address of the variable later on.

// 5: mov [bp + 2], global; mov [global + 4], global

else if (((operand0 is Register) || (operand0 is string)) &&

(operand1 is int) && (operand2 is string)) {

int offset = (int) operand1;

int size = (operand0 is Register) ? SizeOfValue(offset)

: TypeSize.PointerSize;

List<byte> byteList = LookupByteArray(Operator, operand0,

size, TypeSize.PointerSize);

LoadByteList(byteList, byteList.Count -

(size + TypeSize.PointerSize), size, offset);

LoadByteList(byteList, byteList.Count - TypeSize.PointerSize,

TypeSize.PointerSize, 0);

return byteList;

}

In category six, we assign an integral value to an address. We need to determine the size of both the offset and the value, and then load them both into the byte list. We also have the special case where the static variable is null, which occurs in the initialization in the beginning of the code, such as mov word [65534], 65534.

// 6: mov [bp + 2], 123; mov [global + 4], 123

// mov [null + 4], 123; Special case

else if (((operand0 is Register) || (operand0 is string) ||

(operand0 == null)) && (operand1 is int) &&

(operand2 is BigInteger)) {

int offset = (int) operand1;

BigInteger value = (BigInteger) operand2;

int offsetSize = (operand0 is Register) ? SizeOfValue(offset)

: TypeSize.PointerSize,

valueSize = SizeOfValue(value, Operator);

List<byte> byteList =

LookupByteArray(Operator, operand0, offsetSize, valueSize);

LoadByteList(byteList, byteList.Count - (offsetSize + valueSize),

offsetSize, offset);

LoadByteList(byteList, byteList.Count - valueSize,

valueSize, value);

return byteList;

}

In category seven, we assign a register the value stored on an address, defined by a register and an offset or a static variable and an offset.

// 7: mov ax, [bp + 2]; mov ax, [global + 4]

else if ((operand0 is Register) && ((operand1 is Register) ||

(operand1 is string)) && (operand2 is int)) {

Register register = (Register) operand0;

int offset = (int) operand2;

int size = (operand1 is Register) ? SizeOfValue(offset)

: TypeSize.PointerSize;

List<byte> byteList =

LookupByteArray(Operator, register, operand1, size);

LoadByteList(byteList, byteList.Count - size, size, offset);

return byteList;

}

}

The LoadByteList method load the byte list with a value.

public static void LoadByteList(IList<byte> byteList, int index,

int size, BigInteger value) {

switch (size) {

case 1: {

if (value < 0) {

byteList[index] = (byte) ((sbyte) value);

}

else {

byteList[index] = (byte) value;

}

}

break;

case 2: {

if (value < 0) {

short shortValue = (short) value;

byteList[index] = (byte) ((sbyte) shortValue);

byteList[index + 1] = (byte) ((sbyte) (shortValue >> 8));

}

else {

ushort ushortValue = (ushort) value;

byteList[index] = (byte) ushortValue;

byteList[index + 1] = (byte) (ushortValue >> 8);

}

}

break;

case 4: {

if (value < 0) {

int intValue = (int) value;

byteList[index] = (byte) ((sbyte) intValue);

byteList[index + 1] = (byte) ((sbyte) (intValue >> 8));

byteList[index + 2] = (byte) ((sbyte) (intValue >> 16));

byteList[index + 3] = (byte) ((sbyte) (intValue >> 24));

}

else {

uint uintValue = (uint) value;

byteList[index] = (byte) uintValue;

byteList[index + 1] = (byte) (uintValue >> 8);

byteList[index + 2] = (byte) (uintValue >> 16);

byteList[index + 3] = (byte) (uintValue >> 24);

}

}

break;

}

}

The LookupByteArray method looks up the assembly code instruction in the MainArrayMap map in ObjectCodeTable.

public static List<byte> LookupByteArray(AssemblyOperator objectOp,

object operand1 = null, object operand2 = null,

object operand3 = null) {

if ((objectOp == AssemblyOperator.shl) ||

(objectOp == AssemblyOperator.shr)) {

operand1 = (operand1 is BigInteger) ? 0L : operand1;

operand2 = (operand2 is BigInteger) ? 0L : operand2;

operand3 = (operand3 is BigInteger) ? 0L : operand3;

}

If any of the operands is a string, it shall be changed to null.

operand0 = (operand0 is string) ? null : operand0;

operand1 = (operand1 is string) ? null : operand1;

operand2 = (operand2 is string) ? null : operand2;

ObjectCodeInfo info =

new ObjectCodeInfo(objectOp, operand1, operand2, operand3);

byte[] byteArray = ObjectCodeTable.MainArrayMap[info];

Assert.ErrorXXX(byteArray != null);

List<byte> byteList = new List<byte>();

foreach (byte b in byteArray) {

byteList.Add(b);

}

return byteList;

}

## Linking

The linker if the final part of the compilation process. Its merges together the compiled files and generates an executable file. To be more exact, it has the following tasks:

* To resolve access to static symbols
* To resolve function calls.
* To resolve return addresses.

### The Linker Class

The LinkerInfo class holds the information of a function definition or a static variable:

* m\_fullName: the full name of the function or static variable. The full name of a function or global variable with external linkage is just its regular name, while the full name of
* m\_entryPoint: the function call entry point, which may differ from the start address of the function. Actually, a function’s start address and its the entry point differs in one case only: when main is defined with the argc and argv parameters and is called recursively, which is allowed in the C standard. In that case, the first part of the main function holds code for initializing argc and argv, which shall not be executed in a recursive call.
* m\_accessMap: the map of all accesses of function or static variables.
* m\_callMap: the map of all function calls. The reason we distinguish between m\_accessMap and m\_callMap is that accesses are absolute while function calls are relative.
* m\_returnSet: the set of all return address assignments.
* m\_byteList: the actual code of the function or initialization data of the static variable. If the static variable is uninitialized, all data is zero.

The Linker class merge code of the object files and modifies the global accesses, function calls, and return assignments and saves the final code in a com-file, which is executable in 16-bits Windows. More specific, the linker has the following tasks:

* First we load all function and static variables from all the object files and to make sure that two elements do not share the same name.
* The we identify the main function and trace all functions and global variables reachable from main. All elements not reachable from main are discarded.
* For each function or global variable we modifies all accesses, and for each function we modify all calls and return assignments.
* Finally, we save the code and data for all elements reachable from main in the com-file.

There are several maps:

* m\_globalMap: holds all functions and static variables of all object files.
* m\_globalList: holds all functions and static variables reachable from main, where main is always placed first.
* m\_addressMap holds the beginning of each function or static variables in the final code.
* m\_entryMap holds the entry point of each function. As mention in Section 13.2, a function’s start address and its the entry point differs only when main is defined with the argc and argv parameters and is called recursively.

Linker.cs

using System;

using System.IO;

using System.Collections.Generic;

namespace CCompiler {

public class Linker {

The m\_globalMap map holds all the static symbols of the source code, while m\_globalList holds only the symbols that are reachable (directly or indirectly) from the main function. The final code is generated from the symbols of m\_globalList, the other symbols are omitted from the final code. The m\_addressMap holds the address of each symbol in m\_globalList. Finally, m\_totalSize holds the current total size of the symbols of m\_globalList and is used to add the positions of the symbols to m\_addressMap.

private IDictionary<string,StaticSymbolWindows> m\_globalMap =

new Dictionary<string,StaticSymbolWindows>();

private List<StaticSymbolWindows> m\_globalList =

new List<StaticSymbolWindows>();

private IDictionary<string,int> m\_addressMap =

new Dictionary<string,int>();

private int m\_totalSize = 256;

The Add method is called for each static symbol of the source code. If two symbols have the same unique name, that does not end with the number identification character (‘#’), we have two symbols with external linkage, which is not allowed. However, if the name ends with the identification character, it only means that two numerical constants with the same value is present in two different files, with is allowed. We simply refrain from adding the second symbol to the map.

public void Add(StaticSymbolWindows staticSymbol) {

string uniqueName = staticSymbol.UniqueName;

if (!m\_globalMap.ContainsKey(uniqueName)) {

m\_globalMap.Add(uniqueName, staticSymbol);

}

else {

Assert.Error(uniqueName.EndsWith(Symbol.NumberId),

SimpleName(uniqueName), Message.Duplicate\_global\_name);

}

}

The Generate method writes the final executable code to the target file. Its task is to identify the symbols reachable from the main function, and to replace the names of accessed symbols with called functions with proper addresses as well as to replace the function returns addresses with proper addresses. However, we need to start by including the code for the initialization execution code and, optionally, the code for handling the command line arguments.

public void Generate(FileInfo targetFile) {

The code for initialization shall be added to the code at the beginning, before the main function

{ StaticSymbolWindows initializerInfo =

m\_globalMap[AssemblyCodeGenerator.InitializerName];

m\_globalList.Add(initializerInfo);

m\_totalSize += initializerInfo.ByteList.Count;

m\_addressMap.Add(AssemblyCodeGenerator.InitializerName, 0);

}

In case of command line arguments, we add its symbol after the initialization code symbol and before the main function code.

StaticSymbolWindows pathNameSymbol = null;

if (m\_globalMap.ContainsKey(AssemblyCodeGenerator.ArgsName)) {

{ StaticSymbolWindows argsInfo =

m\_globalMap[AssemblyCodeGenerator.ArgsName];

m\_globalList.Add(argsInfo);

m\_totalSize += argsInfo.ByteList.Count;

m\_addressMap.Add(AssemblyCodeGenerator.ArgsName, 0);

}

We also need to add the path name of final executable file.

{ List<byte> byteList = new List<byte>();

IDictionary<int, string> accessMap = new Dictionary<int, string>();

pathNameSymbol = (StaticSymbolWindows)

ConstantExpression.Value(AssemblyCodeGenerator.PathName,

Type.StringType, targetFile.FullName);

m\_globalMap.Add(AssemblyCodeGenerator.PathName, pathNameSymbol);

}

}

We look up the main function symbol and report an error if it is not present. If it is present, we add it after the initiation symbol and potential command line symbol. Then we call GenerateTrace, which traces all symbols reachable from the main function, and adds them to m\_globalList and m\_addressMap.

StaticSymbolWindows mainInfo;

Assert.Error(m\_globalMap.TryGetValue(AssemblyCodeGenerator.MainName,

out mainInfo), "non-static main",

Message.Function\_missing);

GenerateTrace(mainInfo);

In case of command line arguments, we add the symbol for the executable file path name to m\_globalList and m\_addressMap.

if (pathNameSymbol != null) {

Assert.ErrorXXX(!m\_globalList.Contains(pathNameSymbol));

m\_globalList.Add(pathNameSymbol);

m\_addressMap.Add(pathNameSymbol.UniqueName, m\_totalSize);

m\_totalSize += (int) pathNameSymbol.ByteList.Count;

}

Finally, we add the stack top name, which will be the address of the activation record of the main function.

m\_addressMap.Add(AssemblyCodeGenerator.CallStackStart, m\_totalSize);

We iterate through the global list and, for each of its symbols, we call GenerateAccess, GenerateCall, GenerateReturn, which replace the names of accessed static symbols and called functions as well as return addresses with proper addresses.

foreach (StaticSymbolWindows staticSymbol in m\_globalList) {

List<byte> byteList = staticSymbol.ByteList;

int startAddress = m\_addressMap[staticSymbol.UniqueName];

GenerateAccess(staticSymbol.AccessMap, byteList);

GenerateCall(startAddress, staticSymbol.CallMap, byteList);

GenerateReturn(startAddress, staticSymbol.ReturnSet, byteList);

}

Finally, we write the byte list of each symbol in the global list to the target file.

{ Console.Out.WriteLine("Generating \"" + targetFile.FullName + "\".");

targetFile.Delete();

BinaryWriter targetStream =

new BinaryWriter(File.OpenWrite(targetFile.FullName));

foreach (StaticSymbolWindows staticSymbol in m\_globalList) {

foreach (sbyte b in staticSymbol.ByteList) {

targetStream.Write(b);

}

}

targetStream.Close();

}

}

The GenerateTrace method adds the symbol to the m\_globalList iterates through the names of access name set and function call name set and calls GenerateTrace recursively for each symbol not yet present in the m\_globalList.

private void GenerateTrace(StaticSymbolWindows staticSymbol) {

if (!m\_globalList.Contains(staticSymbol)) {

m\_globalList.Add(staticSymbol);

m\_addressMap.Add(staticSymbol.UniqueName, m\_totalSize);

m\_totalSize += (int) staticSymbol.ByteList.Count;

ISet<string> accessNameSet =

new HashSet<string>(staticSymbol.AccessMap.Values);

foreach (string accessName in accessNameSet) {

StaticSymbolWindows accessSymbol;

Assert.Error(m\_globalMap.TryGetValue(accessName, out accessSymbol),

accessName, Message.Object\_missing\_in\_linking);

Assert.ErrorXXX(accessSymbol != null);

GenerateTrace(accessSymbol);

}

ISet<string> callNameSet =

new HashSet<string>(staticSymbol.CallMap.Values);

foreach (string callName in callNameSet) {

StaticSymbolWindows funcSymbol;

Assert.Error(m\_globalMap.TryGetValue(callName, out funcSymbol),

callName, Message.Function\_missing\_in\_linking);

Assert.Error(funcSymbol != null, SimpleName(callName),

Message.Missing\_external\_function);

GenerateTrace(funcSymbol);

}

}

}

The GenerateAccess methods iterators through the access map and modifies the addresses.

For each element, reachable from main, we need to modify its static accesses, function calls and return assignments (for a global variable the call map and return set are empty) by calling GenerateAccess, GenerateCall, and GenerateReturn.

private void GenerateAccess(IDictionary<int,string> accessMap,

List<byte> byteList) {

foreach (KeyValuePair<int,string> entry in accessMap) {

int sourceAddress = entry.Key;

string targetName = entry.Value;

We obtain the target address from the low and high byte of the source address. Note that the target address does not need to be zero, it may hold an offset. For instance, the target address may be a pointer to a value in an array of a non-zero index, in which case its offset is non-zero.

byte lowByte = byteList[sourceAddress],

highByte = byteList[sourceAddress + 1];

int targetAddress = ((int) highByte << 8) + lowByte;

We modify the target address by adding the address of the name to target address, which we restore to the source address.

targetAddress += m\_addressMap[targetName];

byteList[sourceAddress] = (byte) targetAddress;

byteList[sourceAddress + 1] = (byte) (targetAddress >> 8);

}

}

The GenerateCall method iterates through the names of the functions to called. There are two differences between this method and GenerateAccess above. First, we do not need to take inspect the original address of the function. It is always zero since it is not possible to add an offset to a function call. Second, the address of the call is not absolute, it is relative the next assembly code instruction.

We could change the calls where the address is eight bytes from long jump to short jumps. However, this may cause local jumps in the function to be changed from short jumps to long jumps.

private void GenerateCall(int callerStartAddress,

IDictionary<int,string> callMap,

List<byte> byteList) {

foreach (KeyValuePair<int,string> entry in callMap) {

int sourceAddress = entry.Key;

string sourceName = entry.Value;

The source address is the last two bytes of the function assembly code instruction, why the address of the next instruction is the source address plus two. However, the source address is relative the beginning of the function. Therefore, we need to add the global start address of the caller function to the caller address, in order to make the address relative the whole executable file.

int nextAddress = (sourceAddress + 2) + callerStartAddress;

int calleeAddress = m\_addressMap[sourceName];

The relative address of the call is the difference of the address of the callee function and the next instruction (the address of the instruction following the call).

int relativeAddress = calleeAddress - nextAddress;

byteList[sourceAddress] = (byte) ((sbyte) relativeAddress);

byteList[sourceAddress + 1] = (byte) ((sbyte) (relativeAddress >> 8));

}

}

The GenerateReturn method changes the return address from the address relative the beginning of the function to the address relative the beginning of the executable code. Similar to the GenerateAccess case above, we need to inspect the original address, and add the address of the beginning function.

private void GenerateReturn(int functionStartAddress, ISet<int> returnSet,

List<byte> byteList) {

foreach (int sourceAddress in returnSet) {

int lowByte = byteList[sourceAddress],

highByte = byteList[sourceAddress + 1];

int targetAddress = (highByte << 8) + lowByte;

targetAddress += functionStartAddress;

byteList[sourceAddress] = (byte) targetAddress;

byteList[sourceAddress + 1] = (byte) (targetAddress >> 8);

}

}

The SimpleName method returns the part to the left of the leftmost separator identity characters (‘$’), if present. Otherwise, the whole string is returned.

public static string SimpleName(string name) {

int index = name.LastIndexOf(Symbol.SeparatorId);

return (index != -1) ? name.Substring(0, index) : name;

}

}

}

1. The Standard Library

In the limits, setjmp, stdio, stdlib, and time standard libraries there is different C code for the Linux and Windows environment. We use the \_\_LINUX\_\_ and \_\_WINDOWS\_\_ macros to distinguish between them.

## Integral and Floating Limits

The file limits.h holds limits for integral values for the Linux and the Windows environment.

limits.h

#ifndef \_\_LIMITS\_H\_\_

#define \_\_LIMITS\_H\_\_

#ifdef \_\_WINDOWS\_\_

#define CHAR\_BIT 8

#define CHAR\_MIN -128S

#define CHAR\_MAX 127S

#define UCHAR\_MAX 255US

#define SHRT\_MIN -128S

#define SHRT\_MAX 127S

#define USHRT\_MAX 255US

#define INT\_MIN -32768

#define INT\_MAX 32767

#define UINT\_MAX 65535U

#define LONG\_MIN -2147483648L

#define LONG\_MAX 2147483647L

#define ULONG\_MAX 4294967295UL

#endif

#ifdef \_\_LINUX\_\_

#define CHAR\_BIT 8

#define CHAR\_MIN -128S

#define CHAR\_MAX 127S

#define UCHAR\_MAX 255US

#define SHRT\_MIN -32768S

#define SHRT\_MAX 32767S

#define USHRT\_MAX 65535US

#define INT\_MIN -2147483648

#define INT\_MAX 2147483647

#define UINT\_MAX 4294967295U

#define LONG\_MIN -9223372036854775808L

#define LONG\_MAX 9223372036854775807L

#define ULONG\_MAX 0xFFFFFFFFFFFFFFFFUL

#define ULONG\_MAX 18446744073709551615UL

#endif

#endif

The file float.h holds limits for floating-point values. Note that there is no difference between the Linux and the Windows environment.

float.h

#ifndef \_\_FLOAT\_H\_\_

#define \_\_FLOAT\_H\_\_

#define FLT\_ROUNDS 1

#define FLT\_RADIX 2

#define FLT\_MANT\_DIG 2

#define DBL\_MANT\_DIG 2

#define LDBL\_MANT\_DIG 2

#define FLT\_DIG 6

#define FLT\_EPSILON 1e-6

#define FLT\_MIN\_EXP -38

#define FLT\_MIN 1.2E-38

#define FLT\_MAX\_EXP 38

#define FLT\_MAX 3.4E+38

#define DBL\_DIG 6

#define DBL\_EPSILON 1e-6

#define DBL\_MANT\_DIG 2

#define DBL\_MIN\_EXP -308

#define DBL\_MIN 2.3E-308

#define DBL\_MAX\_EXP 308

#define DBL\_MAX 1.7E+308

#endif

## The assert Macro

The assert macro aborts the execution with an error message if the expression is false. We use the \_\_FILE\_\_ and \_\_LINE\_\_ predefined macros to obtain the current file name and line number. The user can explicitly set the NDEBUG macro to shorten the definition of the assert macro.

assert.h

#ifndef \_\_ASSERT\_H\_\_

#define \_\_ASSERT\_H\_\_

#ifndef NDEBUG

#include <stdio.h>

#include <stdlib.h>

#define assert(expression) if (!(expression)) { \

fprintf(stderr, "Assertion failed: \"%s\" in file %s at line %i\n", \

#expression, \_\_FILE\_\_, \_\_LINE\_\_); abort(); }

#else

#define assert(expression)

#endif

#endif

## Locale Data

The file locale library holds functions for locale data.

locale.h

#ifndef \_\_LOCALE\_H\_\_

#define \_\_LOCALE\_H\_\_

#define LC\_COLLATE 0x01

#define LC\_CTYPE 0x02

#define LC\_MONETARY 0x04

#define LC\_NUMERIC 0x08

#define LC\_TIME 0x10

#define LC\_ALL 0x1F

The lconv struct holds the information of the locale environment. The summerTimeZone and winter-TimeZone holds the number of hours beyond the Greenwich time for the local summer and winter time zones. To decide whether there actually is winter or summer time at the current time of the year, we need the functions of the time standard library.

struct lconv {

int summerTimeZone, winterTimeZone;

The shortDayList, longDaysList, shortMonthList, and longMonthList fields hold the names of the days of the week and the names of the months of the year, abbreviated and in full text.

char \*\*shortDayList;

char \*\*longDayList;

char \*\*shortMonthList;

char \*\*longMonthList;

The lowerCase and upperCase fields hold the alphabet of the current language.

char \*lowerCase;

char \*upperCase;

The messageList field holds the error message on the current language.

char \*\*messageList;

};

The error message enumeration constants are assigned the values from zero and upwards, which corresponds to the indexes of the enMessageList array in the locale.c file.

extern enum {NO\_ERROR, NO\_FUNCTION, NO\_FILE, NO\_PATH, NO\_HANDLE, NO\_ACCESS,

EDOM, ERANGE, EILSEQ, FOPEN, FFLUSH, FCLOSE, NO\_MODE, FWRITE,

FREAD, FSEEK, FTELL, FSIZE, FREMOVE,FRENAME, FTEMPNAME,FTEMPFILE};

extern char\* setlocale(int flag, char\* name);

extern struct lconv \*localeconv(void);

#endif

locale.c

#include <stdio.h>

#include <stdlib.h>

#include <string.h>

#include <locale.h>

We initialize the enShortDayList, enLongDayList, enShortMonthList, and enLongMonthList with the English names of the days of the week and the months of the year.

static char\* enShortDayList[] = {"Sun", "Mon", "Tue", "Wed",

"Thu", "Fri", "Sat"};

static char\* enLongDayList[] ={"Sunday", "Monday", "Tuesday", "Wednesday",

"Thursday", "Friday", "Saturday"};

static char\* enShortMonthList[] = {"Jan", "Feb", "Mar", "Apr", "May", "Jun",

"Jul", "Aug", "Sep", "Oct", "Nov", "Dec"};

static char\* enLongMonthList[] = {"January", "February", "March", "April",

"May", "June", "July", "August",

"September", "October", "November", "December"};

The indexes of the error messages of the enMessageList array correspond to the enumeration values of the locale.h file.

char\* enMessageList[] = {"no error", "function not found",

"file not found", "path not found",

"no handle available", "access denied",

"out of domain", "out of range",

"invalid multibyte sequence", "error while opening",

"error while flushing", "error while closing",

"open mode invalid", "error while writing",

"error while reading", "error while seeking",

"error while telling", "error while sizing",

"error while removing file",

"error while renaming file" };

We initialize the en\_US\_utf8 structure with the information of the English language and the time setting of the American east coast.

static struct lconv en\_US\_utf8 = {-5, -4, enShortDayList, enLongDayList,

enShortMonthList, enLongMonthList,

"abcdefghijklmnopqrstuvwxyz",

"ABCDEFGHIJKLMNOPQRSTUVWXYZ",

enMessageList};

static const struct \_s {

char\* name;

struct lconv\* localePtr;

} g\_sArray[] = {{"", &en\_US\_utf8}, {"C", &en\_US\_utf8}, {"US", &en\_US\_utf8}};

The g\_sSize constant holds the size of the g\_sArray array, which we calculate by dividing the total size of the array by the size of its first value.

static const int g\_sSize = (sizeof g\_sArray) / (sizeof g\_sArray[0]);

We let the static pointer g\_currStructPtr point at the first value of g\_sArray, we assume that there is always at least one value in the array.

static struct \_s\* g\_currStructPtr = &g\_sArray[0];

The setLocale function look up the locale settings with the given name, which is later returned by localeconv below. The function returns the name of the previous name settings, or null if there is no settings.

char\* setlocale(int /\*flag\*/, char\* newName) {

int index;

char \*oldName = (g\_currStructPtr != NULL) ? g\_currStructPtr->name : NULL;

g\_currStructPtr = NULL;

if (newName != NULL) {

for (index = 0; index < g\_sSize; ++index) {

if (strcmp(newName, g\_sArray[index].name) == 0) {

g\_currStructPtr = &g\_sArray[index];

break;

}

}

}

return oldName;

}

The localeconv function returns the current locale settings.

struct lconv\* localeconv(void) {

return (g\_currStructPtr != NULL) ? g\_currStructPtr->localePtr : NULL;

}

## Character Types

The ctype.h and ctype.c files hold a set functions that test different character conditions.

ctype.h

#ifndef \_\_CTYPE\_H\_\_

#define \_\_CTYPE\_H\_\_

extern int islower(int c);

extern int isupper(int c);

extern int isalpha(int c);

extern int isdigit(int c);

extern int isalnum(int c);

extern int isxdigit(int c);

extern int isgraph(int c);

extern int isprint(int c);

extern int ispunct(int c);

extern int iscntrl(int c);

extern int isspace(int c);

extern int tolower(int c);

extern int toupper(int c);

#endif

ctype.c

#include <ctype.h>

#include <stdio.h>

#include <locale.h>

#include <string.h>

#include <stddef.h>

The islower function tests whether the character is a lowercase character. If there are locale settings defined (see Section 13.5) available, we use it. Otherwise, we regard a character to be lowercase if its ASCII value is between the ASCII values for the ‘a’ and ‘z’ characters, inclusive.

int islower(int c) {

struct lconv\* localeConvPtr = localeconv();

if (localeConvPtr != NULL) {

return (strchr(localeConvPtr->lowerCase, c) != NULL);

}

else {

return ((c >= 'a') && (c <= 'z'));

}

}

The isupper function tests whether the character is an uppercase character in the same way as islower above. If there are locale settings available, we use it. Otherwise, we regard a character to be lowercase if its ASCII value is between the ASCII values for the ‘A’ and ‘Z’ character, inclusive.

int isupper(int c) {

struct lconv\* localeConvPtr = localeconv();

if (localeConvPtr != NULL) {

return (strchr(localeConvPtr->upperCase, c) != NULL);

}

else {

return ((c >= 'A') && (c <= 'Z'));

}

}

The isalpha function tests whether the character is a letter; that is, a lowercase or an uppercase character.

int isalpha(int c) {

return islower(c) || isupper(c);

}

The isdigit function tests whether the character is a digit. In this case, we do not use a potential locale convention; we just test whether the ASCII value for the character is between the ASCII values of the ‘0’ and ‘9’ characters, inclusive. Note that we test against the ASCII value of the ‘0’ character, which is not the same as the ASCIII value of the zero-character (‘\0’), which is zero.

int isdigit(int c) {

return (c >= '0') && (c <= '9');

}

The isalnum function tests whether the character is an alphanumerical character; that is, a lowercase or an uppercase character, or a digit.

int isalnum(int c) {

return isalpha(c) || isdigit(c);

}

The isxdigit function tests whether the character is a hexadecimal digit. Both lowercase and uppercase letters are allowed.

int isxdigit(int c) {

return isdigit(c) || ((c >= 'a') && (c <= 'f'))

|| ((c >= 'A') && (c <= 'F'));

}

The isgraph function tests whether the character is a graphical character; that is, a character visible when printed (including space).

int isgraph(int c) {

return (c >= 32) && (c <= 126);

}

The isprint function tests whether the character is a printable character; that is, a character visible when printed (excluding space).

int isprint(int c) {

return isgraph(c) && (c != ' ');

}

The ispunct function tests whether the character is a punctation character; that is, a character that is graphical but no alphanumerical, resulting is all visible characters that are not letters or digits.

int ispunct(int c) {

return isgraph(c) && !isalnum(c);

}

The iscntrl function tests whether the character is a control character, which is all non-printable characters. Note that space is a control character.

int iscntrl(int c) {

return !isprint(c);

}

The isspace function tests whether the character is a white-space character; that is, a space, form feed, new line, return, horizontal tabulator, or vertical tabulator.

int isspace(int c) {

return (c == ' ') || (c == '\f') || (c == '\n') ||

(c == '\r') || (c == '\t') || (c == '\v');

}

The tolower function returns the character converted to uppercase if it is a lowercase character. We use a locale conversion, is available, to locate the lowercase and uppercase character strings.

int tolower(int c) {

if (isupper(c)) {

struct lconv\* localeConvPtr = localeconv();

We obtain the strings of lowercase and uppercase characters from the locale conversion, look up the index of the character by in the uppercase string by calling strchr, and return the lowercase character at the same index in the lowercase string.

if (localeConvPtr != NULL) {

char \*lowerCase = localeConvPtr->lowerCase,

\*upperCase = localeConvPtr->upperCase;

int index = (strchr(upperCase, c) - upperCase);

return ((int) lowerCase[index]);

}

If there is no locale convention available, we return the character with ASCII value plus 32, since the lowercase value are located on values 32 above the uppercase values in the ASCII table.

else {

return (c + 32);

}

}

If the character is not an uppercase character, we simply return it.

else {

return c;

}

}

The toupper function returns the character converted to lowercase if it is an uppercase character in the same way as tolower above. Again, we use a locale conversion, is available, to locate the lowercase and uppercase character string, look up the index of character in the lowercase string and return the character at the same index in the uppercase string.

int toupper(int c) {

if (islower(c)) {

struct lconv\* localeConvPtr = localeconv();

if (localeConvPtr != NULL) {

char \*lowerCase = localeConvPtr->lowerCase,

\*upperCase = localeConvPtr->upperCase;

int index = (strchr(lowerCase, c) - lowerCase);

return ((int) upperCase[index]);

}

If there is no locale convention available, we return the character with ASCII value minus 32, since the uppercase value are located on values 32 below the lowercase values in the ASCII table.

else {

return (c - 32);

}

}

If the character is not a lowercase character, we simply return it.

else {

return c;

}

}

## Strings

The string standard library holds function for searching and modifying strings.

string.h

#ifndef \_\_STRING\_H\_\_

#define \_\_STRING\_H\_\_

#define size\_t int

extern char\* strcpy(char\* target, const char\* source);

extern char\* strncpy(char\* target, const char\* source, size\_t size);

extern char\* strcat(char\* target, const char\* source);

extern char\* strncat(char\* target, const char\* source, size\_t size);

extern int strcmp(const char\* left, const char\* right);

extern int strncmp(const char\* left, const char\* right, size\_t size);

extern char\* strchr(const char\* text, int i);

extern char\* strrchr(const char\* text, int i);

extern size\_t strspn(const char\* mainString, const char\* charSet);

extern size\_t strcspn(const char\* mainString, const char\* charSet);

extern char\* strpbrk(const char\* mainString, const char\* charSet);

extern char\* strstr(const char\* mainString, const char\* subString);

extern size\_t strlen(const char\* string);

extern char\* strerror(int error);

extern char\* strtok(char\* string, const char\* charSet);

extern char\* memcpy(char\* target, const char\* source, size\_t size);

extern char\* memmove(char\* target, const char\* source, size\_t size);

extern int memcmp(const char\* left, const char\* right, size\_t size);

extern char\* memchr(const char\* block, int i, size\_t size);

extern char\* memset(char\* block, int i, size\_t size);

#endif

string.c

#include <errno.h>

#include <stdio.h>

#include <stddef.h>

#include <string.h>

#include <locale.h>

The strcpy function copies the text of the source parameter to the target parameter.

char\* strcpy(char\* target, const char\* source) {

int index;

We iterate throught the source text and copies the characters, until we reach the terminating zero character (‘\0’).

for (index = 0; source[index] != '\0'; ++index) {

target[index] = source[index];

}

We add the terminating zero-character to the target string, and return it.

target[index] = '\0';

return target;

}

The strcpy function copies the text of the source parameter to the target parameter, with size as the maximum number of copied characters.

char\* strncpy(char\* target, const char\* source, size\_t size) {

int index;

for (index = 0; (index < size) && (source[index] != '\0'); ++index) {

target[index] = source[index];

}

We add the terminating zero character only if there is room in the target string; that is, if we have not already copied size characters.

if (index < size) {

target[index] = '\0';

}

return target;

}

The strcat function adds the source string to the target string.

char\* strcat(char\* target, const char\* source) {

int index, targetLength = strlen(target);

for (index = 0; source[index] != '\0'; ++index) {

target[targetLength + index] = source[index];

}

We add the terminating zero-character at the end of the target string.

target[targetLength + index] = '\0';

return target;

}

The strcpy function copies the text of the source parameter to the target parameter, with size as the maximum number of copied characters.

char\* strncat(char\* target, const char\* source, size\_t size) {

int index;

const int targetLength = strlen(target);

for (index = 0; (index < size) && (source[index] != '\0'); ++index) {

target[targetLength + index] = source[index];

}

if ((targetLength + index) < size) {

target[targetLength + index] = '\0';

}

return target;

}

The strcmp function compares the left and right string. It returns minus oner if the left string is smaller than the right string, zero if they are equal, and one if the right string is smaller than the left string.

int strcmp(const char\* left, const char\* right) {

int index;

The only way the regard the two string as equal is if they both reach the terminating zero-character at the same time.

for (index = 0; TRUE; ++index) {

if ((left[index] == '\0') && (right[index] == '\0')) {

return 0;

}

If one of the strings reaches the zero-character first, it is the smaller string. Technically, we could omit these two if statements, since the zero character has ASCII value zero and thereby less than all other characters.

else if (left[index] == '\0') {

return -1;

}

else if (right[index] == '\0') {

return 1;

}

Until one or both strings has reached the zero-character, we compare the current characters. If the left character is smaller the right character, the left string is smaller. If the right character is smaller the left character, the right string is smaller. When comparing two characters, their ASCII values are compared.

else if (left[index] < right[index]) {

return -1;

}

else if (left[index] > right[index]) {

return 1;

}

When we have reached this point, where we have not yet reached the end of either of the string, and they have so far been equal, we just iterate afresh and compare the next characters of the strings.

}

}

The strcmp function compares the left and right string as strcmp above. The difference is that it compares up until size number of characters. If the first size number of characters are equal, the strings are regarded equal.

int strncmp(const char\* left, const char\* right, size\_t size) {

int index;

for (index = 0; index < size; ++index) {

if ((left[index] == '\0') && (right[index] == '\0')) {

return 0;

}

else if (left[index] == '\0') {

return -1;

}

else if (right[index] == '\0') {

return 1;

}

else if (left[index] < right[index]) {

return -1;

}

else if (left[index] > right[index]) {

return 1;

}

}

If we have iterated size number of times without finding any difference between the characters, the strings are regarded equal, and we return zero.

return 0;

}

The strchr function looks up the character and returns the address its first occurrence in the string if it is present, and null if it is not present.

char\* strchr(const char\* text, int i) {

int index;

char c = (char) i;

for (index = 0; text[index] != '\0'; ++index) {

if (text[index] == c) {

return &text[index];

}

}

return NULL;

}

The strrchr function looks up the character and returns the address its last occurrence in the string if it is present, and null if it is not present.

char\* strrchr(const char\* text, int i) {

int index;

char\* result = NULL;

char c = (char) i;

Note that we do not return any value inside the loop in this function. Instead we assign result the address of each occurrence. In this way, result will eventually be given the address of the last occurrence of the character, if it is present in the string.

for (index = 0; text[index] != '\0'; ++index) {

if (text[index] == c) {

result = &text[index];

}

}

As result was initialized by null, null is return if there was no occurrence of the character in the string.

return result;

}

The strspn function returns number of initial characters of the main string made up of the characters of the character set.

size\_t strspn(const char\* mainString, const char\* charSet) {

int index;

for (index = 0; mainString[index] != '\0'; ++index) {

if (strchr(charSet, mainString[index]) == NULL) {

return index;

}

}

return -1;

}

The strcspn function return the index in the string of the first occurrence of one of the characters not present in the character set, or minus one if there is no such occurrence.

size\_t strcspn(const char\* mainString, const char\* charSet) {

int index;

for (index = 0; mainString[index] != '\0'; ++index) {

if (strchr(charSet, mainString[index]) != NULL) {

return index;

}

}

return -1;

}

The strpbrk function

char\* strpbrk(const char\* mainString, const char\* charSet) {

int index;

for (index = 0; mainString[index] != '\0'; ++index) {

if (strchr(charSet, mainString[index]) != NULL) {

return &mainString[index];

}

}

return NULL;

}

The strstr function returns the index of the first occurance of the substring in the main string, or null if there is no occurrence.

char\* strstr(const char\* mainString, const char\* subString) {

int index;

const int subStringSize = strlen(subString);

We iterate through the main string and call strncmp for each index to test whether the sub stirng is indeed a sub string of the main string.

for (index = 0; mainString[index] != '\0'; ++index) {

if (strncmp(mainString + index, subString, subStringSize) == 0) {

return &mainString[index];

}

}

return NULL;

}

The strlen function returns the number of characters of the string, not counting the terminating zero-character.

size\_t strlen(const char\* string) {

int index;

We simply iterate through the string until we reach the terminating zero-character.

for (index = 0; string[index] != '\0'; ++index) {

// Empty.

}

return index;

}

The strerror function return the error message corresponding to the error value in the local settings if there are locale settings with an error message list, otherwise null.

char\* strerror(int errno) {

struct lconv\* localeConvPtr = localeconv();

if (localeConvPtr != NULL) {

char\*\* messageList = localeConvPtr->messageList;

if (messageList != NULL) {

return messageList[errno];

}

}

return NULL;

}

The strtok function divide the string in tokens, separated by any of the characters in character set. The static global variable token holds the address of the first token of the string.

static char\* token = NULL;

char\* strtok(char\* string, const char\* charSet) {

int index;

char\* tokenStart;

if (string != NULL) {

if (string[0] == '\0') {

return NULL;

}

for (index = 0; string[index] != '\0'; ++index) {

if (strchr(charSet, string[index]) != NULL) {

string[index] = '\0';

token = &string[index + 1];

return string;

}

}

token = &string[index];

return string;

}

else if (token == NULL) {

return NULL;

}

else {

if (token[0] == '\0') {

return NULL;

}

for (index = 0; token[index] != '\0'; ++index) {

if (strchr(charSet, token[index]) != NULL) {

char\* tokenStart2 = token;

token[index] = '\0';

token = &token[index + 1];

return tokenStart2;

}

}

tokenStart = token;

token = &token[index];

return tokenStart;

}

}

Finally, we have the three memory functions memcpy, memmove, memcmp, memchr, and memset. Basically, the do the same thing as their string function counterparts. However, there is no terminating zero-character. Instead, there is always a size in bytes. The memcpy function copies a memory block of a certain size (in bytes) from a source address to a target address. However, it takes no consideration in whether the memory blocks overlaps.

void\* memcpy(void\* target, const void\* source, size\_t size) {

char\* charTarget = (char\*) target;

const char\* charSource = (const char\*) source;

int index;

for (index = 0; index < size; ++index) {

charTarget[index] = charSource[index];

}

return ((void\*) target);

}

Similar to the memcpy function above, memmove copies a memory block of a certain size (in bytes) from a source address to a target address. The difference is that memmove takes consideration in whether the memory blocks overlaps.

void\* memmove(void\* target, const void\* source, size\_t size) {

char\* charTarget = (char\*) target;

const char\* charSource = (const char\*) source;

If the source address is lower than the target address, we copy the bytes in reversed order to make sure the bytes are copied in valid order in case of overlapping.

int index;

if (source < target) {

for (index = (size - 1); index >= 0; --index) {

charTarget[index] = charSource[index];

}

}

else {

for (index = 0; index < size; ++index) {

charTarget[index] = charSource[index];

}

}

return ((void\*) target);

}

The memcmp function compares the two memory block, and return minus one if the left block is the smaller one, zero if they are equal, and plus one if the right block is the smaller one.

int memcmp(const void\* left, const void\* right, size\_t size) {

const char\* charLeft = (const char\*) left;

const char\* charRight = (const char\*) right;

int index;

for (index = 0; index < size; ++index) {

if (charLeft[index] < charRight[index]) {

return -1;

}

else if (charLeft[index] > charRight[index]) {

return 1;

}

}

return 0;

}

The memchr function searches the memory block after a byte and returns the address of the first occurrence if it is present in the block, otherwise null.

void\* memchr(const void\* block, int i, size\_t size) {

int index;

const char\* charBlock = (const char\*) block;

char c = (char) i;

for (index = 0; index < size; ++index) {

if (charBlock[index] == c) {

return (void\*) &charBlock[index];

}

}

return NULL;

}

The memset function sets the memory block with the value.

void\* memset(void\* block, int i, size\_t size) {

char\* charBlock = (char\*) block;

char c = (char) i;

int index;

for (index = 0; index < size; ++index) {

charBlock[index] = c;

}

return block;

}

## Long Jumps

The setjmp.h and setjmp.c files hold functions for preparing and performing long jumps; that is, jumps through a function call chain.

setjmp.h

typedef void\* jmp\_buf[3];

int setjmp(jmp\_buf env);

void longjmp(jmp\_buf env, int value);

setjmp.c

#include <setjmp.h>

The setjmp function stores the return address, regular frame pointer, and variadic frame pointer in the buf static variable.

#ifdef \_\_LINUX\_\_

int setjmp(jmp\_buf buf) {

void\*\* rbp\_pointer = register\_rbp;

buf[0] = rbp\_pointer[0];

buf[1] = rbp\_pointer[1];

buf[2] = rbp\_pointer[2];

return 0;

}

The longjmp function resets the return address, regular frame pointer, and variadic frame pointer by looking up their values from the buf static variable by setjmp above. The last line jumps back to the return address of the previous setjmp call, with the return value given as a parameter to longjmp. In this way the previous setjmp call return zero and the following longjmp call returns the (presumably non-zero) value of its parameter.

void longjmp(jmp\_buf buf, int return\_value) {

register\_ebx = return\_value;

register\_rcx = buf[0];

register\_rdi = buf[2];

register\_rbp = buf[1];

jump\_register(register\_rcx);

}

#endif

The code for the Windows environment is similar to the Linux environment. The difference is that we use 16-bit registers rather than 32-bit and 64-bit registers.

#ifdef \_\_WINDOWS\_\_

int setjmp(jmp\_buf buf) {

void\*\* bp\_pointer = register\_bp;

buf[0] = bp\_pointer[0];

buf[1] = bp\_pointer[1];

buf[2] = bp\_pointer[2];

return 0;

}

void longjmp(jmp\_buf buf, int return\_value) {

register\_bx = return\_value;

register\_cx = buf[0];

register\_di = buf[2];

register\_bp = buf[1];

jump\_register(register\_cx);

}

#endif

The following code demonstrates how setjmp and longjmp can be used by jumping back through a function call chain.

setjmptest.c

#include <stdio.h>

#include <setjmp.h>

jmp\_buf buffer;

double inverse(double x);

double divide(double x, double y);

void main() {

char\* message;

double x;

printf("Please input a value: ");

scanf("%lf", &x);

if ((message = setjmp(buffer)) == 0) {

printf("1.0 / %f = %f\n", x, inverse(x));

}

else {

printf("%s\n", message);

}

}

double inverse(double x) {

return divide(1, x);

}

In case of a non-zero denominator, we return the quitent in a normal way and setjmp returns zero. In case of a zero denominator, we call longjmp with an error message. The address of the error message is then be returned by the call to setjmp, instead of zero.

double divide(double x, double y) {

if (y != 0) {

return x / y;

}

else {

longjmp(buffer, "Division by Zero.");

return 0;

}

}

## Mathematical Functions

The mathematical functions are implemented completely in C, they do not make system calls.

math.h

#ifndef \_\_MATH\_H\_\_

#define \_\_MATH\_H\_\_

#define PI 3.1415926535897932384626433

#define E 2.7182818284590452353602874

extern double exp(double x);

extern double log(double x);

extern double log10(double x);

extern double pow(double x, double y);

extern double ldexp(double x, int exponent);

extern double frexp(double x, int\* exponent);

extern double sqrt(double x);

extern double modf(double x, double\* integral);

extern double fmod(double x, double y);

extern double sin(double x);

extern double cos(double x);

extern double tan(double x);

extern double sinh(double x);

extern double cosh(double x);

extern double tanh(double x);

extern double asin(double x);

extern double acos(double x);

extern double atan(double x);

extern double atan2(double x, double y);

extern double floor(double x);

extern double ceil(double x);

extern double round(double x);

extern double fabs(double x);

#endif

### Exponent and Logarithm Functions

The Sine, Cosine, Exponent, and Logarithm functions are calculated by iterative methods, and the rest of the functions call these functions.

math.c

#include <math.h>

#include <errno.h>

#include <stdio.h>

#include <stddef.h>

#include <stdlib.h>

The exponent function can be calculated be the following iterative formula for any values of .

Naturally, we cannot iterate indefinitely in the code. Instead, we break the iteration when the term is smaller than the EPSILON constant.

#define EPSILON 1e-9

double exp(double x) {

double index = 1, term, sum = 1, faculty = 1, power = x;

do {

term = power / faculty;

sum += term;

power \*= x;

faculty \*= ++index;

} while (fabs(term) >= EPSILON);

return sum;

}

The logarithm function can be calculated be the following formula for any values , which gives .

To begin with, we must make sure that the input value is more than zero and less than two. Moreover, in order to keep the number of iterations down, the value shall not be too close to either zero or two. Therefore, we make sure the value holds the interval . We can do that by multiply or divide by , and use the facts that

and

This gives that we can divide with until , and we can multiply with until , as long as we keep track of the number of divisions and multiplications.

#define E\_INVERSE (1 / E)

double log(double x) {

if (x > 0) {

int n = 0;

If is more than zero and not exact one, we test whether x is more than one. If it is, we divide with until is less than one.

if (x > 1) {

while (x > 1) {

x /= E;

++n;

}

}

If is less than , we multiple with until it is more than more . In both cases we keep track of the number of divisions and multiplications by updating .

else if (x < E\_INVERSE) {

while (x < E\_INVERSE) {

x \*= E;

--n;

}

}

When we have made sure that , we iterate with the formula .

{ double index = 1, term, sum = 0, sign = 1,

x\_minus\_1 = x - 1, power = x\_minus\_1;

do {

term = sign \* power / index++;

sum += term;

power \*= x\_minus\_1;

sign \*= -1.0;

} while (fabs(term) >= EPSILON);

The result is the sum of the iteration plus the number of division and multiplication () before the iteration.

return sum + n;

}

}

If is less than zero, we report an error by setting errno domain error and return zero.

else {

errno = EDOM;

return 0;

}

}

The common logarithm, with base 10, is defined as follows:

We define a constant value for , which we divide with.

#define LN\_10 2.3025850929940456840179914

double log10(double x) {

return log(x) / LN\_10;

}

### Power Functions

The power function is defined for all real values as follows:

If and is an integer value, then

double pow(double x, double y) {

if (x > 0) {

return exp(y \* log(x));

}

else if ((x == 0) && (y == 0)) {

return 1;

}

else if ((x == 0) && (y > 0)) {

return 0;

}

We test if is an integer value by comparing its floor and ceiling values. If they are equal, the value is an integer.

else if ((x < 0) && (floor(y) == ceil(y))) {

long long\_y = (long) y;

If the integer value of modulo two is zero, it is even. Otherwise, it is odd.

if ((long\_y % 2) == 0) {

return exp(y \* log(-x));

}

else {

return -exp(y \* log(-x));

}

}

else {

errno = EDOM;

return 0;

}

}

The function ldexp is defined as .

double ldexp(double x, int n) {

return x \* pow(2, n);

}

The function frexp splits x into normalized fraction of x. The return value is within the interval from 0.5, inclusive, to 1, exclusive.

#define LN\_2 0.6931471805599453094172321

static log2(double x) {

return log(x) / LN\_2;

}

double frexp(double x, int\* p) {

if (x != 0) {

int exponent = (int) log2(fabs(x));

if (pow(2, exponent) < x) {

++exponent;

}

if (p != NULL) {

\*p = exponent;

}

return (x / pow(2, exponent));

}

else {

if (p != NULL) {

\*p = 0;

}

return 0;

}

}

### Square Root

The square root of is for all is iterative defined as:

double sqrt(double x) {

if (x >= 0) {

double root\_i, root\_i\_plus\_1 = 1;

do {

root\_i = root\_i\_plus\_1;

root\_i\_plus\_1 = (root\_i + (x / root\_i)) / 2;

} while (fabs(root\_i\_plus\_1 - root\_i) >= EPSILON);

return root\_i\_plus\_1;

}

else {

errno = EDOM;

return 0;

}

}

### Modulo Functions

The modf function split x into a integral and fractional part. The integral is returned, and the fractional part is assigned the value that p points at, if it is not null. The integral and fractional values hold the same sign as the original value.

double modf(double x, double\* p) {

double abs\_x = fabs(x),

integral = (double) ((long) abs\_x),

fractional = abs\_x - integral;

if (p != NULL) {

\*p = (x > 0) ? integral : -integral;

}

return (x > 0) ? fractional : -fractional;

}

The fmod function returns the floating-point remainder of x / y, with the same sign as x.

double fmod(double x, double y) {

if (y != 0) {

double remainder = fabs(x - (y \* ((int) (x / y))));

return (x > 0) ? remainder : -remainder;

}

If y is zero, we have division by zero, we report a domain error and return zero.

else {

errno = EDOM;

return 0;

}

}

### Trigonometric Functions

The sine function is defined for all as follows:

double sin(double x) {

if (fabs(x) > (2 \* PI)) {

x = fmod(x, 2 \* PI);

}

{ double index = 1, term, sum = 0, sign = 1, power = x, faculty = 1;

do {

term = sign \* power / faculty;

sum += term;

sign \*= -1;

power \*= x \* x;

faculty \*= ++index \* ++index;

} while (fabs(term) >= EPSILON);

return sum;

}

}

The cosine function is defined for all as follows:

double cos(double x) {

if (fabs(x) > (2 \* PI)) {

x = fmod(x, 2 \* PI);

}

{ double index = 0, term, sum = 0, sign = 1, power = 1, faculty = 1;

do {

term = sign \* power / faculty;

sum += term;

sign \*= -1;

power \*= x \* x;

faculty \*= ++index \* ++index;

} while (fabs(term) >= EPSILON);

return sum;

}

}

The tangent function is defined as for all such as .

double tan(double x) {

double cos\_of\_x = cos(x);

printf("cos(%f) = %f\n", x, cos(x));

if (cos\_of\_x != 0) {

return (sin(x) / cos\_of\_x);

}

else {

errno = EDOM;

return 0;

}

}

### Inverted Trigonometric Functions

The arcsine function is defined for all such as as follows:

double asin(double x) {

if (x == 1) {

return PI / 2;

}

else if (x < 0) {

return -asin(-x);

}

else if (x < 1) {

return atan(x / sqrt(1 - (x \* x)));

}

else {

errno = EDOM;

return 0;

}

}

The arccosine function is defined for all such as as follows:

double acos(double x) {

if (x == 0) {

return PI / 2;

}

else if (x < 0) {

return PI - acos(-x);

}

else if (x <= 1) {

return atan(sqrt(1 - (x \* x)) / x);

}

else {

errno = EDOM;

return 0;

}

}

The arctangent function is defined for all as follows:

double atan(double x) {

if (x < 0) {

return -atan(-x);

}

else if (x > 1) {

return PI / 2 - atan(1 / x);

}

The number of iterations of the formula above is very high for values close to one. Therefore, we use the following formula for The function argument of the right-hand call is always less than or equal to , since the numerator is at most one and the denominator is at least two.

else if (x > 0.5) {

return 2 \* atan(x / (1 + sqrt(1 + (x \* x))));

}

We have finally reached the iterative formula. There is at most 30 iterations for values close to , and less iteration for values closer to zero.

else {

double term, sum = 0, sign = 1, denominator = 1, product = x;

do {

term = sign \* product / denominator;

sum += term;

sign = -sign;

product \*= x \* x;

denominator += 2;

} while (fabs(term) >= EPSILON);

return sum;

}

}

The second arctangent function return arctangent of , if .

double atan2(double x, double y) {

if (y > 0) {

return atan(x / y);

}

else if ((x >= 0) && (y < 0)) {

return PI + atan(x / y);

}

else if ((x < 0) && (y < 0)) {

return (-PI) + atan(x / y);

}

else if ((x > 0) && (y == 0)) {

return PI / 2;

}

else if ((x < 0) && (y == 0)) {

return (-PI) / 2;

}

else {

errno = EDOM;

return 0;

}

}

### Hyperbolic Trigonometric Functions

The hyperbolic sine function is defined for all as follows:

double sinh(double x) {

return (exp(x) - exp(-x)) / 2;

}

The hyperbolic cosine function is defined as follows:

double cosh(double x) {

return (exp(x) + exp(-x)) / 2;

}

The hyperbolic tangent function is defined for all as follows:

Not that we do not have to check for division by zero in this case, since cannot be zero.

double tanh(double x) {

return sinh(x) / cosh(x);

}

### Floor, Ceiling, Absolute, and Rounding Functions

The floor function returns the integer value closer to zero.

double floor(double x) {

if (x < 0) {

return -ceil(-x);

}

return (double) ((long) x);

}

double ceil(double x) {

if (x < 0) {

return -floor(-x);

}

return (double) ((long) (x + 0.999999999999));

}

double round(double x) {

return (double) ((long) ((x < 0) ? (x - 0.5) : (x + 0.5)));

}

double fabs(double x) {

return (x < 0) ? -x : x;

}

## Standard Input/Output

stdio.h

#ifndef \_\_STDIO\_H\_\_

#define \_\_STDIO\_H\_\_

#include <math.h>

#include <ctype.h>

#include <stdarg.h>

#include <stddef.h>

#include <file.h>

#include <temp.h>

#include <scanf.h>

#include <printf.h>

#endif

### Printing

printf.h

#ifndef \_\_PRINTF\_H\_\_

#define \_\_PRINTF\_H\_\_

#define DEVICE 0

#define STRING 1

extern int g\_outStatus, g\_charCount;

extern void\* g\_outDevice;

int putc(int c, FILE\* stream);

int fputc(int c, FILE\* stream);

int putchar(int c);

int printf(const char\* format, ...);

int vprintf(const char\* format, va\_list arg\_list);

int fprintf(FILE\* outStream, const char\* format, ...);

int vfprintf(FILE\* outStream, const char\* format, va\_list arg\_list);

int sprintf(char\* outString, const char\* format, ...);

int vsprintf(char\* outString, const char\* format, va\_list arg\_list);

#endif

printf.c

#include <math.h>

#include <ctype.h>

#include <stdio.h>

#include <stddef.h>

#include <stdarg.h>

#include <stdlib.h>

#include <scanf.h>

#include <printf.h>

#define DEVICE 0

#define STRING 1

#define BLANK 2

#define PRINT(x,y) { printf(#x " = <%" #y ">\n", (x)); }

int g\_outStatus, g\_outChars;

void\* g\_outDevice;

#define DEFAULT\_PRECISION 6

int putc(int i, FILE\* stream) {

g\_outStatus = DEVICE;

g\_outDevice = (void\*) stream;

printChar((char) i);

return 1;

}

int fputc(int i, FILE\* stream) {

g\_outStatus = DEVICE;

g\_outDevice = (void\*) stream;

printChar((char) i);

return 1;

}

int putchar(int i) {

g\_outStatus = DEVICE;

g\_outDevice = (void\*) stdout;

printChar((char) i);

return 1;

}

void printChar(char c) {

int handle;

char\* outString;

switch (g\_outStatus) {

case DEVICE: {

FILE\* stream = (FILE\*) g\_outDevice;

#ifdef \_\_WINDOWS\_\_

register\_ah = 0x40;

register\_bx = stream->handle;

register\_cx = 1;

register\_dx = &c;

interrupt(0x21s);

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 0x01;

register\_rdi = (unsigned long) stream->handle;

register\_rsi = (unsigned long) &c;

register\_rdx = 1L;

syscall();

#endif

++g\_outChars;

break;

}

case STRING: {

outString = (char\*) g\_outDevice;

outString[g\_outChars++] = c;

}

break;

case BLANK:

g\_outChars++;

break;

}

}

void printString(const char\* s, int precision) {

if (s != NULL) {

int index;

if (precision == 0) {

for (index = 0; s[index] != '\0'; ++index) {

printChar(s[index]);

}

}

else {

for (index = 0; (precision-- > 0) && (s[index] != '\0'); ++index) {

printChar(s[index]);

}

}

}

else {

printChar('<');

printChar('N');

printChar('U');

printChar('L');

printChar('L');

printChar('>');

}

}

void printLongIntRec(long longValue) {

if (longValue != 0) {

int digit = (int) (longValue % 10L);

printLongIntRec(longValue / 10L);

printChar((char)(digit + '0'));

}

}

void printLongInt(long longValue, BOOL plus, BOOL space) {

if (longValue < 0L) {

longValue = -longValue;

printChar('-');

}

else if (space) {

printChar(' ');

}

else if (plus) {

printChar('+');

}

if (longValue == 0L) {

printChar('0');

}

else {

printLongIntRec(longValue);

}

}

char digitToChar(int digit, BOOL capital) {

if (digit < 10) {

return ((char) ('0' + digit));

}

else if (capital) {

return ((char) ('A' + (digit - 10)));

}

else {

return ((char) ('a' + (digit - 10)));

}

}

void printUnsignedLongRec(unsigned long unsignedValue,

unsigned long base, BOOL capital) {

if (unsignedValue > 0ul) {

int digit = (int) (unsignedValue % base);

printUnsignedLongRec(unsignedValue / base, base, capital);

{ char c = digitToChar(digit, capital);

printChar(c);

}

}

}

void printUnsignedLong(unsigned long unsignedValue, BOOL plus, BOOL space,

BOOL grid, unsigned long base, BOOL capital) {

if (plus) {

printChar('+');

}

if (space) {

printChar(' ');

}

if (grid) {

if (base == 8ul) {

printChar('0');

}

if (base == 16ul) {

printChar('0');

printChar(capital ? 'X' : 'x');

}

}

if (unsignedValue == 0ul) {

printChar('0');

}

else {

printUnsignedLongRec(unsignedValue, base, capital);

}

}

void printLongDoubleFraction(long double longDoubleValue,

BOOL grid, int precision) {

longDoubleValue -= (long) longDoubleValue;

if (precision == 0) {

precision = DEFAULT\_PRECISION;

}

if (grid || (precision > 0)) {

printChar('.');

}

while (precision-- > 0) {

long double longDoubleValue10 = 10.0L \* longDoubleValue;

int digitValue = (int) longDoubleValue10;

printChar((char) (digitValue + '0'));

longDoubleValue = longDoubleValue10 - ((long double) digitValue);

}

}

void printLongDoublePlain(long double longDoubleValue, BOOL plus,

BOOL space, BOOL grid, int precision) {

if (longDoubleValue < 0.0L) {

printChar('-');

longDoubleValue = -longDoubleValue;

plus = FALSE;

space = FALSE;

}

{ long longValue = (long) longDoubleValue;

printLongInt(longValue, plus, space);

longDoubleValue -= (long double) longValue;

printLongDoubleFraction(longDoubleValue, grid, precision);

}

}

void printLongDoubleExpo(long double value, BOOL plus, BOOL space,

BOOL grid, int precision, BOOL capital) {

if (value == 0.0L) {

printChar('0');

printLongDoubleFraction(0.0L, precision, grid);

printChar(capital ? 'E' : 'e');

printChar('0');

}

else {

if (value < 0.0L) {

printChar('-');

value = -value;

}

{ int expo = (int) log10(value);

value /= pow(10.0, expo);

printLongDoublePlain(value, plus, space, grid, precision);

printChar(capital ? 'E' : 'e');

printLongInt(expo, TRUE, FALSE);

}

}

}

va\_list checkWidthAndPrecision(va\_list arg\_list, int\* widthPtr,

int\* precisionPtr) {

if ((widthPtr != NULL) && (\*widthPtr == -1)) {

\*widthPtr = va\_arg(arg\_list, int);

}

if ((precisionPtr != NULL) && (\*precisionPtr == -1)) {

\*precisionPtr = va\_arg(arg\_list, int);

}

return arg\_list;

}

va\_list printArgument(const char\* format, va\_list arg\_list, BOOL plus,

BOOL space, BOOL grid, int\* widthPtr, int precision,

BOOL shortInt, BOOL longInt, BOOL longDouble, BOOL sign,

BOOL\* negativePtr) {

char c = format[0], charValue;

int \*intPtr;

long double longDoubleValue;

void\* ptrValue;

switch (c) {

case 'i':

case 'd': {

long longValue;

if (shortInt) {

longValue = (long) (short) va\_arg(arg\_list, int);

}

else if (longInt) {

longValue = va\_arg(arg\_list, long);

}

else {

longValue = (long) va\_arg(arg\_list, int);

}

if (negativePtr != NULL) {

\*negativePtr = (longValue < 0);

}

if (!sign) {

longValue = labs(longValue);

}

arg\_list = checkWidthAndPrecision(arg\_list, widthPtr, &precision);

printLongInt(longValue, plus, space);

}

break;

case 'c':

charValue = (char) va\_arg(arg\_list, int);

arg\_list = checkWidthAndPrecision(arg\_list, widthPtr, &precision);

printChar(charValue);

break;

case 's': {

char\* stringValue = va\_arg(arg\_list, char\*);

arg\_list = checkWidthAndPrecision(arg\_list, widthPtr, &precision);

printString(stringValue, precision);

}

break;

case 'u':

case 'o':

case 'b':

case 'x':

case 'X': {

unsigned long base = ((c == 'u') ? 10ul : ((c == 'o') ? 8ul :

((c == 'b') ? 2ul : 16ul)));

unsigned long value;

if (shortInt) {

value = (unsigned long) (unsigned short)

va\_arg(arg\_list, unsigned int);

}

else if (longInt) {

value = va\_arg(arg\_list, unsigned long);

}

else {

value = (unsigned long) va\_arg(arg\_list, unsigned int);

}

arg\_list = checkWidthAndPrecision(arg\_list, widthPtr, &precision);

printUnsignedLong(value, plus, space, grid, base, isupper(c));

}

break;

case 'f':

case 'e':

case 'E':

case 'g':

case 'G':

if (longDouble) {

longDoubleValue = va\_arg(arg\_list, long double);

printLongDoublePlain(longDoubleValue, FALSE, FALSE, FALSE, 3);

}

else {

longDoubleValue = (long double) va\_arg(arg\_list, double);

}

if (negativePtr != NULL) {

\*negativePtr = (longDoubleValue < 0);

}

if (!sign) {

longDoubleValue = fabs(longDoubleValue);

}

arg\_list = checkWidthAndPrecision(arg\_list, widthPtr, &precision);

if (c == 'f') {

printLongDoublePlain(longDoubleValue, plus, space, grid, precision);

}

else if (tolower(c) == 'e') {

printLongDoubleExpo(longDoubleValue, plus, space,

grid, precision, isupper(c));

}

else {

int expo = (int) log10(fabs(longDoubleValue));

if ((expo >= -3) && (expo <= 2)) {

printLongDoublePlain(longDoubleValue, plus, space, grid, precision);

}

else {

printLongDoubleExpo(longDoubleValue, plus, space,

grid, precision, isupper(c));

}

}

break;

case 'p':

ptrValue = va\_arg(arg\_list, void\*);

arg\_list = checkWidthAndPrecision(arg\_list, widthPtr, &precision);

printUnsignedLong((void\*) ptrValue, FALSE, FALSE, FALSE, 10u, FALSE);

break;

case 'n':

ptrValue = va\_arg(arg\_list, void\*);

intPtr = va\_arg(arg\_list, int\*);

arg\_list = checkWidthAndPrecision(arg\_list, widthPtr, &precision);

\*intPtr = g\_outChars;

break;

case '%':

arg\_list = checkWidthAndPrecision(arg\_list, widthPtr, &precision);

printChar('%');

break;

}

return arg\_list;

}

int printFormat(const char\* format, va\_list arg\_list) {

int index, width = 0, precision = 0;

BOOL percent = FALSE, plus = FALSE, minus = FALSE, space = FALSE,

zero = FALSE, grid = FALSE, widthStar = FALSE,

period = FALSE, precisionStar = FALSE,

shortInt = FALSE, longInt = FALSE, longDouble = FALSE;

g\_outChars = 0;

for (index = 0; format[index] != '\0'; ++index) {

char c = format[index];

if (percent) {

switch (c) {

case '+':

plus = TRUE;

break;

case '-':

minus = TRUE;

break;

case ' ':

space = TRUE;

break;

case '0':

zero = TRUE;

break;

case '#':

grid = TRUE;

break;

case '.':

period = TRUE;

break;

case '\*':

if (!period) {

width = -1;

}

else {

precision = -1;

}

break;

case 'h':

shortInt = TRUE;

break;

case 'l':

longInt = TRUE;

break;

case 'L':

longDouble = TRUE;

break;

case 'i':

case 'd':

case 'u':

case 'b':

case 'o':

case 'x':

case 'X':

case 'c':

case 's':

case 'f':

case 'e':

case 'E':

case 'g':

case 'G':

case 'p':

case 'n':

case '%': {

if (minus) {

int startChars = g\_outChars;

arg\_list = printArgument(&format[index], arg\_list, plus, space,

grid, &width, precision, shortInt,

longInt, longDouble, TRUE, NULL);

{ int field = g\_outChars - startChars;

while (field++ < width) {

printChar(' ');

}

}

}

else if (zero) {

int startChars = g\_outChars, oldOutStatus = g\_outStatus;

BOOL negative = FALSE;

g\_outStatus = BLANK;

printArgument(&format[index], arg\_list, FALSE, FALSE, grid,

&width, precision, shortInt, longInt,

longDouble, FALSE, &negative);

g\_outStatus = oldOutStatus;

{ int field = g\_outChars - startChars;

g\_outChars = startChars;

if (negative) {

printChar('X');

printChar('-');

++field;

}

else if (plus) {

printChar('+');

++field;

}

else if (space) {

printChar(' ');

++field;

}

while (field++ < width) {

printChar('0');

}

arg\_list = printArgument(&format[index], arg\_list, FALSE, FALSE,

grid, NULL, precision, shortInt,

longInt, longDouble, FALSE, NULL);

}

}

else {

int startChars = g\_outChars, oldOutStatus = g\_outStatus;

g\_outStatus = BLANK;

printArgument(&format[index], arg\_list, plus, space, grid,

&width, precision, shortInt, longInt,

longDouble, TRUE, NULL);

g\_outStatus = oldOutStatus;

{ int field = g\_outChars - startChars;

g\_outChars = startChars;

while (field++ < width) {

printChar(' ');

}

arg\_list = printArgument(&format[index], arg\_list, plus, space,

grid, NULL, precision, shortInt,

longInt, longDouble, TRUE, NULL);

}

}

percent = FALSE;

}

break;

default: {

int value = 0;

while (isdigit(c)) {

value = (10 \* value) + (c - '0');

c = format[++index];

}

--index;

if (!period) {

width = value;

}

else {

precision = value;

}

}

break;

}

}

else {

if (c == '%') {

percent = TRUE;

plus = FALSE;

minus = FALSE;

space = FALSE;

zero = FALSE;

grid = FALSE;

widthStar = FALSE;

period = FALSE;

precisionStar = FALSE;

shortInt = FALSE;

longInt = FALSE;

longDouble = FALSE;

width = 0;

precision = 0;

}

else {

printChar(c);

}

}

}

if (g\_outStatus == STRING) {

char\* outString = (char\*) g\_outDevice;

outString[g\_outChars] = '\0';

}

return g\_outChars;

}

int printf(char\* format, ...) {

va\_list arg\_list;

va\_start(arg\_list, format);

return vprintf(format, arg\_list);

}

int vprintf(char\* format, va\_list arg\_list) {

return vfprintf(stdout, format, arg\_list);

}

int fprintf(FILE\* outStream, char\* format, ...) {

va\_list arg\_list;

va\_start(arg\_list, format);

return vfprintf(outStream, format, arg\_list);

}

int vfprintf(FILE\* outStream, char\* format, va\_list arg\_list) {

g\_outStatus = DEVICE;

g\_outDevice = (void\*) outStream;

return printFormat(format, arg\_list);

}

int sprintf(char\* outString, char\* format, ...) {

va\_list arg\_list;

va\_start(arg\_list, format);

return vsprintf(outString, format, arg\_list);

}

int vsprintf(char\* outString, char\* format, va\_list arg\_list) {

g\_outStatus = STRING;

g\_outDevice = (void\*) outString;

return printFormat(format, arg\_list);

}

### Scanning

scanf.h

#ifndef \_\_SCANF\_H\_\_

#define \_\_SCANF\_H\_\_

#define DEVICE 0

#define STRING 1

#define EOF -1

char scanChar(void);

void unscanChar(char c);

void scanString(char\* string, int precision);

long scanLongInt(int base);

unsigned long scanUnsignedLongInt(int base);

long double scanLongDouble(void);

int scanf(const char\* format, ...);

int vscanf(const char\* format, va\_list arg\_list);

int fscanf(FILE\* inStream, const char\* format, ...);

int vfscanf(FILE\* inStream, const char\* format, va\_list arg\_list);

int sscanf(char\* inString, const char\* format, ...);

int vsscanf(char\* inString, const char\* format, va\_list arg\_list);

#endif

scanf.c

#include <math.h>

#include <ctype.h>

#include <stdio.h>

#include <stddef.h>

#include <stdarg.h>

#include <String.h>

#include <scanf.h>

#include <printf.h>

#define PRINT(x,y) { printf(#x " = <%" #y ">\n", (x)); }

int g\_inStatus, g\_inChars;

void\* g\_inDevice;

int g\_inCount;

char scanChar(void) {

char c = '\0';

FILE\* stream;

int handle;

char\* inString;

switch (g\_inStatus) {

case DEVICE:

stream = (FILE\*) g\_inDevice;

handle = stream->handle;

#ifdef \_\_WINDOWS\_\_

register\_ah = 0x3Fs;

register\_bx = handle;

register\_cx = 1;

register\_dx = &c;

interrupt(0x21s);

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 0x00L;

register\_rdi = (unsigned long) stream->handle;

register\_rsi = (unsigned long) &c;

register\_rdx = 1L;

syscall();

#endif

++g\_inChars;

return c;

case STRING:

inString = (char\*) g\_inDevice;

return inString[g\_inChars++];

default:

return '\0';

}

}

void unscanChar(char /\* c \*/) {

switch (g\_inStatus) {

case DEVICE:

--g\_inChars;

break;

case STRING:

--g\_inChars;

break;

}

}

static char\* strnchr(const char\* text, int size, int i) {

int index;

char c = (char) i;

for (index = 0; index < size; ++index) {

if (text[index] == c) {

return &text[index];

}

}

return NULL;

}

void scanPattern(char\* string, char\* pattern, int size, BOOL not) {

int index = 0;

char input = scanChar();

while (isspace(input)) {

input = scanChar();

}

if (string != NULL) {

while ((!not && strnchr(pattern, size, input)) ||

(not && !strnchr(pattern, size, input))) {

string[index++] = input;

input = scanChar();

}

string[index] = '\0';

}

else {

while ((!not && strnchr(pattern, size, input)) ||

(not && !strnchr(pattern, size, input))) {

input = scanChar();

}

}

}

void scanString(char\* string, int precision) {

int index = 0;

char input = scanChar();

BOOL found = FALSE;

while (isspace(input)) {

input = scanChar();

}

if (string != NULL) {

if (precision == 0) {

while (!isspace(input) && (input != EOF) && (input != '\n')) {

string[index++] = input;

input = scanChar();

found = TRUE;

++g\_inChars;

}

string[index] = '\0';

++g\_inChars;

}

else {

while ((precision-- > 0) && (!isspace(input) &&

(input != EOF) && (input != '\n'))) {

string[index++] = input;

input = scanChar();

found = TRUE;

++g\_inChars;

}

if (precision > 0) {

string[index] = '\0';

++g\_inChars;

}

}

}

else {

if (precision == 0) {

while (!isspace(input) && (input != EOF) &&

(input != '\n')) {

input = scanChar();

found = TRUE;

++g\_inChars;

}

++g\_inChars;

}

else {

while ((precision-- > 0) && (!isspace(input) &&

(input != EOF) && (input != '\n'))) {

input = scanChar();

found = TRUE;

++g\_inChars;

}

if (precision > 0) {

++g\_inChars;

}

}

}

if (found) {

++g\_inCount;

}

}

static BOOL isDigitInBase(char c, int base) {

if (isdigit(c)) {

int value = c - '0';

return ((value >= 0) && (value < base));

}

else if (islower(c)) {

int value = (c - 'a') + 10;

return ((value >= 0) && (value < base));

}

else if (isupper(c)) {

int value = (c - 'A') + 10;

return ((value >= 0) && (value < base));

}

else {

return FALSE;

}

}

static int digitToValue(char c) {

if (isdigit(c)) {

return (c - '0');

}

else if (islower(c)) {

return ((c - 'a') + 10);

}

else if (isupper(c)) {

return ((c - 'A') + 10);

}

else {

return 0;

}

}

long scanLongInt(int base) {

long longValue = 0l;

BOOL minus = FALSE, found = FALSE;

char input = scanChar();

while (isspace(input)) {

input = scanChar();

}

if (input == '+') {

input = scanChar();

}

else if (input == '-') {

minus = TRUE;

input = scanChar();

}

if (base == 0) {

if (input == '0') {

input = scanChar();

if (tolower(input) == 'x') {

base = 16;

input = scanChar();

}

else {

base = 8;

}

}

else {

base = 10;

}

}

while (isDigitInBase(input, base)) {

longValue \*= base;

longValue += digitToValue(input);

input = scanChar();

found = TRUE;

}

if (minus) {

longValue = -longValue;

}

if (found) {

++g\_inCount;

}

unscanChar(input);

return longValue;

}

unsigned long scanUnsignedLongInt(int base) {

unsigned long unsignedLongValue = 0, digit;

char input = scanChar();

BOOL found = TRUE;

while (isspace(input)) {

input = scanChar();

}

if (input == '+') {

input = scanChar();

}

if (base == 0) {

if (input == '0') {

input = scanChar();

if (tolower(input) == 'x') {

base = 16;

input = scanChar();

}

else {

base = 8;

}

}

else {

base = 10;

}

}

while (isDigitInBase(input, base)) {

unsignedLongValue \*= base;

unsignedLongValue += digitToValue(input);

found = TRUE;

input = scanChar();

}

if (found) {

++g\_inCount;

}

unscanChar(input);

return unsignedLongValue;

}

long double scanLongDouble(void) {

BOOL minus = FALSE, found = FALSE;

long double value = 0.0L, factor = 1.0L;

char input = scanChar();

while (isspace(input)) {

input = scanChar();

}

if (input == '+') {

input = scanChar();

}

else if (input == '-') {

minus = TRUE;

input = scanChar();

}

while (isdigit(input)) {

value = (10.0L \* value) + ((long double) (input - '0'));

input = scanChar();

found = TRUE;

}

if (input == '.') {

input = scanChar();

while (isdigit(input)) {

factor /= 10.0L;

value += factor \* ((long double) (input - '0'));

input = scanChar();

found = TRUE;

}

}

if (tolower(input) == 'e') {

double exponent = (double) scanLongInt(10);

value \*= pow(10.0, exponent);

}

else {

unscanChar(input);

}

if (minus) {

value = -value;

}

if (found) {

++g\_inCount;

}

return value;

}

int scanFormat(char\* format, va\_list arg\_list) {

char c, \*charPtr;

BOOL percent = FALSE, shortInt = FALSE, longIntOrDouble = FALSE,

longDouble = FALSE, star = FALSE;

long longValue, \*longPtr;

short\* shortPtr;

int index, \*intPtr, \*charsPtr;

unsigned long unsignedLongValue, \*unsignedLongPtr;

unsigned short\* unsignedShortPtr;

unsigned int\* unsignedIntPtr;

long double longDoubleValue;

g\_inCount = 0;

g\_inChars = 0;

for (index = 0; format[index] != '\0'; ++index) {

c = format[index];

{ int d = c + 1;

if (percent) {

switch (d - 1) {

case 'h':

shortInt = TRUE;

break;

case 'l':

longIntOrDouble = TRUE;

break;

case 'L':

longDouble = TRUE;

break;

case '\*':

star = TRUE;

break;

case 'c': {

char charValue = scanChar();

if (!star) {

charPtr = va\_arg(arg\_list, char\*);

\*charPtr = charValue;

}

percent = FALSE;

if (charValue != EOF) {

++g\_inCount;

}

}

break;

case 's':

if (!star) {

charPtr = va\_arg(arg\_list, char\*);

scanString(charPtr, 0);

}

else {

scanString(NULL, 0);

}

percent = FALSE;

break;

case 'i':

case 'd':

longValue = scanLongInt(10);

if (!star) {

if (shortInt) {

shortPtr = va\_arg(arg\_list, short\*);

\*shortPtr = (short) longValue;

}

else if (!longIntOrDouble) {

intPtr = va\_arg(arg\_list, int\*);

\*intPtr = (int) longValue;

}

else {

longPtr = va\_arg(arg\_list, long\*);

\*longPtr = longValue;

}

}

percent = FALSE;

break;

case 'o':

unsignedLongValue = scanUnsignedLongInt(8);

if (!star) {

if (shortInt) {

unsignedShortPtr = va\_arg(arg\_list, unsigned short\*);

\*unsignedShortPtr = (short) unsignedLongValue;

}

else if (!longIntOrDouble) {

unsignedIntPtr = va\_arg(arg\_list, unsigned int\*);

\*unsignedIntPtr = (int) unsignedLongValue;

}

else {

unsignedLongPtr = va\_arg(arg\_list, unsigned long\*);

\*unsignedLongPtr = unsignedLongValue;

}

}

percent = FALSE;

break;

case 'x':

unsignedLongValue = scanUnsignedLongInt(16);

if (!star) {

if (shortInt) {

unsignedShortPtr = va\_arg(arg\_list, unsigned short\*);

\*unsignedShortPtr = (short) unsignedLongValue;

}

else if (!longIntOrDouble) {

unsignedIntPtr = va\_arg(arg\_list, unsigned int\*);

\*unsignedIntPtr = (int) unsignedLongValue;

}

else {

unsignedLongPtr = va\_arg(arg\_list, unsigned long\*);

\*unsignedLongPtr = unsignedLongValue;

}

}

percent = FALSE;

break;

case 'u':

unsignedLongValue = scanUnsignedLongInt(0);

if (!star) {

if (shortInt) {

unsignedShortPtr = va\_arg(arg\_list, unsigned short\*);

\*unsignedShortPtr = (short) unsignedLongValue;

}

else if (!longIntOrDouble) {

unsignedIntPtr = va\_arg(arg\_list, unsigned int\*);

\*unsignedIntPtr = (int) unsignedLongValue;

}

else {

unsignedLongPtr = va\_arg(arg\_list, unsigned long\*);

\*unsignedLongPtr = unsignedLongValue;

}

}

percent = FALSE;

break;

case 'e':

case 'f':

case 'g':

longDoubleValue = scanLongDouble();

if (!star) {

if (longIntOrDouble) {

double\* doublePtr = va\_arg(arg\_list, double\*);

\*doublePtr = (double) longDoubleValue;

}

else if (longDouble) {

long double\* longDoublePtr = va\_arg(arg\_list, long double\*);

\*longDoublePtr = longDoubleValue;

}

else {

float\* floatPtr = va\_arg(arg\_list, float\*);

\*floatPtr = (float) longDoubleValue;

}

}

percent = FALSE;

break;

case '[': {

BOOL not = FALSE;

++index;

if (format[index] == '^') {

not = TRUE;

++index;

}

{ int startIndex = index;

while (format[index] != ']') {

++index;

}

{ int size = index - startIndex;

char c = format[index];

format[index] = '\0';

if (!star) {

char\* string = va\_arg(arg\_list, char\*);

scanPattern(string, &format[startIndex], size, not);

}

else {

scanPattern(NULL, &format[startIndex], size, not);

}

format[index] = c;

}

}

}

break;

case 'n':

charsPtr = va\_arg(arg\_list, int\*);

\*charsPtr = g\_inChars;

percent = FALSE;

break;

default:

printf("scanFormat c = '%c'\n", c);

break;

}

}

else {

if (c == '%') {

percent = TRUE;

shortInt = FALSE;

longIntOrDouble = FALSE;

longDouble = FALSE;

star = FALSE;

}

}

}

}

return g\_inCount;

}

int scanf(const char\* format, ...) {

va\_list arg\_list;

va\_start(arg\_list, format);

return vscanf(format, arg\_list);

}

int vscanf(const char\* format, va\_list arg\_list) {

return vfscanf(stdin, format, arg\_list);

}

int fscanf(FILE\* inStream, const char\* format, ...) {

va\_list arg\_list;

va\_start(arg\_list, format);

return vfscanf(inStream, format, arg\_list);

}

int vfscanf(FILE\* inStream, const char\* format, va\_list arg\_list) {

g\_inStatus = DEVICE;

g\_inDevice = (void\*) inStream;

return scanFormat(format, arg\_list);

}

int sscanf(char\* inString, const char\* format, ...) {

va\_list arg\_list;

va\_start(arg\_list, format);

return vsscanf(inString, format, arg\_list);

}

int vsscanf(char\* inString, const char\* format, va\_list arg\_list) {

g\_inStatus = STRING;

g\_inDevice = (void\*) inString;

return scanFormat(format, arg\_list);

}

### File Handling

file.h

#ifndef \_\_FILE\_H\_\_

#define \_\_FILE\_H\_\_

#define FOPEN\_MAX 20

#define FILENAME\_MAX 16

#define fpos\_t int

#define EOF -1

typedef struct {

BOOL open;

unsigned int handle;

char name[FILENAME\_MAX], ungetc;

int errno;

unsigned int position, size;

BOOL temporary;

} FILE;

extern FILE \*stdin, \*stdout, \*stderr;

extern enum {EEXIST, ENOENT, EACCES};

extern enum {SEEK\_SET, SEEK\_CUR, SEEK\_END};

extern enum {READ, WRITE, READ\_WRITE};

#ifdef \_\_LINUX\_\_

#define O\_RDONLY 0x0000L /\* open for reading only \*/

#define O\_WRONLY 0x0001L /\* open for writing only \*/

#define O\_CREAT 0x0200L /\* create if nonexistant \*/

#define O\_TRUNC 0x0400L /\* truncate to zero length \*/

#define FILE\_DESC\_STDOUT 1

#define SYS\_EXIT 1

#define SYS\_READ 3

#define SYS\_WRITE 4

#define SYS\_OPEN 5

#define SYS\_CLOSE 6

#endif

#define getc(stream) fgetc(stream)

BOOL fileexists(const char\* name);

FILE\* fopen(const char\* filename, const char\* mode);

FILE\* freopen(const char\* filename, const char\* mode, FILE\* stream);

int fflush(FILE\* stream);

int fclose(FILE\* stream);

int remove(const char\* name);

int rename(const char\* oldName, const char\* newName);

int setvbuf(FILE\* stream, char\* buffer, int mode, size\_t size);

void setbuf(FILE\* stream, char\* buffer);

int fgetc(FILE\* stream);

char\* fgets(char\* s, int n, FILE\* stream);

int fputc(int i, FILE\* stream);

int fputs(const char\* s, FILE\* stream);

int getchar(void);

char\* gets(char\* s);

int putchar(int c);

int puts(const char\* s);

int ungetc(int c, FILE\* stream);

size\_t fread(void\* ptr, size\_t size, size\_t nobj, FILE\* stream);

size\_t fwrite(const void\* ptr, size\_t size, size\_t nobj, FILE\* stream);

int fseek(FILE\* stream, int offset, int origin);

int ftell(FILE\* stream);

void rewind(FILE\* stream);

int fgetpos(FILE\* stream, fpos\_t\* ptr);

int fsetpos(FILE\* stream, const fpos\_t\* ptr);

void clearerr(FILE\* stream);

BOOL feof(FILE\* stream);

int ferror(FILE\* stream);

void perror(const char\* s);

#endif

file.c

#include <stdio.h>

#include <errno.h>

#include <locale.h>

#include <string.h>

FILE g\_fileArray[FOPEN\_MAX] = {{TRUE, 0}, {TRUE, 1}, {TRUE, 2}};

FILE \*stdin = &g\_fileArray[0], \*stdout = &g\_fileArray[1],

\*stderr = &g\_fileArray[2];

enum {EEXIST, ENOENT, EACCES};

enum {SEEK\_SET = 0, SEEK\_CUR = 1, SEEK\_END = 2};

enum {READ = 0x40, WRITE = 0x41, READ\_WRITE = 0x42};

#define MAX(a,b) (((a) > (b)) ? (a) : (b))

#define FILE\_NOT\_FOUND 0x02

#define PRINT(x,y) { printf(#x " = <%" #y ">\n", (x)); }

static int filecreate(const char\* name) {

#ifdef \_\_WINDOWS\_\_

register\_ah = 0x3Cs;

register\_cx = 0x00;

register\_dx = name;

interrupt(0x21s);

{ int handle = register\_ax;

if (carry\_flag) {

errno = FOPEN;

return -1;

}

return handle;

}

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 85;

register\_rdi = (unsigned long) name;

register\_rsi = 0777L; // octal

syscall();

return register\_eax;

#endif

}

BOOL fileexists(const char\* name) {

#ifdef \_\_WINDOWS\_\_

register\_ah = 0x43s;

register\_al = 0x00s;

register\_dx = name;

interrupt(0x21s);

return !carry\_flag;

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 5;

register\_rdi = (unsigned long) name;

register\_rsi = NULL;

syscall();

return register\_rbx;

#endif

}

static int fileopen(const char\* name, unsigned short mode) {

#ifdef \_\_WINDOWS\_\_

register\_ah = 0x3Ds;

register\_al = mode;

register\_dx = name;

interrupt(0x21s);

if (carry\_flag) {

errno = FOPEN;

return -1;

}

else {

return register\_ax;

}

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 2;

register\_rdi = (unsigned long) name;

register\_rsi = (unsigned long) mode;

syscall();

return register\_rax;

#endif

}

FILE\* fopen(const char\* name, const char\* mode) {

int index;

for (index = 0; index < FOPEN\_MAX; ++index) {

if (!g\_fileArray[index].open) {

return freopen(name, mode, &g\_fileArray[index]);

}

}

return NULL;

}

FILE\* freopen(const char\* name, const char\* mode, FILE\* stream) {

int handle = -1;

if (strcmp(mode, "r") == 0) {

handle = fileopen(name, (unsigned short) READ);

}

else if (strcmp(mode, "w") == 0) {

handle = filecreate(name);

}

else if (strcmp(mode, "a") == 0) {

handle = fileopen(name, (unsigned short) WRITE);

if (handle != -1) {

fseek(stream, 0L, (int) SEEK\_END);

}

else {

handle = filecreate(name);

}

}

else if (strcmp(mode, "r+") == 0) {

handle = fileopen(name, (unsigned short) READ\_WRITE);

}

else if (strcmp(mode, "w+") == 0) {

if (fileexists(name)) {

handle = fileopen(name, (unsigned short) READ\_WRITE);

}

else {

handle = filecreate(name);

}

}

else if (strcmp(mode, "a+") == 0) {

handle = fileopen(name, (unsigned short) READ\_WRITE);

if (handle != -1) {

fseek(stream, 0L, (int) SEEK\_END);

}

else {

handle = filecreate(name);

}

}

if (handle != -1) {

stream->open = TRUE;

stream->handle = handle;

stream->size = 0l; // filesize(handle);

strcpy(stream->name, name);

stream->temporary = FALSE;

return stream;

}

else {

stream->open = FALSE;

return NULL;

}

}

int fflush(FILE\* stream) {

if (stream == NULL) {

int index;

for (index = 0; index < FOPEN\_MAX; ++index) {

if (g\_fileArray[index].open) {

if (fflush(&g\_fileArray[index]) == EOF) {

return EOF;

}

}

}

}

// ...

return 0;

}

int fclose(FILE\* stream) {

if (stream != NULL) {

#ifdef \_\_WINDOWS\_\_

register\_ah = 0x3Es;

register\_bx = stream->handle;

interrupt(0x21s);

if (carry\_flag) {

errno = FCLOSE;

return -1;

}

if (stream->temporary) {

remove(stream->name);

}

stream->open = FALSE;

return 0;

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 3L;

register\_rdi = (unsigned long) stream->handle;

syscall();

return 0;

#endif

}

else {

int index;

for (index = 0; index < FOPEN\_MAX; ++index) {

if (g\_fileArray[index].open) {

if (fclose(&g\_fileArray[index]) == -1) {

return -1;

}

}

}

return 0;

}

}

int remove(const char\* name) {

#ifdef \_\_WINDOWS\_\_

register\_ah = 0x41s;

register\_cl = 0s;

register\_dx = name;

interrupt(0x21s);

if (!carry\_flag) {

return 0;

}

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 88L;

register\_rdi = (unsigned long) name;

syscall();

if (register\_ebx == 0) {

return 0;

}

#endif

errno = FREMOVE;

return -1;

}

int rename(const char\* oldName, const char\* newName) {

#ifdef \_\_WINDOWS\_\_

register\_ah = 0x56s;

register\_cl = 0s;

register\_dx = oldName;

register\_di = newName;

interrupt(0x21s);

if (!carry\_flag) {

return 0;

}

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 82L;

register\_rdi = (unsigned long) oldName;

register\_rsi = (unsigned long) newName;

syscall();

if (register\_eax == 0) {

return 0;

}

#endif

errno = FRENAME;

return -1;

}

int setvbuf(FILE\* /\* stream \*/, char\* /\* buffer \*/,

int /\* mode \*/, size\_t /\* size \*/) {

return 0;

}

void setbuf(FILE\* /\* stream \*/, char\* /\* buffer \*/) {

// Empty.

}

int fgetc(FILE\* stream) {

char c = '\0';

if (fread(&c, sizeof (char), 1, stream) > 0) {

return (int) c;

}

return -1;

}

char\* fgets(char\* text, int size, FILE\* stream) {

int count = 0;

char prevChar = '\0';

while ((count < (size - 1))) {

char currChar = '\0';

fscanf(stream, "%c", &currChar);

if ((prevChar == '\r') && (currChar == '\n')) {

text[count] = '\0';

break;

}

if (currChar == -1) {

text[count] = '\0';

break;

}

if ((currChar != '\r') && (currChar != '\n')) {

text[count++] = currChar;

}

prevChar = currChar;

}

return text;

}

int fputs(const char\* s, FILE\* stream) {

int size = (strlen(s) + 1) \* sizeof (char);

return (fwrite(s, size, 1, stream) == size) ? 0 : EOF;

}

int getchar(void) {

return fgetc(stdin);

}

char\* gets(char\* s) {

if (fgets(s, -1, stdin) != NULL) {

int size = strlen(s);

if (size > 0) {

s[size - 1] = '\0';

}

return s;

}

else {

return NULL;

}

}

int puts(const char\* s) {

if (fputs(s, stdout) != 0) {

return fputc('\n', stdout);

}

return EOF;

}

int ungetc(int c, FILE\* stream) {

if (stream->ungetc != EOF) {

stream->ungetc = (char) c;

}

return c;

}

size\_t fread(void\* ptr, size\_t size, size\_t nobj, FILE\* stream) {

#ifdef \_\_WINDOWS\_\_

register\_bx = stream->handle;

register\_cx = size \* nobj;

register\_ah = 0x3Fs;

register\_dx = ptr;

interrupt(0x21s);

if (carry\_flag) {

stream->errno = errno = FREAD;

return 0;

}

else {

return register\_ax;

}

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 0L;

register\_rdi = (unsigned long) stream->handle;

register\_rsi = (unsigned long) ptr;

register\_rdx = (unsigned long) (size \* nobj);

syscall();

return register\_eax;

#endif

}

size\_t fwrite(const void\* ptr, size\_t size, size\_t nobj, FILE\* stream) {

#ifdef \_\_WINDOWS\_\_

register\_bx = stream->handle;

register\_cx = size \* nobj;

register\_ah = 0x40s;

register\_dx = ptr;

interrupt(0x21s);

if (carry\_flag) {

stream->errno = errno = FWRITE;

return 0;

}

else {

return register\_ax;

}

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 1;

register\_rdi = (unsigned long) stream->handle;

register\_rsi = (unsigned long) ptr;

register\_rdx = (unsigned long) (size \* nobj);

syscall();

return register\_eax;

#endif

}

int fseek(FILE\* stream, int offset, int origin) {

#ifdef \_\_WINDOWS\_\_

register\_al = (short) origin;

register\_ah = 0x42s;

register\_bx = stream->handle;

register\_cx = 0;

register\_dx = (int) offset;

interrupt(0x21s);

if (!carry\_flag) {

stream->position = register\_ax;

return stream->position;

}

else {

stream->errno = FSEEK;

return -1;

}

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 8;

register\_rdi = (unsigned long) stream->handle;

register\_rsi = (unsigned long) offset;

register\_rdx = (unsigned long) origin;

syscall();

return register\_eax;

#endif

}

int ftell(FILE\* stream) {

return fseek(stream, 0, SEEK\_CUR);

}

void rewind(FILE\* stream) {

(void) fseek(stream, 0, SEEK\_SET);

}

int fgetpos(FILE\* stream, fpos\_t\* ptr) {

\*ptr = (fpos\_t) ftell(stream);

return 0;

}

int fsetpos(FILE\* stream, const fpos\_t\* ptr) {

return ((int) fseek(stream, \*ptr, (int) SEEK\_SET));

}

void clearerr(FILE\* stream) {

stream->errno = errno = 0;

}

BOOL feof(FILE\* stream) {

long unsigned currPosition = fseek(stream, 0L, (int) SEEK\_CUR);

long unsigned lastPosition = fseek(stream, 0L, (int) SEEK\_END);

fseek(stream, currPosition, (int) SEEK\_SET);

{ BOOL endOfFile = (currPosition == lastPosition);

return endOfFile;

}

}

int ferror(FILE\* stream) {

return stream->errno;

}

void perror(const char\* s) {

printf("%s: %s.\n", s, strerror(errno));

}

## The Standard Library

The standard library holds functions for type casting, random number generating, absolute value, division and remainder as well as searching and sorting.

stdlib.h

#ifndef \_\_STDLIB\_H\_\_

#define \_\_STDLIB\_H\_\_

#define NULL ((void\*) 0)

double atof(const char\* s);

int atoi(const char\* s);

long atol(const char\* s);

double strtod(const char\* s, char\*\* endp);

long strtol(const char\* s, char\*\* endp, int base);

unsigned long strtoul(const char\* s, char\*\* endp, int base);

int rand(void);

void srand(unsigned int seed);

char\* getenv(const char\* name);

int system(const char\* command);

void abort(void);

void exit(int status);

typedef void (\*FUNC\_PTR)(void);

int atexit(FUNC\_PTR fcn);

#define FUNC\_MAX 256

#define OPEN\_MAX 16

int abs(int value);

long labs(long value);

void\* malloc(size\_t size);

void\* realloc(void\* ptr, size\_t newSize);

void\* calloc(size\_t num, size\_t size);

void free(void\* ptr);

void qsort(void\* valueList, size\_t listSize, size\_t valueSize,

int (\*compare)(const void\*, const void\*));

void\* bsearch(const void\* key, const void\* valueList,

size\_t listSize, size\_t valueSize,

int (\*compare)(const void\*, const void\*));

int abs(int value);

long labs(long value);

typedef struct {

int quot, rem;

} div\_t;

div\_t div(int num, int denum);

typedef struct {

long quot, rem;

} ldiv\_t;

ldiv\_t ldiv(long num, long denum);

#endif

stdlib.c

#include <math.h>

#include <ctype.h>

#include <errno.h>

#include <stdarg.h>

#include <stddef.h>

#include <string.h>

#include <stdlib.h>

#include <stdio.h>

extern FILE g\_fileArray[];

The atoi (ASCII-to-integer) and atol (ASCII-to-long) functions cast a string to an integer value by calling strtol.

int atoi(const char\* s) {

return (int) strtol(s, (const char\*\*) NULL, 10);

}

long atol(const char\* s) {

return strtol(s, (const char\*\*) NULL, 10);

}

The strtol (string-to-long)

extern int g\_inStatus, g\_inChars;

extern void\* g\_inDevice;

long strtol(const char\* s, char\*\* endp, int base) {

g\_inStatus = STRING;

g\_inDevice = s;

g\_inChars = 0;

{ long value = scanLongInt(base);

if (endp != NULL) {

\*endp = s + g\_inChars;

}

return value;

}

}

The strtoul (string-to-unsigned-long)

unsigned long strtoul(const char\* s, char\*\* endp, int base) {

g\_inStatus = STRING;

g\_inDevice = s;

g\_inChars = 0;

{ unsigned long unsignedLongValue = scanUnsignedLongInt(base);

if (endp != NULL) {

\*endp = s + g\_inChars;

}

return unsignedLongValue;

}

}

double atof(const char\* s) {

return strtod(s, (char\*\*) NULL);

}

The strtod (string-to-double)

double strtod(const char\* s, char\*\* endp) {

int chars = '\0';

double value = 0;

sscanf(s, "%lf%n", &value, &chars);

if (endp != NULL) {

\*endp = s + chars;

}

return value;

}

void abort(void) {

#ifdef \_\_WINDOWS\_\_

register\_ah = 0x4Cs;

register\_al = -1s;

interrupt(0x21s);

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 60L;

register\_rdi = -1L;

syscall();

#endif

}

char\* getenv(const char\* /\* name \*/) {

return NULL;

}

int system(const char\* /\* command \*/) {

return -1;

}

void memswp(void\* value1, void\* value2, int valueSize) {

char\* charValue1 = (char\*) value1;

char\* charValue2 = (char\*) value2;

int index;

for (index = 0; index < valueSize; ++index) {

char tempValue = charValue1[index];

charValue1[index] = charValue2[index];

charValue2[index] = tempValue;

}

}

void\* bsearch(const void\* keyPtr, const void\* valueList,

size\_t listSize, size\_t valueSize,

int (\*compare)(const void\*, const void\*)) {

int firstIndex = 0, lastIndex = listSize - 1;

if (listSize == 0) {

return NULL;

}

while (TRUE) {

{ char\* firstValuePtr = ((char\*) valueList) + (firstIndex \* valueSize);

int firstCompare = compare(keyPtr, firstValuePtr);

if (firstCompare < 0) {

return NULL;

}

else if (firstCompare == 0) {

return firstValuePtr;

}

}

{ char\* lastValuePtr = ((char\*) valueList) + (lastIndex \* valueSize);

int lastCompare = compare(keyPtr, lastValuePtr);

if (lastCompare > 0) {

return NULL;

}

else if (lastCompare == 0) {

return lastValuePtr;

}

}

{ int middleIndex = (firstIndex + lastIndex) / 2;

char\* middleValuePtr = ((char\*)valueList) + (middleIndex \* valueSize);

int middleCompare = compare(keyPtr, middleValuePtr);

if (middleCompare < 0) {

lastIndex = middleIndex;

}

else if (middleCompare > 0) {

firstIndex = middleIndex;

}

else {

return middleValuePtr;

}

}

}

}

static long g\_randValue;

#define A 1664525l

#define C 1013904223l

#define RAND\_MAX 127

int rand(void) {

g\_randValue = ((A \* g\_randValue) + C) % RAND\_MAX;

return (int) g\_randValue;

}

void srand(unsigned int seed) {

g\_randValue = (long) seed;

}

#define STACK\_TOP\_ADDRESS 32766

#define HEAP\_BOTTOM\_ADDRESS 32764

#define HEAP\_TOP\_ADDRESS 32762

#define HEADER\_SIZE (2 \* sizeof (unsigned int))

FUNC\_PTR g\_funcArray[FUNC\_MAX] = { NULL };

int atexit(FUNC\_PTR fcn) {

int index;

for (index = 0; index < FUNC\_MAX; ++index) {

if (g\_funcArray[index] == NULL) {

g\_funcArray[index] = fcn;

return 0;

}

}

return -1;

}

void exit(int status) {

int index;

for (index = (FUNC\_MAX - 1); index >= 0; --index) {

if (g\_funcArray[index] != NULL) {

g\_funcArray[index]();

}

}

#ifdef \_\_WINDOWS\_\_

register\_al = (short) status;

register\_ah = 0x4Cs;

interrupt(0x21s);

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 60L;

register\_rdi = (unsigned long) status;

syscall();

#endif

}

void swap(char\* leftValuePtr, char\* rightValuePtr, int valueSize) {

int index;

for (index = 0; index < valueSize; ++index) {

char tempValue = leftValuePtr[index];

leftValuePtr[index] = rightValuePtr[index];

rightValuePtr[index] = tempValue;

}

}

void qsort(void\* valueList, size\_t listSize, size\_t valueSize,

int (\*compare) (const void\*, const void\*)) {

char\* charList = (char\*) valueList;

int size;

for (size = (listSize - 1); size > 0; --size) {

int index;

BOOL update = FALSE;

for (index = 0; index < size; ++index) {

char\* valuePtr1 = charList + (index \* valueSize);

char\* valuePtr2 = charList + ((index + 1) \* valueSize);

if (compare(valuePtr1, valuePtr2) > 0) {

memswap(valuePtr1, valuePtr2, valueSize);

update = TRUE;

}

}

if (!update) {

break;

}

}

}

int abs(int value) {

return (value < 0) ? -value : value;

}

long labs(long value) {

return (value < 0l) ? -value : value;

}

div\_t div(int num, int denum) {

div\_t result = {0, 0};

if (denum == 0) {

errno = EDOM;

return result;

}

result.quot = num / denum;

result.rem = num % denum;

return result;

}

ldiv\_t ldiv(long num, long denum) {

ldiv\_t result = {0, 0};

if (denum == 0) {

errno = EDOM;

return result;

}

result.quot = num / denum;

result.rem = num % denum;

return result;

}

## Time

time.h

#ifndef \_\_TIME\_H\_\_

#define \_\_TIME\_H\_\_

#define size\_t int

#define time\_t unsigned long

#define clock\_t long

struct tm {

int tm\_sec;

int tm\_min;

int tm\_hour;

int tm\_mday;

int tm\_mon;

int tm\_year;

int tm\_wday;

int tm\_yday;

int tm\_isdst;

};

extern clock\_t clock(void);

extern time\_t time(time\_t\* time);

extern double difftime(time\_t time2, time\_t time1);

extern time\_t mktime(struct tm\* timeStruct);

extern char\* asctime(const struct tm\* timeStruct);

extern char\* ctime(const time\_t\* time);

extern struct tm\* gmtime(const time\_t\* time);

extern struct tm\* localtime(const time\_t\* time);

extern struct tm\* localtimeX(const time\_t\* time);

extern size\_t strftime(char\* buffer, size\_t size,

const char\* format, const struct tm\* timeStruct);

#endif

time.c

#include <time.h>

#include <stdio.h>

#include <stdlib.h>

#include <string.h>

#include <locale.h>

#include <assert.h>

clock\_t clock(void) {

return -1;

}

time\_t time(time\_t\* timePtr) {

time\_t time;

#ifdef \_\_WINDOWS\_\_

int year;

short month, monthDay;

short hour, min, sec;

register\_ah = 0x2As;

interrupt(0x21s);

year = register\_cx - 1900;

month = register\_dh - 1s;

monthDay = register\_dl;

register\_ah = 0x2Cs;

interrupt(0x21s);

hour = register\_ch;

min = register\_cl;

sec = register\_dh;

{ const BOOL leapYear = (year % 4) == 0;

const int daysOfMonths[] = {31, leapYear ? 29 : 28, 31, 30,

31, 30, 30, 31, 30, 31, 30, 31};

int yearDay = monthDay - 1, mon;

for (mon = 0; mon < month; ++mon) {

yearDay += daysOfMonths[mon];

}

{ struct tm s = {sec, min, hour, monthDay, month, year, 0, yearDay, 0};

time = mktime(&s);

}

}

#endif

#ifdef \_\_LINUX\_\_

register\_rax = 201L;

register\_rdi = (unsigned long) &time;

syscall();

#endif

if (timePtr != NULL) {

\*timePtr = time;

}

return time;

}

time\_t mktime(struct tm\* tp) {

if (tp != NULL) {

const long leapDays = (tp->tm\_year - 69) / 4;

const long totalDays = 365 \* (tp->tm\_year - 70) + leapDays + tp->tm\_yday;

return (86400L \* totalDays) + (3600L \* tp->tm\_hour) +

(60L \* tp->tm\_min) + tp->tm\_sec;

}

return 0;

}

static struct tm g\_timeStruct;

struct tm\* gmtime(const time\_t\* timePtr) {

int year = 1970;

if (timePtr != NULL) {

time\_t time = \*timePtr;

const long secondsOfDay = time % 86400L;

long totalDays = time / 86400L;

g\_timeStruct.tm\_hour = secondsOfDay / 3600;

g\_timeStruct.tm\_min = (secondsOfDay % 3600) / 60;

g\_timeStruct.tm\_sec = (secondsOfDay % 3600) % 60;

// January 1, 1970, was a Thursday

if (totalDays < 3) {

g\_timeStruct.tm\_wday = totalDays + 4;

}

else {

g\_timeStruct.tm\_wday = (totalDays - 3) % 7;

}

while (TRUE) {

const BOOL leapYear = (((year % 4) == 0) &&

((year % 100) != 0)) || ((year % 400) == 0);

const int daysOfYear = leapYear ? 366 : 365;

if (totalDays < daysOfYear) {

const int daysOfMonths[] = {31, leapYear ? 29 : 28, 31, 30,

31, 30, 30, 31, 30, 31, 30, 31};

int month = 0;

g\_timeStruct.tm\_year = year - 1900;

g\_timeStruct.tm\_yday = totalDays;

while (totalDays >= daysOfMonths[month]) {

totalDays -= daysOfMonths[month];

++month;

}

g\_timeStruct.tm\_mon = month;

g\_timeStruct.tm\_mday = totalDays + 1;

g\_timeStruct.tm\_isdst = -1;

return &g\_timeStruct;

}

++year;

totalDays -= daysOfYear;

}

}

return NULL;

}

double difftime(time\_t time1, time\_t time2) {

return (double) (time2 - time1);

}

static char g\_timeString[256];

static char\* g\_defaultShortDayList[] = {"Sun", "Mon", "Tue", "Wed",

"Thu", "Fri", "Sat"};

static char\* g\_defaultLongDayList[] = {"Sunday", "Monday", "Tuesday",

"Wednesday", "Thursday", "Friday", "Saturday"};

static char\* g\_defaultShortMonthList[] =

{"Jan", "Feb", "Mar", "Apr", "May", "Jun",

"Jul", "Aug", "Sep", "Oct", "Nov", "Dec"};

static char\* g\_defaultLongMonthList[] =

{"January", "February", "March", "April", "May", "June",

"July", "August", "September", "October", "November", "December"};

char\* asctime(const struct tm\* tp) {

struct lconv\* localeConvPtr = NULL;

char\*\* shortDayList = (localeConvPtr != NULL) ? localeConvPtr->shortDayList

: NULL;

char\*\* shortMonthList = (localeConvPtr != NULL)

? localeConvPtr->shortMonthList : NULL;

shortDayList = (shortDayList != NULL) ? shortDayList

: g\_defaultShortDayList;

shortMonthList = (shortMonthList != NULL) ? shortMonthList

: g\_defaultShortMonthList;

sprintf(g\_timeString, "%s %s %2i %02i:%02i:%02i %04i",

shortDayList[tp->tm\_wday], shortMonthList[tp->tm\_mon],

tp->tm\_mday, tp->tm\_hour, tp->tm\_min,

tp->tm\_sec, tp->tm\_year + 1900);

return g\_timeString;

}

char\* ctime(const time\_t\* time) {

return asctime(localtime(time));

}

struct tm\* localtime(const time\_t\* timePtr) {

struct tm\* tmPtr = gmtime(timePtr);

struct lconv\* localeConvPtr = localeconv();

int timeZone = 0;

if (localeConvPtr != NULL) {

timeZone = tmPtr->tm\_isdst ? localeConvPtr->summerTimeZone

: localeConvPtr->winterTimeZone;

}

{ time\_t time = \*timePtr + (3600 \* timeZone);

return gmtime(&time);

}

}

size\_t strftime(char\* s, size\_t smax, const char\* fmt, const struct tm\* tp) {

struct lconv\* localeConvPtr = localeconv();

char\*\* shortDayList = (localeConvPtr != NULL)

? (localeConvPtr->shortDayList) : NULL;

char\*\* shortMonthList = (localeConvPtr != NULL)

? (localeConvPtr->shortMonthList) : NULL;

char\*\* longDayList = (localeConvPtr != NULL)

? (localeConvPtr->longDayList) : NULL;

char\*\* longMonthList = (localeConvPtr != NULL)

? (localeConvPtr->longMonthList) : NULL;

const BOOL leapDays = (tp->tm\_year - 69) / 4;

const long totalDays = 365 \* (tp->tm\_year - 70) + leapDays + tp->tm\_yday;

int yearDaySunday, yearDayMonday;

strcpy(s, "");

shortDayList = (shortDayList != NULL)

? shortDayList : g\_defaultShortDayList;

longDayList = (longDayList != NULL) ? longDayList : g\_defaultLongDayList;

shortMonthList = (shortMonthList != NULL)

? shortMonthList : g\_defaultShortMonthList;

longMonthList = (longMonthList != NULL)

? longMonthList : g\_defaultLongMonthList;

// January 1, 1970, was a Thursday

if (totalDays < 3) {

yearDaySunday = totalDays + 4;

}

else {

yearDaySunday = (totalDays - 3) % 7;

}

if (totalDays < 4) {

yearDayMonday = totalDays + 3;

}

else {

yearDayMonday = (totalDays - 4) % 7;

}

{ int index;

for (index = 0; fmt[index] != '\0'; ++index) {

char add[20];

if (fmt[index] == '%') {

switch (fmt[++index]) {

case 'a':

strcpy(add, shortDayList[tp->tm\_wday]);

break;

case 'A':

strcpy(add, longDayList[tp->tm\_wday]);

break;

case 'b':

strcpy(add, shortMonthList[tp->tm\_mon]);

break;

case 'B':

strcpy(add, longMonthList[tp->tm\_mon]);

break;

case 'c':

sprintf(add, "%04d-%02d-%02d %02d:%02d:%02d",

1900 + tp->tm\_year, tp->tm\_mon + 1, tp->tm\_mday,

tp->tm\_hour, tp->tm\_min, tp->tm\_sec);

break;

case 'd':

sprintf(add, "%02d", tp->tm\_mday);

break;

case 'H':

sprintf(add, "%02d", tp->tm\_hour);

break;

case 'I':

sprintf(add, "%02d", tp->tm\_hour % 12);

break;

case 'j':

sprintf(add, "%03d", tp->tm\_yday);

break;

case 'm':

sprintf(add, "%02d", tp->tm\_mon + 1);

break;

case 'M':

sprintf(add, "%02d", tp->tm\_min);

break;

case 'p':

sprintf(add, "%s", (tp->tm\_hour < 12) ? "AM" : "PM");

break;

case 'S':

sprintf(add, "%02d", tp->tm\_sec);

break;

case 'U':

sprintf(add, "%02d", yearDaySunday);

break;

case 'w':

sprintf(add, "%02d", tp->tm\_wday);

break;

case 'W':

sprintf(add, "%02d", yearDayMonday);

break;

case 'x':

sprintf(add, "%04d-%02d-%02d", 1900 + tp->tm\_year,

tp->tm\_mon + 1, tp->tm\_mday);

break;

case 'X':

sprintf(add, "%02d:%02d:%02d", tp->tm\_hour,

tp->tm\_min, tp->tm\_sec);

break;

case 'y':

sprintf(add, "%02d", tp->tm\_year % 100);

break;

case 'Y':

sprintf(add, "%04d", 1900 + tp->tm\_year);

break;

case 'Z':

strcpy(add, "");

break;

case '%':

strcpy(add, "%");

}

}

else {

add[0] = fmt[index];

add[1] = '\0';

}

if ((strlen(s) + strlen(add)) < smax) {

strcat(s, add);

}

else {

break;

}

}

}

return strlen(s);

}

1. The Preprocessor

The preprocessor processes the source code before the compiler. This chapter is divided into two parts, where the first part is a scanner and a parser, and the second part takes care of comments, string, and characters as well as handling macros and conditional programming.

## The Expression Scanner and Parser

Before we consider the preprocessor itself, we need to find a way to decide the whether the expressions given in the #if or #elif directives equals zero. For that we need the expression parser, which can be viewed as a smaller version of the main parser of Chapter 3. We look into the grammar, scanner, and parser.

### The Grammar

The grammar of an expression parser follows below. Similar to the expression part of the main parser of Chapter 3, we add a new grammar rule for each precedence level.

expression ::=

logical\_or\_expression

| logical\_or\_expression **?** expression **:** expression

logical\_or\_expression ::=

logical\_and\_expression

| logical\_or\_expression **||** logical\_and\_expression

logical\_and\_expression ::=

Bitwise\_or\_expression

| bitwise\_and\_expression **&&** bitwise\_or\_expression

bitwise\_or\_expression ::=

bitwise\_xor\_expression

| bitwise\_or\_expression **|** bitwise\_xor\_expression

bitwise\_xor\_expression ::=

bitwise\_and\_expression

| bitwise\_xor\_expression **^** bitwise\_and\_expression

bitwise\_and\_expression ::=

equality\_expression

| bitwise\_and\_expression **&** equality\_expression

equality\_expression ::=

relation\_expression

| equality\_expression **==** relation\_expression

| equality\_expression **!=** relation\_expression

relation\_expression ::=

shift\_expression

| relation\_expression **<** shift\_expression

| relation\_expression **<=** shift\_expression

| relation\_expression **>** shift\_expression

| relation\_expression **>=** shift\_expression

shift\_expression ::=

add\_expression

| shift\_expression **<<** add\_expression

| shift\_expression **>>** add\_expression

add\_expression ::=

multiply\_expression

| add\_expression **+** multiply\_expression

| add\_expression **-** multiply\_expression

multiply\_expression ::=

prefix\_expression

| multiply\_expression **\*** prefix\_expression

| multiply\_expression **/** prefix\_expression

| multiply\_expression **%** prefix\_expression

prefix\_expression ::=

primary\_expression

| **+** prefix\_expression

| **-** prefix\_expression

| **~** prefix\_expression

| **!** prefix\_expression

primary\_expression ::=

**identifier**

| **value**

| **defined** **identifier**

| **defined** **( identifier )**

| **(** expression **)**

### The Parser

The result of the parser is an integer value, floating values are not used by the preprocessor.

ExpressionParser.gppg

%namespace CCompiler\_Exp

%partial

%using CCompiler;

%{

// Empty.

%}

%token DEFINED ADD SUBTRACT MULTIPLY DIVIDE MODULO EQUAL NOT\_EQUAL LESS\_THAN

LESS\_THAN\_EQUAL GREATER\_THAN GREATER\_THAN\_EQUAL LEFT\_SHIFT RIGHT\_SHIFT

QUESTION\_MARK COLON LEFT\_PARENTHESIS RIGHT\_PARENTHESIS LOGICAL\_OR

LOGICAL\_AND LOGICAL\_NOT BITWISE\_XOR BITWISE\_OR BITWISE\_AND BITWISE\_NOT

EOL

%union {

public string name;

public int integer\_value;

}

%token <name> NAME

%token <integer\_value> VALUE

%type <integer\_value> expression logical\_or\_expression

logical\_and\_expression bitwise\_or\_expression

bitwise\_xor\_expression bitwise\_and\_expression

equality\_expression relation\_expression

shift\_expression add\_expression

multiply\_expression prefix\_expression

primary\_expression

%start expression

%%

The result of the parsing is placed in the PreProcessorResult of the Preprocessor class. The return value of the parser is a Boolean value representing whether the expression complied with the grammar.

expression:

logical\_or\_expression {

$$ = $1;

Preprocessor.PreProcessorResult = $$;

}

| logical\_or\_expression QUESTION\_MARK expression

COLON expression {

$$ = ($1 != 0) ? $3 : $5;

Preprocessor.PreProcessorResult = $$;

};

The logical\_or\_expression and logical\_and\_expression return one or zero

logical\_or\_expression:

logical\_and\_expression {

$$ = $1;

}

| logical\_or\_expression LOGICAL\_OR logical\_and\_expression {

$$ = (($1 != 0) || ($3 != 0)) ? 1 : 0;

};

logical\_and\_expression:

bitwise\_or\_expression {

$$ = $1;

}

| bitwise\_and\_expression LOGICAL\_AND bitwise\_or\_expression {

$$ = (($1 != 0) && ($3 != 0)) ? 1 : 0;

};

bitwise\_or\_expression:

bitwise\_xor\_expression {

$$ = $1;

}

| bitwise\_or\_expression BITWISE\_OR bitwise\_xor\_expression {

$$ = $1 | $3;

};

bitwise\_xor\_expression:

bitwise\_and\_expression {

$$ = $1;

}

| bitwise\_xor\_expression BITWISE\_XOR bitwise\_and\_expression {

$$ = $1 ^ $3;

};

bitwise\_and\_expression:

equality\_expression {

$$ = $1;

}

| bitwise\_and\_expression BITWISE\_AND equality\_expression {

$$ = $1 & $3;

};

equality\_expression:

relation\_expression {

$$ = $1;

}

| equality\_expression EQUAL relation\_expression {

$$ = ($1 == $3) ? 1 : 0;

}

| equality\_expression NOT\_EQUAL relation\_expression {

$$ = ($1 != $3) ? 1 : 0;

};

relation\_expression:

shift\_expression { $$ = $1; }

| relation\_expression LESS\_THAN shift\_expression {

$$ = ($1 < $3) ? 1 : 0;

}

| relation\_expression LESS\_THAN\_EQUAL shift\_expression {

$$ = ($1 <= $3) ? 1 : 0;

}

| relation\_expression GREATER\_THAN shift\_expression {

$$ = ($1 > $3) ? 1 : 0;

}

| relation\_expression GREATER\_THAN\_EQUAL shift\_expression {

$$ = ($1 >=$3) ? 1 : 0;

};

shift\_expression:

add\_expression {

$$ = $1;

}

| shift\_expression LEFT\_SHIFT add\_expression {

$$ = $1 << $3;

}

| shift\_expression RIGHT\_SHIFT add\_expression {

$$ = $1 >> $3;

};

add\_expression:

multiply\_expression {

$$ = $1;

}

| add\_expression ADD multiply\_expression {

$$ = $1 + $3;

}

| add\_expression SUBTRACT multiply\_expression {

$$ = $1 - $3;

};

multiply\_expression:

prefix\_expression {

$$ = $1;

}

| multiply\_expression MULTIPLY prefix\_expression {

$$ = $1 \* $3;

}

| multiply\_expression DIVIDE prefix\_expression {

$$ = $1 / $3;

}

| multiply\_expression MODULO prefix\_expression {

$$ = $1 % $3;

};

prefix\_expression:

primary\_expression {

$$ = $1;

}

| ADD prefix\_expression {

$$ = $2;

}

| SUBTRACT prefix\_expression {

$$ = -$2;

}

| BITWISE\_NOT prefix\_expression {

$$ = ~$2;

}

An expression is considered to be true if it does not equal zero.

| LOGICAL\_NOT prefix\_expression {

$$ = ($2 == 0) ? 1 : 0;

};

In the preprocessor directive, every identifier is treated as the value zero.

primary\_expression:

NAME {

$$ = 1;

}

| VALUE {

$$ = $1;

}

The defined directive works both with and without parenthesis.

|  |  |
| --- | --- |
| #if defined (x) | #if defined x |
| (a) With parenthesis | (b) Without parenthesis |

ExpressionParser.gppg

| DEFINED NAME {

$$ = Preprocessor.MacroMap.ContainsKey($2) ? 1 : 0;

}

| DEFINED LEFT\_PARENTHESIS NAME RIGHT\_PARENTHESIS {

$$ = Preprocessor.MacroMap.ContainsKey($3) ? 1 : 0;

}

| LEFT\_PARENTHESIS expression RIGHT\_PARENTHESIS {

$$ = $2;

};

### The Scanner

As mention in Section 0, tokens are the smallest parts of the language. However, the parser cannot know that, for instance, a sequence of letters is an identifier, or that a sequence of digits is an integer value. Instead, that is the task of a scanner.

ExpressionScanner.gplex

%namespace CCompiler\_Exp

%using CCompiler;

%{

// Empty.

%}

OCTAL\_VALUE (\+|\-)?[0][0-7]\*([uU]|[sSlL]|[uU][sSlL]|[sSlL][uU])?

DECIMAL\_VALUE (\+|\-)?[1-9][0-9]\*([uU]|[sSlL]|[uU][sSlL]|[sSlL][uU])?

HEXADECIMAL\_VALUE (\+|\-)?[0][xX][0-9a-fA-F]+([uU]|[sSlL]|[uU][sSlL]|[sSlL][uU])?

NAME [a-zA-Z\_][a-zA-Z0-9\_]\*

WHITE\_SPACE [ \t\r\n\f]

%%

"defined" { return ((int) Tokens.DEFINED); }

"?" { return ((int) Tokens.QUESTION\_MARK); }

":" { return ((int) Tokens.COLON); }

"||" { return ((int) Tokens.LOGICAL\_OR); }

"&&" { return ((int) Tokens.LOGICAL\_AND); }

"!" { return ((int) Tokens.LOGICAL\_NOT); }

"&" { return ((int) Tokens.BITWISE\_AND); }

"^" { return ((int) Tokens.BITWISE\_XOR); }

"|" { return ((int) Tokens.BITWISE\_OR); }

"~" { return ((int) Tokens.BITWISE\_NOT); }

"==" { return ((int) Tokens.EQUAL); }

"!=" { return ((int) Tokens.NOT\_EQUAL); }

"<" { return ((int) Tokens.LESS\_THAN); }

"<=" { return ((int) Tokens.LESS\_THAN\_EQUAL); }

">" { return ((int) Tokens.GREATER\_THAN); }

">=" { return ((int) Tokens.GREATER\_THAN\_EQUAL); }

"<<" { return ((int) Tokens.LEFT\_SHIFT); }

">>" { return ((int) Tokens.RIGHT\_SHIFT); }

"+" { return ((int) Tokens.ADD); }

"-" { return ((int) Tokens.SUBTRACT); }

"\*" { return ((int) Tokens.MULTIPLY); }

"/" { return ((int) Tokens.DIVIDE); }

"%" { return ((int) Tokens.MODULO); }

"(" { return ((int) Tokens.LEFT\_PARENTHESIS); }

")" { return ((int) Tokens.RIGHT\_PARENTHESIS); }

"\0" { return ((int) Tokens.EOL); }

In case of an octal, decimal, or hexadecimal value, we just remove the letters u (unsigned), s (short), l (long), and x (hexadecimal value marker) from the text and let the Integer class interpret the value.

{NAME} {

yylval.name = yytext;

return ((int) Tokens.NAME);

}

{OCTAL\_VALUE} {

{ string text = yytext.ToLower().Replace("u", "").Replace("s", "")

.Replace("l", "");

yylval.integer\_value = Convert.ToInt32(text, 8);

return ((int) Tokens.VALUE);

}

}

{DECIMAL\_VALUE} {

{ string text = yytext.ToLower().Replace("u", "").Replace("s", "")

.Replace("l", "");

yylval.integer\_value = Convert.ToInt32(text, 10);

return ((int) Tokens.VALUE);

}

}

{HEXADECIMAL\_VALUE} {

{ string text = yytext.ToLower().Replace("x", "").Replace("u", "").

Replace("s", "").Replace("l", "");

yylval.integer\_value = Convert.ToInt32(text, 16);

return ((int) Tokens.VALUE);

}

}

It is important to detect white spaces, otherwise the user would not be allowed to add any spaces to the code.

{WHITE\_SPACE} {

if (yytext.Equals("\n")) {

++CCompiler\_Main.Scanner.Line;

}

}

If we detect a character we do not recognize, we just stop the execution with an error message.

. { Assert.Error(yytext, Message.Unknown\_character); }

## The Preprocessor Scanner and Parser

There is actually a second parser and scanner, that parsers the source code. The parser is very simple. Its only task is to define a set of terminals.

PreParser.gppg

%namespace CCompiler\_Pre

%partial

%{

// Empty.

%}

%union {

public string name;

}

%token <name> NAME STRING LEFT\_PARENTHESIS RIGHT\_PARENTHESIS

COMMA SHARP DOUBLE\_SHARP TOKEN MARK // EOL

%start source\_code\_file

%%

source\_code\_file:

/\* Empty. \*/;

%%

The scanner is a little bit more complicated. Its task is to define a set of tokens. Similar to the scanner of Chapter 2, it defines a set of values tokens are to be used of the preprocessor. The idea is that we divide the

PreScanner.gplex

%namespace CCompiler\_Pre

%{

public static int NewlineCount = 0;

public static bool Whitespace = false;

%}

OCTAL\_VALUE (\+|\-)?[0][0-7]\*([uU]|[sSlL]|[uU][sSlL]|[sSlL][uU])?

DECIMAL\_VALUE (\+|\-)?[1-9][0-9]\*([uU]|[sSlL]|[uU][sSlL]|[sSlL][uU])?

HEXADECIMAL\_VALUE (\+|\-)?[0][xX][0-9a-fA-F]+

([uU]|[sSlL]|[uU][sSlL]|[sSlL][uU])? // XXX

FLOATING\_VALUE (\+|\-)?(([0-9]+"."[0-9]\*|"."[0-9]+)([eE][\+\-]?[0-9]+)?|[0-9]+[eE][\+\-]?[0-9]+)([fF]|[lL])?

CHAR\_VALUE \'("\n"|"\\\'"|[^\'])\*\'

STRING\_VALUE \"("\n"|"\\\""|[^\"])\*\"

NAME [a-zA-Z\_][a-zA-Z0-9\_]\*

WHITE\_SPACE [ \t\r\n\f]

TOKEN [^a-zA-Z0-9()#,\"\' \t\r\n\f\0]+

%%

"(" { yylval.name = "("; return ((int) Tokens.LEFT\_PARENTHESIS); }

")" { yylval.name = ")"; return ((int) Tokens.RIGHT\_PARENTHESIS); }

"##" { yylval.name = "##"; return ((int) Tokens.DOUBLE\_SHARP); }

"#" { yylval.name = "#"; return ((int) Tokens.SHARP); }

"," { yylval.name = ","; return ((int) Tokens.COMMA); }

//"\0" { /\*CCompiler\_Pre.Scanner.NewlineCount = 0;\*/ yylval.name = ""; return ((int) Tokens.EOL); }

{NAME} {

yylval.name = yytext;

return ((int) Tokens.NAME);

}

{OCTAL\_VALUE} {

yylval.name = yytext;

return ((int) Tokens.TOKEN);

}

{DECIMAL\_VALUE} {

yylval.name = yytext;

return ((int) Tokens.TOKEN);

}

{HEXADECIMAL\_VALUE} {

yylval.name = yytext;

return ((int) Tokens.TOKEN);

}

{FLOATING\_VALUE} {

yylval.name = yytext;

return ((int) Tokens.TOKEN);

}

{CHAR\_VALUE} {

yylval.name = yytext;

return ((int) Tokens.TOKEN);

}

{STRING\_VALUE} {

yylval.name = yytext;

return ((int) Tokens.STRING);

}

{TOKEN} {

yylval.name = yytext;

return ((int) Tokens.TOKEN);

}

{WHITE\_SPACE} {

if (yytext.Equals("\n")) {

++CCompiler\_Pre.Scanner.NewlineCount;

}

CCompiler\_Pre.Scanner.Whitespace = true;

}

. { CCompiler.Assert.Error("<" + yytext + ">",

CCompiler.Message.Unknown\_character); }

### If-Else-Chain

An If-Else-Chain is a sequence of if, elif, else, and endif preprocessor directives. We use the IfElseChain class to keep track of the sequence. We need to keep track of three statues:

Former Status. Has an earlier if or elif directive been evaluated to true? In that case, all succeeding if, elif, or else directive shall be evaluated to false.

Current Status. Has the current if, elif, or else directive been evaluated to true? In that case the code of this directive is visible; that is, the code shall be included. If the directive has been evaluated to false, the code is invisible and shall be excluded.

Else Status. A chain can only hold one else directive. If we encounter a second else directive, we report an error.

The IfElseChain class holds the former, current, and else status of an if-else-chain.

IfElseChain.cs

namespace CCompiler {

class IfElseChain {

bool m\_formerStatus, m\_currentStatus, m\_elseStatus;

public IfElseChain(bool formerStatus, bool currentStatus, bool elseStatus){

m\_formerStatus = formerStatus;

m\_currentStatus = currentStatus;

m\_elseStatus = elseStatus;

}

public bool FormerStatus {

get { return m\_formerStatus; }

}

public bool CurrentStatus {

get { return m\_currentStatus; }

}

public bool ElseStatus {

get { return m\_elseStatus; }

}

}

}

## The Preprocessor

The preprocessor has several tasks:

**Tri Graphs**. When C was originally introduced, some keyboards had a limited set of keys. Therefore, a special set of double question mark character sequence was introduced, which are replaced by modern equivalents.

**Comments**. The line comments are removed, and each block comment is replaced by a blank character.

**Slashes in strings and characters**. Each backslash is processed and transformed into a regular character. Thereafter, to ease the macro expansion later we transform each character into the format of a backslash followed by three octal digits.

**Include files**. The system include files (encapsulated by ‘<’ and ‘>’) and internal include files (encapsulated by quotes) are read and included in the final code.

**Conditional programming**. The #if, #ifdef, #ifndef, #elif, and #endif directives are processed.

**Macro expansion**. The source code is traversed, and each macro is expanded. Since the contents of each character and string constant has been translated into slash codes, we do not need to check whether the macro to be expanded is part of a character of string.

**String sequences**. Finally, every sequence of string is concatenated into one string. For instance, the sequence "ab" "cd" "ef" is concatenated into "abcded".

The first phase of the preprocessor is to read the source file and place in a text buffer (a StringBuilder object), and then replace the tri graph sequences, remove the comments and replace every character in string and characters with its equivalent backslash code. Then the buffer is divided into lines and each line starting with a preprocessor directive is evaluated and every line is search for macros to expand.

Preprocessor.cs

using System;

using System.IO;

using System.Text;

using System.Collections.Generic;

namespace CCompiler {

public class Preprocessor {

private IDictionary<string,Macro> m\_macroMap =

new Dictionary<string,Macro>();

private Stack<FileInfo> m\_includeStack = new Stack<FileInfo>();

private ISet<FileInfo> m\_includeSet = new HashSet<FileInfo>();

private Stack<IfElseChain> m\_ifElseChainStack = new Stack<IfElseChain>();

private StringBuilder m\_outputBuffer = new StringBuilder();

public string PreprocessedCode {

get { return m\_outputBuffer.ToString(); }

}

public ISet<FileInfo> IncludeSet {

get { return m\_includeSet; }

}

We add the \_\_LINUX\_\_ macro for the compiler to compile the source code in accordance with the Linux environment, and the \_\_WINDOWS\_\_ macro for the Windows environment. There are several parts of the standard library source code that are different depending on whether the macro is present.

public Preprocessor(FileInfo file) {

if (Start.Linux) {

m\_macroMap.Add("\_\_LINUX\_\_", new Macro(0, new List<Token>()));

}

if (Start.Windows) {

m\_macroMap.Add("\_\_WINDOWS\_\_", new Macro(0, new List<Token>()));

}

DoProcess(file);

Assert.Error(m\_ifElseChainStack.Count == 0, Message.

If\_\_\_ifdef\_\_\_\_or\_ifndef\_directive\_without\_matching\_endif);

}

public void DoProcess(FileInfo file) {

StreamReader streamReader = new StreamReader(file.FullName);

StringBuilder inputBuffer =

new StringBuilder(streamReader.ReadToEnd());

streamReader.Close();

CCompiler\_Main.Scanner.Path = file;

CCompiler\_Main.Scanner.Line = 2;

GenerateTriGraphs(inputBuffer);

TraverseBuffer(inputBuffer);

List<string> lineList =

GenerateLineList(inputBuffer.ToString());

CCompiler\_Main.Scanner.Line = 1;

int stackSize = m\_ifElseChainStack.Count;

TraverseLineList(lineList);

Assert.Error(m\_ifElseChainStack.Count ==

stackSize, Message.Unbalanced\_if\_and\_endif\_directive\_structure);

}

### Tri Graphs

A tri graph character sequence is three-character sequence that starts with two question marks and represent another character. It is a remain from old times, when the keyboards hold a reduced set of characters. The following table shows the tri graphs character sequences and their modern equivalences:

|  |  |
| --- | --- |
| **Tri-Graph Character Sequence** | **Modern Equivalence** |
| ??= | # |
| ??/ | \ |
| ??´ | ^ |
| ??( | [ |
| ??) | ] |
| ??! | | |
| ??< | { |
| ??> | } |
| ??- | ~ |

We need to replace the tri graphs sequences with their modern equivalences, which is easy to do since we do not need to take any consideration in whether the sequences are placed inside strings, characters, or comments.

private static IDictionary<string,string> m\_triGraphMap =

new Dictionary<string,string>()

{{"??=", "#"}, {"??/", "\\"}, {"??\'", "^"},

{"??(", "["}, {"??)", "]"}, {"??!", "|"},

{"??<", "{"}, {"??>", "}"}, {"??-", "~"}};

public void GenerateTriGraphs(StringBuilder buffer) {

foreach (KeyValuePair<string,string> pair in m\_triGraphMap) {

buffer.Replace(pair.Key, pair.Value);

}

}

### Comments, Strings, and Characters

The next step is to take care of block comments (/\* to \*/) and line comments (// to the end of the line), strings, and characters. Since they may be nested, they need to be evaluated in the same phase.

The idea is that we go through the buffer, and when we find the beginning of a line comment, a block comment, a string, or a character we call the corresponding method, which modify the buffer and returns the index of the next character to be inspected.

To make the inspection of the source code buffer easier we start by adding three zero characters, which are removed at the end of the method.

The traverseLineComment method is quite easy, we just continue until we find the end of the line (‘\n’) or the end of the buffer (‘\0’), and replace every preceding character with a blank.

The traverseBlockComment method is a little bit more complicated. We continue until we find the end of the comment (\*/). If we instead find the end of the buffer (‘\0’) the comment is unterminated, which result in an error message. Each character is replaced by a space character. However, newlines (‘\n’) are not replaced, in order for the line count to work properly.

The traverseString method needs to keep track of the end of the string (‘\”’), the end of the buffer (‘\0’), or newline in the string (‘\n’). The end of the buffer or newline result in error messages. Moreover, when we find a backslash (‘\\’) we call doSlash, that transforms the slash code into a regular character. Finally, when the string has been traversed each regular character is modified into a backslash followed by three octal digits, in order to ease the macro replacement process in a later phase.

The traverseCharacter method is similar to traverseString above, it generates error messages for end-of-line or newline, calls doSlash for each backslash, and finally changes each regular character into a backslash followed by three octal digits.

public void TraverseBuffer(StringBuilder buffer) {

for (int index = 0; index < buffer.Length; ++index) {

if ((buffer[index] == '/') && (buffer[index + 1] == '\*')) {

buffer[index++] = ' ';

buffer[index++] = ' ';

for (; true; ++index) {

if (index == buffer.Length) {

Assert.Error(Message.Unfinished\_block\_comment);

}

else if ((buffer[index] == '\*') && (buffer[index + 1] == '/')) {

buffer[index++] = ' ';

buffer[index] = ' ';

break;

}

else if (buffer[index] == '\n') {

++CCompiler\_Main.Scanner.Line;

}

else {

buffer[index] = ' ';

}

}

}

else if ((buffer[index] == '/') && (buffer[index + 1] == '/')) {

buffer[index++] = ' ';

buffer[index++] = ' ';

for (; true; ++index) {

if ((index == buffer.Length) || (buffer[index] == '\n')) {

break;

}

else {

buffer[index] = ' ';

}

}

}

else if ((buffer[index] == '\'')) {

++index;

for (; true; ++index) {

if (index == buffer.Length) {

Assert.Error(Message.Unfinished\_character);

}

else if (buffer[index] == '\n') {

Assert.Error(Message.Newline\_in\_character);

}

else if ((buffer[index] == '\\') && (buffer[index] == '\'')) {

++index;

}

else if (buffer[index] == '\'') {

break;

}

}

}

else if ((buffer[index] == '\"')) {

++index;

for (; true; ++index) {

if (index == buffer.Length) {

Assert.Error(Message.Unfinished\_string);

}

else if (buffer[index] == '\n') {

Assert.Error(Message.Newline\_in\_string);

}

else if ((buffer[index] == '\\') && (buffer[index] == '\"')) {

++index;

}

else if (buffer[index] == '\"') {

break;

}

}

}

else if (buffer[index] == '\n') {

++CCompiler\_Main.Scanner.Line;

}

}

}

### Slash Codes

The slashToChar method inspect the character succeeding the slash.

If the characters following the slash are a lowercase ‘x’ or an uppercase ‘X’ and one or two hexadecimal digits, or three octal digits XXX the following two characters are inspected, and their value calculated and the two hexadecimal digits are replaced three octal digits. However, if the two following characters are not hexadecimal digits, an error message occurs.

If the three, two, or one characters following the slash are octal digits, the value of the digits is calculated and the slash sequence is replaced by the character with the ASCII value of the octal digits. If the value of the digits exceeds 255, we report an error.

If the character following the slash is a lowercase ‘x’ or an uppercase ‘X’ and two or one hexadecimal digits, the value of the digits is calculated, and the slash sequence is replaced by the character with the ASCII value of the hexadecimal digits. A lowercase ‘x’ or an uppercase ‘X’ not followed by at least one hexadecimal digit results in an error being reported.

If none of the cases above applies, we report an error.

When the character or string has been translated from slash codes to regular characters, charToOctal is called to translate them to octal slash codes.

The octalToChar method is called by the scanner to tralform the octal shalch codes into regular characters.

### The Line List

When the buffer has been modified in accordance with the methods above, it is time to transform it into a list of lines to process the preprocessors directives. The transformation is simple, we just use the split method. We add an extra blank line in the last line is a macro definition that ends with two slashes (“\\”).

private List<string> GenerateLineList(string text) {

List<string> trimList = new List<string>();

foreach (string line in text.Split('\n')) {

trimList.Add(line.Trim());

}

int index = 0;

List<string> resultList = new List<string>();

while (index < trimList.Count) {

if (trimList[index].StartsWith("#")) {

StringBuilder buffer = new StringBuilder();

for (; (index < trimList.Count) && trimList[index].EndsWith("\\");

++index) {

buffer.Append(trimList[index].

Substring(0, trimList[index].Length - 1) + "\n");

}

if (index < trimList.Count) {

buffer.Append(trimList[index++] + "\n");

}

resultList.Add(buffer.ToString());

}

else {

StringBuilder buffer = new StringBuilder();

for (; (index < trimList.Count) && !trimList[index].StartsWith("#");

++index) {

buffer.Append(trimList[index] + "\n");

}

resultList.Add(buffer.ToString());

}

}

return resultList;

}

However, we also need to merge lines that ends with two backslashes together:

|  |  |
| --- | --- |
| #define min(x,y) \\  (((x) < (y)) ? \\  (x) : (y)) | #define min(x,y) (((x) < (y)) ? (x) : (y)) |
| (a) Before | (b) After |

If we encounter lines ending with two backslashes, we continue to append the lines to lineBuffer. Since an empty line was added at the end of the line list in generateLineList above, there is no risk that that the last line ends with a backslash.

When traversing the lines, we check whether the line starts with a sharp (‘#’). If it does, we consider the word following the sharp sign (#). If it is a preprocessor directive, we call the corresponding method. If the line does not start with a sharp and we are in a visible part of source code, we call searchForMacros, which expands macros. If we are not in a visible part of the source code, the line is replaced by an empty line.

private List<Token> Scan(string text) {

byte[] byteArray = Encoding.ASCII.GetBytes(text);

MemoryStream memoryStream = new MemoryStream(byteArray);

CCompiler\_Pre.Scanner scanner = new CCompiler\_Pre.Scanner(memoryStream);

List<Token> tokenList = new List<Token>();

while (true) {

CCompiler\_Pre.Tokens tokenId = (CCompiler\_Pre.Tokens) scanner.yylex();

tokenList.Add(new Token(tokenId, scanner.yylval.name));

if (tokenId == CCompiler\_Pre.Tokens.EOF) {

break;

}

}

memoryStream.Close();

return tokenList;

}

private string TokenListToString(List<Token> tokenList) {

StringBuilder buffer = new StringBuilder();

foreach (Token token in tokenList) {

buffer.Append(((buffer.Length > 0) ? " " : "") + token.ToString());

}

return buffer.ToString();

}

private List<Token> CloneList(List<Token> tokenList) {

List<Token> resultList = new List<Token>();

foreach (Token token in tokenList) {

resultList.Add((Token) token.Clone());

}

return resultList;

}

public void TraverseLineList(List<string> lineList) {

foreach (string line in lineList) {

List<Token> tokenList = Scan(line);

if (tokenList[0].Id == CCompiler\_Pre.Tokens.SHARP) {

Token secondToken = tokenList[1];

if (secondToken.Id == CCompiler\_Pre.Tokens.NAME) {

string secondTokenName = (string) secondToken.Value;

if (secondTokenName.Equals("ifdef")) {

DoIfDefined(tokenList);

}

else if (secondTokenName.Equals("ifndef")) {

DoIfNotDefined(tokenList);

}

else if (secondTokenName.Equals("if")) {

DoIf(tokenList);

}

else if (secondTokenName.Equals("elif")) {

DoElseIf(tokenList);

}

else if (secondTokenName.Equals("else")) {

DoElse(tokenList);

}

else if (secondTokenName.Equals("endif")) {

DoEndIf(tokenList);

}

else if (IsVisible()) {

if (secondTokenName.Equals("include")) {

DoInclude(tokenList);

}

else if (secondTokenName.Equals("define")) {

DoDefine(tokenList);

}

else if (secondTokenName.Equals("undef")) {

DoUndef(tokenList);

}

else if (secondTokenName.Equals("line")) {

DoLine(tokenList);

}

else if (secondTokenName.Equals("error")) {

Assert.Error(TokenListToString

(tokenList.GetRange(1, tokenList.Count - 1)));

}

}

}

AddNewlinesToBuffer(tokenList);

}

else {

if (IsVisible()) {

SearchForMacros(tokenList, new Stack<string>());

ConcatTokens(tokenList);

MergeStrings(tokenList);

AddTokenListToBuffer(tokenList);

}

else {

AddNewlinesToBuffer(tokenList);

}

}

}

}

private void AddTokenListToBuffer(List<Token> tokenList) {

foreach (Token token in tokenList) {

m\_outputBuffer.Append(token.ToNewlineString() + token.ToString());

}

}

private void AddNewlinesToBuffer(List<Token> tokenList) {

foreach (Token token in tokenList) {

m\_outputBuffer.Append(token.ToNewlineString());

CCompiler\_Main.Scanner.Line += token.GetNewlineCount();

}

}

The IsVisible method checks whether we are in a visible part of the source code; that is, a part that has not been earlier omitted by a #if, #ifdef, #ifndef, #elif, or #else preprocessor directive. We use the Main.IfStack stack since conditional programming can be nested. If the source code is not visible at any nesting level (that is, anywhere in the stack), the area is not visible, and we return false.

private bool IsVisible() {

foreach (IfElseChain ifElseChain in m\_ifElseChainStack) {

if (!ifElseChain.CurrentStatus) {

return false;

}

}

return true;

}

The trimLeft and trimRight methods simple removes all white-spaces to the left or right of the string.

The countChar method simple counts the number of occurrences of the character.

### Lines

The doLine method handles the #line directive by setting the Main.Path and Main.Line fields and returns a text with the path and line that starts and ends with dollar signs (‘$’).

|  |  |  |
| --- | --- | --- |
| #line 100 C:\Temp\Test.c |  |  |
| $C:\Temp\Test.c,100$ |  |  |

private void DoLine(List<Token> tokenList) {

int listSize = tokenList.Count;

if ((listSize == 4) || (listSize == 5)) {

string lineText = (string) tokenList[2].Value;

Assert.Error(int.TryParse(lineText, out CCompiler\_Main.Scanner.Line),

lineText, Message.Invalid\_line\_number);

if (listSize == 5) {

CCompiler\_Main.Scanner.Path =

new FileInfo((string) tokenList[3].Value);

}

}

else {

Assert.Error(listSize == 3, TokenListToString(tokenList),

Message.Invalid\_preprocessor\_directive);

}

m\_outputBuffer.Append(Symbol.SeparatorId + CCompiler\_Main.Scanner.Path +

"," + CCompiler\_Main.Scanner.Line +

Symbol.SeparatorId + "\n");

}

### Includes

There are two kinds of include files: system files (‘<’ and ‘>’) and internal files (‘\”’). The systems files are included from the path given by Main.IncludePath, and the internal files are included locally.

The Main.IncludeSet set is used to track the included file so that the source file include the files shall be recompiled if it or any of its included files has been changed after the object file. The Main.IncludeStack is used to make sure that there is no circular inclusion.

The Main.Path and Main.Line fields are temporary replaced by the path and line of the included file. After the include line has been read, the original path and line is added to the included source code, in the dollar sign format (for example: “$C:\Temp\Test.c,100$”). The method returns the source code of the included file, which will replace the original include line.

private void DoInclude(List<Token> tokenList) {

FileInfo includeFile = null;

if ((tokenList[2].Id == CCompiler\_Pre.Tokens.STRING) &&

(tokenList[3].Id == CCompiler\_Pre.Tokens.EOF)) {

string text = tokenList[2].ToString();

string file = text.ToString().Substring(1, text.Length - 1);

includeFile = new FileInfo(Start.SourcePath + file);

}

else {

StringBuilder buffer = new StringBuilder();

foreach (Token token in tokenList.GetRange(2, tokenList.Count - 2)) {

buffer.Append(token.ToString());

}

string text = buffer.ToString();

if (text.StartsWith("<") && text.EndsWith(">")) {

string file = text.ToString().Substring(1, text.Length - 2);

includeFile = new FileInfo(Start.SourcePath + file);

}

else {

Assert.Error(TokenListToString(tokenList),

Message.Invalid\_preprocessor\_directive);

}

}

Assert.Error(!m\_includeStack.Contains(includeFile),

includeFile.FullName, Message.Repeted\_include\_statement);

m\_includeStack.Push(includeFile);

m\_includeSet.Add(includeFile);

FileInfo oldPath = CCompiler\_Main.Scanner.Path;

int oldLine = CCompiler\_Main.Scanner.Line;

CCompiler\_Main.Scanner.Path = includeFile;

CCompiler\_Main.Scanner.Line = 1;

m\_outputBuffer.Append(Symbol.SeparatorId + CCompiler\_Main.Scanner.Path +

",0" + Symbol.SeparatorId + "\n");

DoProcess(includeFile);

CCompiler\_Main.Scanner.Line = oldLine;

CCompiler\_Main.Scanner.Path = oldPath;

m\_outputBuffer.Append(Symbol.SeparatorId + CCompiler\_Main.Scanner.Path +

"," + (CCompiler\_Main.Scanner.Line - 1) +

Symbol.SeparatorId + "\n");

m\_includeStack.Pop();

}

### Macros

The Macro class keep track of a macro, a macro has a possible empty list of parameters and a text.

Macro.cs

using System.Linq;

using System.Collections.Generic;

namespace CCompiler {

public class Macro {

private int m\_parameters;

private List<Token> m\_tokenList;

public Macro(int parameters, List<Token> tokenList) {

m\_parameters = parameters;

m\_tokenList = new List<Token>(tokenList);

for (int index = (m\_tokenList.Count - 1); index >= 0; --index) {

if (m\_tokenList[index].Id == CCompiler\_Pre.Tokens.EOF) {

m\_tokenList.RemoveAt(index);

}

}

}

public int Parameters() {

return m\_parameters;

}

public List<Token> TokenList() {

return m\_tokenList;

}

public override int GetHashCode() {

return base.GetHashCode();

}

public override bool Equals(object obj) {

if (obj is Macro) {

Macro macro = (Macro) obj;

return (m\_parameters == macro.m\_parameters) &&

(m\_tokenList.SequenceEqual(macro.m\_tokenList));

}

return false;

}

}

}

Token.cs

using System.Text;

namespace CCompiler {

public class Token {

private CCompiler\_Pre.Tokens m\_id;

private object m\_value;

private bool m\_whitespace;

private int m\_newlineCount;

public Token(CCompiler\_Pre.Tokens id, object value) {

m\_id = id;

m\_value = value;

m\_newlineCount = CCompiler\_Pre.Scanner.NewlineCount;

m\_whitespace = CCompiler\_Pre.Scanner.Whitespace;

CCompiler\_Pre.Scanner.NewlineCount = 0;

CCompiler\_Pre.Scanner.Whitespace = false;

}

public Token(CCompiler\_Pre.Tokens id, object value, int newlineCount) {

m\_id = id;

m\_value = value;

m\_newlineCount = newlineCount;

m\_whitespace = false;

}

public void AddNewlineCount(int newlineCount) {

m\_newlineCount += newlineCount;

}

public object Clone() {

Token token = new Token(m\_id, m\_value);

token.m\_newlineCount = 0;

return token;

}

public CCompiler\_Pre.Tokens Id {

get { return m\_id; }

set { m\_id = value; }

}

public object Value {

get { return m\_value; }

set { m\_value = value; }

}

public int GetNewlineCount() {

return m\_newlineCount;

}

public string ToNewlineString() {

StringBuilder buffer = new StringBuilder();

if (m\_newlineCount > 0) {

for (int count = 0; count < m\_newlineCount; ++count) {

buffer.Append('\n');

}

}

else {

buffer.Append(' ');

}

return buffer.ToString();

}

public void ClearNewlineCount() {

m\_newlineCount = 0;

}

public bool HasWhitespace() {

return m\_whitespace;

}

public override int GetHashCode() {

return base.GetHashCode();

}

public override bool Equals(object obj) {

if (obj is Token) {

Token token = (Token) obj;

return (m\_id == token.m\_id) &&

m\_value.Equals(token.m\_value);

}

return false;

}

public override string ToString() {

switch (m\_id) {

case CCompiler\_Pre.Tokens.EOF:

return "";

case CCompiler\_Pre.Tokens.MARK:

return "<mark " + ((string) m\_value) + ">";

default:

return m\_value.ToString();

}

}

}

}

The doDefine method splits the line into the name of the macro, a potential parameter list, and a body. When the first character that is not a letter, digit, or an underline is reached, we check whether the line so far is a proper identifier. Then we have three cases:

1. We have reached the end of the line, in which case we have a macro with a name, but without parameters or body.

2. We have reached a space, in which case we have a macro with a name and a body, but no parameters.

3. We have reached a left parenthesis, in which case we have a list of parameters and we call doParameterDefine.

4. If none of the above applies, we generate an error message.

Preprocessor.cs

The doParameterDefine method look into macros with parameters in two steps. First we extract the parameters by calling scanParameters. Then we go through the parameter list and check that the parameters are identifiers and that no parameters is repeated.

Then we extract the identifiers from the body and replace each occurrence of a parameters with the text “$parameter\_index$”.

Finally, we look for the merge operator ##. When we find one, we remove the operator and trim the text to its left and right.

It is allowed to redefine a macro if it is has the same parameters list and macro.

An identifier is a text starting with a letter or an underline and continuing with letters, digits, or underlines. If the given name is not an identifier, an error occurs.

public void DoDefine(List<Token> tokenList) {

Assert.Error(tokenList[2].Id == CCompiler\_Pre.Tokens.NAME,

TokenListToString(tokenList),

Message.Invalid\_define\_directive);

string name = tokenList[2].ToString();

Macro macro;

if ((tokenList[3].Id == CCompiler\_Pre.Tokens.LEFT\_PARENTHESIS) &&

!tokenList[3].HasWhitespace()) {

int tokenIndex = 4, paramIndex = 0;

IDictionary<string,int> paramMap = new Dictionary<string,int>();

while (true) {

Token nextToken = tokenList[tokenIndex++];

Assert.Error(nextToken.Id == CCompiler\_Pre.Tokens.NAME,

nextToken.ToString(),

Message.Invalid\_macro\_definitializerion);

string paramName = (string) nextToken.Value;

Assert.Error(!paramMap.ContainsKey(paramName),

paramName, Message.Repeated\_macro\_parameter);

paramMap.Add(paramName, paramIndex++);

nextToken = tokenList[tokenIndex++];

if (nextToken.Id == CCompiler\_Pre.Tokens.COMMA) {

// Empty.

}

else if (nextToken.Id == CCompiler\_Pre.Tokens.RIGHT\_PARENTHESIS) {

break;

}

else {

Assert.Error(nextToken.ToString(),

Message.Invalid\_macro\_definitializerion);

}

}

List<Token> macroList =

tokenList.GetRange(tokenIndex, tokenList.Count - tokenIndex);

foreach (Token macroToken in macroList) {

if (macroToken.Id == CCompiler\_Pre.Tokens.NAME) {

string macroName = (string) macroToken.Value;

if (paramMap.ContainsKey(macroName)) {

macroToken.Id = CCompiler\_Pre.Tokens.MARK;

macroToken.Value = paramMap[macroName];

}

}

}

macro = new Macro(paramMap.Count, macroList);

}

else {

macro = new Macro(0, tokenList.GetRange(3, tokenList.Count - 3));

}

if (!m\_macroMap.ContainsKey(name)) {

m\_macroMap.Add(name, macro);

}

else {

Assert.Error(m\_macroMap[name].Equals(macro), name,

Message.Invalid\_macro\_redefinitializerion);

}

}

The doUndef method removes a macro from the MacroMap map. If the macro does not exists a warning is given.

public void DoUndef(List<Token> tokenList) {

Assert.Error((tokenList[2].Id == CCompiler\_Pre.Tokens.NAME) &&

(tokenList[3].Id == CCompiler\_Pre.Tokens.EOF),

TokenListToString(tokenList),

Message.Invalid\_undef\_directive);

string name = tokenList[2].ToString();

Assert.Error(m\_macroMap.Remove(name),

name, Message.Macro\_not\_defined);

}

### Conditional Programming

With conditional programming, it is possible to exclude part of the source code. The doIf method uses the parser of Section 13.13. It calls parseExpression, which creates a scanner and parser, parse the line, and returns a Boolean value that is true in case of a non-zero value.

private void DoIf(List<Token> tokenList) {

bool result = ParseExpression(TokenListToString

(tokenList.GetRange(2, tokenList.Count - 2)));

m\_ifElseChainStack.Push(new IfElseChain(result, result, false));

}

public static object PreProcessorResult;

private bool ParseExpression(string line) {

int result = 0;

try {

byte[] byteArray = Encoding.ASCII.GetBytes(line);

MemoryStream memoryStream = new MemoryStream(byteArray);

CCompiler\_Exp.Scanner expressionScanner =

new CCompiler\_Exp.Scanner(memoryStream);

CCompiler\_Exp.Parser expressionParser =

new CCompiler\_Exp.Parser(expressionScanner, m\_macroMap);

Assert.Error(expressionParser.Parse(), Message.Preprocessor\_parser);

result = (int) PreProcessorResult;

memoryStream.Close();

}

catch (Exception exception) {

Console.Out.WriteLine(exception.StackTrace);

Assert.Error(line, Message.Invalid\_expression);

}

return (result != 0);

}

The doIfDefined and doIfNotDefined checks if the macro is defined or not, respectively, and add the result to the Main.IfStack stack.

private void DoIfDefined(List<Token> tokenList) {

Assert.Error((tokenList[2].Id == CCompiler\_Pre.Tokens.NAME) &&

(tokenList[3].Id == CCompiler\_Pre.Tokens.EOF),

TokenListToString(tokenList),

Message.Invalid\_preprocessor\_directive);

bool result = m\_macroMap.ContainsKey((string) tokenList[2].Value);

m\_ifElseChainStack.Push(new IfElseChain(result, result, false));

}

private void DoIfNotDefined(List<Token> tokenList) {

Assert.Error((tokenList[2].Id == CCompiler\_Pre.Tokens.NAME) &&

(tokenList[3].Id == CCompiler\_Pre.Tokens.EOF),

TokenListToString(tokenList),

Message.Invalid\_preprocessor\_directive);

bool result = !m\_macroMap.ContainsKey((string)tokenList[2].Value);

m\_ifElseChainStack.Push(new IfElseChain(result, result, false));

}

The doIfElseIf method is a little bit more complicated. It checks whether there is a preceding #if directive (Main.IfStack is not empty). Thereafter, it looks up the status of the if-stack. The stack value is a ifElseChain, the first value indicates whether the current if-status is true, and the second value gives whether there has been an earlier true if-status. The third value is of type IfStatus, that can hold the value If and Else. If the value is Else, and #elseif has already occurred, and another on is not allowed. Finally, if this if-chain has ever been true (lastTrue is true) is does not matter whether the condition of directive is true, the new status is false. If not, we consider if this value is true. In both cases, we push the result on the stack.

private void DoElseIf(List<Token> tokenList) {

Assert.Error(m\_ifElseChainStack.Count > 0, Message.

Elif\_directive\_without\_preceeding\_if\_\_\_\_ifdef\_\_\_\_or\_ifndef\_directive);

IfElseChain ifElseChain = m\_ifElseChainStack.Pop();

Assert.Error(!ifElseChain.ElseStatus,

Message.Elif\_directive\_following\_else\_directive);

if (ifElseChain.FormerStatus) {

m\_ifElseChainStack.Push(new IfElseChain(true, false, false));

}

else {

bool result =

ParseExpression(TokenListToString

(tokenList.GetRange(2, tokenList.Count - 2)));

m\_ifElseChainStack.Push(new IfElseChain(result, result, false));

}

}

The doElse method is a little bit easier. It checks whether there has ever been a true part of the if-chain. If not, it pushes true on the stack.

private void DoElse(List<Token> tokenList) {

Assert.Error(m\_ifElseChainStack.Count > 0, Message.

Else\_directive\_without\_preceeding\_if\_\_\_\_ifdef\_\_\_\_or\_ifndef\_directive);

Assert.Error(tokenList[2].Id == CCompiler\_Pre.Tokens.EOF,

TokenListToString(tokenList),

Message.Invalid\_preprocessor\_directive);

IfElseChain ifElseChain = m\_ifElseChainStack.Pop();

Assert.Error(!ifElseChain.ElseStatus,

Message.Else\_directive\_after\_else\_directive);

bool formerStatus = ifElseChain.FormerStatus;

m\_ifElseChainStack.Push(new IfElseChain(!formerStatus,

!formerStatus, true));

}

Finally, the doEndIf just pops the if-stack, subject to it is not already empty.

private void DoEndIf(List<Token> tokenList) {

Assert.Error(m\_ifElseChainStack.Count > 0, Message.

Endif\_directive\_without\_preceeding\_if\_\_\_\_ifdef\_\_\_\_or\_ifndef\_directive);

Assert.Error(tokenList[2].Id == CCompiler\_Pre.Tokens.EOF,

tokenList[2].ToString(),

Message.Invalid\_preprocessor\_directive);

m\_ifElseChainStack.Pop();

}

### Macro Expansion

In the regular source code (line not starting with ‘#’), we need to replace macro calls with their resulting bodies. We also need to recursively replace macro calls in the bodies. However, unlike function, recursive macro calls are not allowed.

We go through the text, and look for identifiers, we do not have to worry about identifier in strings or characters, since every letter or underline (an identifier must begin with a letter or underline) in strings or characters has been changed to a three-octal-digit backslash code by slashToChar in Section 13.15.2.

When we find and identifier that is a stored macro name (either in m\_specialMacroSet or MacroMap), we scan the parameters if the name is followed by a left parenthesis, each parameter is recursively searched for macros. We then look up the macro body and replace its marked parameter locations with the parameters. Finally, we replace the macro call with its body. The nameStack stack is used to prevent recursive macro calls.

private void SearchForMacros(List<Token> tokenList,

Stack<string> nameStack) {

for (int index = 0; index < tokenList.Count; ++index) {

Token thisToken = tokenList[index];

CCompiler\_Main.Scanner.Line += thisToken.GetNewlineCount();

if (thisToken.Id == CCompiler\_Pre.Tokens.NAME) {

string name = (string) thisToken.Value;

int beginNewlineCount = thisToken.GetNewlineCount();

if (!nameStack.Contains(name) && m\_macroMap.ContainsKey(name)) {

Token nextToken = tokenList[index + 1];

if ((nextToken.Id == CCompiler\_Pre.Tokens.LEFT\_PARENTHESIS) &&

!nextToken.HasWhitespace()) {

int countIndex = index + 2, level = 1, totalNewlineCount = 0;

List<Token> subList = new List<Token>();

List<List<Token>> mainList = new List<List<Token>>();

while (true) {

nextToken = tokenList[countIndex];

int newlineCount = nextToken.GetNewlineCount();

totalNewlineCount += newlineCount;

CCompiler\_Main.Scanner.Line += newlineCount;

nextToken.ClearNewlineCount();

Token token = tokenList[countIndex];

Assert.Error(token.Id != CCompiler\_Pre.Tokens.EOF,

Message.Invalid\_end\_of\_macro\_call);

switch (token.Id) {

case CCompiler\_Pre.Tokens.LEFT\_PARENTHESIS:

++level;

subList.Add(token);

break;

case CCompiler\_Pre.Tokens.RIGHT\_PARENTHESIS:

if ((--level) > 0) {

subList.Add(token);

}

break;

default:

if ((level == 1) &&

(token.Id == CCompiler\_Pre.Tokens.COMMA)) {

Assert.Error(subList.Count > 0, name,

Message.Empty\_macro\_parameter);

SearchForMacros(subList, nameStack); // XXX

mainList.Add(subList);

subList = new List<Token>();

}

else {

subList.Add(token);

}

break;

}

if (level == 0) {

Assert.Error(subList.Count > 0, name,

Message.Empty\_macro\_parameter\_list);

mainList.Add(subList);

break;

}

++countIndex;

}

Macro macro = m\_macroMap[name];

Assert.Error(macro.Parameters() == mainList.Count, name,

Message.Invalid\_number\_of\_parameters\_in\_macro\_call);

List<Token> cloneListX = CloneList(macro.TokenList());

for (int macroIndex = (cloneListX.Count - 1);

macroIndex >= 0; --macroIndex) {

Token macroToken = cloneListX[macroIndex];

if (macroToken.Id == CCompiler\_Pre.Tokens.MARK) {

int markIndex = (int) macroToken.Value;

cloneListX.RemoveAt(macroIndex);

List<Token> replaceList = CloneList(mainList[markIndex]);

if ((macroIndex > 0) && (cloneListX[macroIndex - 1].Id ==

CCompiler\_Pre.Tokens.SHARP)) {

string text = "\"" + TokenListToString(replaceList) + "\"";

cloneListX.Insert(macroIndex,

new Token(CCompiler\_Pre.Tokens.STRING, text));

cloneListX.RemoveAt(--macroIndex);

}

else {

cloneListX.InsertRange(macroIndex, replaceList);

}

}

}

nameStack.Push(name);

SearchForMacros(cloneListX, nameStack);

nameStack.Pop();

tokenList.RemoveRange(index, countIndex - index + 1);

tokenList.InsertRange(index, cloneListX);

tokenList[index].AddNewlineCount(beginNewlineCount);

tokenList[index +

cloneListX.Count].AddNewlineCount(totalNewlineCount);

index += cloneListX.Count - 1;

}

else {

Macro macro = m\_macroMap[name];

Assert.Error(macro.Parameters() == 0, name, Message.

Invalid\_number\_of\_parameters\_in\_macro\_call);

List<Token> cloneListX = CloneList(macro.TokenList());

nameStack.Push(name);

SearchForMacros(cloneListX, nameStack);

nameStack.Pop();

tokenList.RemoveAt(index);

tokenList.InsertRange(index, cloneListX);

tokenList[index].AddNewlineCount(beginNewlineCount);

index += cloneListX.Count - 1;

}

}

else if (name.Equals("\_\_STDC\_\_")) {

tokenList[index] =

new Token(CCompiler\_Pre.Tokens.TOKEN, 1, beginNewlineCount);

}

else if (name.Equals("\_\_FILE\_\_")) {

string text = "\"" + CCompiler\_Main.Scanner.Path

.FullName.Replace("\\", "\\\\") + "\"";

tokenList[index] =

new Token(CCompiler\_Pre.Tokens.TOKEN, text, beginNewlineCount);

}

else if (name.Equals("\_\_LINE\_\_")) {

tokenList[index] =

new Token(CCompiler\_Pre.Tokens.TOKEN,

CCompiler\_Main.Scanner.Line, beginNewlineCount);

}

else if (name.Equals("\_\_DATE\_\_")) {

string text = "\"" + DateTime.Now.ToString("MMMM dd yyyy") + "\"";

tokenList[index] =

new Token(CCompiler\_Pre.Tokens.TOKEN, text, beginNewlineCount);

}

else if (name.Equals("\_\_TIME\_\_")) {

string text = "\"" + DateTime.Now.ToString("HH:mm:ss") + "\"";

tokenList[index] =

new Token(CCompiler\_Pre.Tokens.TOKEN, text, beginNewlineCount);

}

}

}

}

The scanParameters extracts the actual parameters of a macro call. The parameters are separated by commas and the list is terminated by a right parenthesis. It becomes a bit complicated since the parameters themselves can hold parenthesis and commas. The field paranCount counts the level of parentheses nesting and only takes parameter into consideration when the nesting level is zero.

The lookupMacro method looks up a given macro. If it stored in MacroMap, the number of parameters is checked and then the parameter locations in the macro body are replaced by the actual parameters.

If the macro name is not stored in MacroMap, it may be one of the predefined special macros. The \_\_STDC\_\_ macro is replaced by the one integer value since the compiler of this book supports standard C.

The \_\_FILE\_\_ macro is replaced by the current path (Main.Path), the \_\_LINE\_\_ macro is replaced by the current line number (Main.Line).

The \_\_DATE\_\_ macro is replaced by the current date on the format “Jan 1, 1970” and the \_\_TIME\_\_ macro is replaced by the current time on the format “01:02:03”.

### String Merging

When the all the macros have been expanded, it is finally time to check whether there are pairs of string constants that need to be merged:

|  |  |  |
| --- | --- | --- |
| "Hello" "World" | "HelloWorld" |  |
| (a) Before | (b) After |  |

private void ConcatTokens(List<Token> tokenList) {

for (int index = 1; index < (tokenList.Count - 1); ++index) {

Token thisToken = tokenList[index];

if (thisToken.Id == CCompiler\_Pre.Tokens.DOUBLE\_SHARP) {

Token prevToken = tokenList[index - 1],

nextToken = tokenList[index + 1];

if ((prevToken.Id == CCompiler\_Pre.Tokens.STRING) ||

(nextToken.Id == CCompiler\_Pre.Tokens.STRING)) {

nextToken.AddNewlineCount(thisToken.GetNewlineCount());

tokenList.RemoveAt(index);

}

else {

prevToken.Value = prevToken.ToString() + nextToken.ToString();

prevToken.AddNewlineCount(thisToken.GetNewlineCount() +

nextToken.GetNewlineCount());

tokenList.RemoveAt(index);

tokenList.RemoveAt(index);

}

}

}

}

private void MergeStrings(List<Token> tokenList) {

for (int index = (tokenList.Count - 2); index >= 0; --index) {

Token thisToken = tokenList[index], nextToken = tokenList[index + 1];

if ((thisToken.Id == CCompiler\_Pre.Tokens.STRING) &&

(nextToken.Id == CCompiler\_Pre.Tokens.STRING)) {

string thisText = thisToken.ToString(),

nextText = nextToken.ToString();

thisToken.Value =

thisText.ToString().Substring(0, thisText.Length - 1) +

nextText.ToString().Substring(1, nextText.Length - 1);

thisToken.AddNewlineCount(nextToken.GetNewlineCount());

tokenList.RemoveAt(index + 1);

}

}

}

}

}

## Register Allocation

The architecture holds a set of registers. On particular feature of the registers is that they to some extent overlaps. More specifically, the a, b, c, and d registers overlaps. Below we see the

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| bits 0-7 | bits 8-15 | bits 16-31 | | bits 32-63 | | | |
| al | ah |  | |  | | | |
| ax | |
| eax | | | |
| rax | | | | | | | |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

1. The Register Set

Below is a description of the register set, where registers of different sizes overlap each other.

|  |  |  |  |
| --- | --- | --- | --- |
| Bits 0-7 | Bits 8-15 | Bits 16-31 | Bits 32-63 |
| al | ah |  |  |
| ax | |
| eax | | |
| rax | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Bits 0-7 | Bits 8-15 | Bits 16-31 | Bits 32-63 |
| bl | bh |  |  |
| bx | |
| ebx | | |
| rbx | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Bits 0-7 | Bits 8-15 | Bits 16-31 | Bits 32-63 |
| cl | ch |  |  |
| cx | |
| ecx | | |
| rcx | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Bits 0-7 | Bits 8-15 | Bits 16-31 | Bits 32-63 |
| dl | ah |  |  |
| dx | |
| edx | | |
| rdx | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Bits 0-7 | Bits 8-15 | Bits 16-31 | Bits 32-63 |
| si | |  |  |
| esi | | |
| rsi | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Bits 0-7 | Bits 8-15 | Bits 16-31 | Bits 32-63 |
| di | |  |  |
| edi | | |
| rdi | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Bits 0-7 | Bits 8-15 | Bits 16-31 | Bits 32-63 |
| bp | |  |  |
| ebp | | |
| rbp | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| Bits 0-7 | Bits 8-15 | Bits 16-31 | Bits 32-63 |
| sp | |  |  |
| esp | | |
| rsp | | | |

1. The C Grammar

This grammar of the C compiler of this book is based on (and to a large extent identical with) the grammar defined in the *American National Standard for Information systems – Programming Language C, X3.159-1989* standard, which is described in the second edition of *The C Programming Language* by Brian W. Kernighan and Dennis M. Ritchie.

## The Preprocessor Grammar

The following grammar is implemented in GPPG in Chapter 1.3.

control\_line ::=

# define identifier token-sequence

| # define identifier( identifier, … , identifier ) token-sequence

| # undef identifier

| # include <filename>

| # include "filename"

| # line constant

| # line constant "filename"

| # error optional\_token\_sequence

| # pragma optional\_token\_sequence

| #

| preprocessor\_conditional

optional\_token\_sequence ::=

/\* empty \*/

| token\_sequence

preprocessor\_conditional ::=

if-line text optional\_elif-parts optional\_else\_line # endif

if-line ::=

# if constant\_expression

| # ifdef identifier

| # ifndef identifier

optional\_elif-parts ::=

/\* empty \*/

| elif-part optional\_elif-parts

elif-part ::=

# elif constant\_expression text

optional\_else\_part ::=

/\* empty \*/

| #else text

## The Language Grammar

The following grammar is implemented in GPPG in Chapter 0. It is the same grammar, written in one block.

*source\_code\_file* ::=

*external*\_*declaration*

| *source\_code\_file* *external*\_*declaration*

*external*\_*declaration* ::=

*function*\_*definition*

| *declaration*

*function*\_*definition* ::=

*direct\_declarator* *optional*\_*declaration*\_*list* **(** *optional*\_*statement*\_*list* **)**

| *declaration*\_*specifier*\_*list* *direct\_declarator* *optional*\_*declaration*\_*list*

**(** *optional*\_*statement*\_*list* **)**

*optional*\_*declaration*\_*list* ::=

/\* *empty* \*/

| *optional*\_*declaration*\_*list* *declaration*

*declaration* ::=

*declaration*\_*specifier*\_*list* **;**

| *declaration*\_*specifier*\_*list* *declarator\_list* **;**

*declaration*\_*specifier*\_*list* ::=

*declaration*\_*specifier*

| *declaration*\_*specifier*\_*list* *declaration*\_*specifier*

*declaration*\_*specifier* ::=

**const** | **volatile** | **auto** | **register** | **static** | **extern** | **typedef** | **void**

| **char** | **wchar\_t** | **short** | **int** | **long** | **float** | **double** | **signed** | **unsigned**

| *struct\_union*\_*specifier |* *enum*\_*specifier* | *typedef*\_*name*

*struct\_union*\_*specifier* ::=

struct\_union *optional*\_*identifier* **(** *declaration*\_*list* **)**

| *struct\_union* **identifier**

*struct\_union ::=*

**struct** | **union**

*optional*\_*identifier* ::=

/\* *empty* \*/

| **identifier**

*declaration*\_*list* ::=

*declaration*

| *declaration*\_*list* *declaration*

*enum*\_*specifier* ::=

**enum** *optional*\_*identifier* **(** *enum*\_*list* **)**

| **enum** **identifier**

*enum*\_*list* ::=

*enum*

| *enum*\_*list* **,** *enum*

*enum* ::=

**identifier**

| **identifier** **=** *const*\_*expression*

*declarator\_list* ::=

*initialization\_bitfield\_direct\_declarator*

| *declarator\_list* **,** *initialization\_bitfield\_direct\_declarator*

*initialization\_bitfield\_direct\_declarator* ::=

*direct\_declarator*

| *direct\_declarator* **=** *initializer*

| *direct\_declarator* **:** *const*\_*expression*

| **:** *const*\_*expression*

*direct\_declarator* ::=

*optional*\_*pointer*\_*list* *direct*\_*direct\_declarator*

*direct*\_*direct\_declarator* ::=

**identifier**

| **(** *direct\_declarator* **)**

| *direct*\_*direct\_declarator* **[** *optional*\_*const*\_*expression* **]**

| *direct*\_*direct\_declarator* **(** *parameter*\_*variadic*\_*list* **)**

| *direct*\_*direct\_declarator* **(** *optional*\_*identifier*\_*list* **)**

*optional*\_*pointer*\_*list* ::=

/\* *empty* \*/

| *pointer*\_*list*

*pointer*\_*list* ::=

*pointer*

| *pointer*\_*list* *pointer*

*pointer* ::=

**\*** | **\*** *declaration*\_*specifier*\_*list*

*parameter*\_*variadic*\_*list* ::=

*parameter*\_*list*

| *parameter*\_*list* **,** **...**

*parameter*\_*list* ::=

*parameter*\_*declaration*

| *parameter*\_*list* **,** *parameter*\_*declaration*

*parameter*\_*declaration* ::=

*declaration*\_*specifier*\_*list*

| *declaration*\_*specifier*\_*list* *direct\_declarator*

| *declaration*\_*specifier*\_*list* *abstract*\_*direct\_declarator*

*optional*\_*identifier*\_*list* ::=

/\* *empty* \*/

| *identifier*\_*list*

*identifier*\_*list* ::=

**identifier**

| *identifier*\_*list* **,** **identifier**

*initializer*\_*list* ::=

*initializer*

| *initializer*\_*list* **,** *initializer*

*initializer* ::=

*assignment*\_*expression*

| **(** *initializer*\_*list* **)**

| **(** *initializer*\_*list* **,** **)**

*type*\_*name* ::=

*declaration*\_*specifier*\_*list*

| *declaration*\_*specifier*\_*list* *abstract*\_*direct\_declarator*

*abstract*\_*direct\_declarator* ::=

*pointer*\_*list*

| *optional*\_*pointer*\_*list* *direct*\_*abstract*\_*direct\_declarator*

*direct*\_*abstract*\_*direct\_declarator* ::=

**(** *abstract*\_*direct\_declarator* **)**

| **[** *optional*\_*const*\_*expression* **]**

| *direct*\_*abstract*\_*direct\_declarator* **[** *optional*\_*const*\_*expression* **]**

| **(** *optional*\_*parameter*\_*variadic*\_*list* **)**

| *direct*\_*abstract*\_*direct\_declarator* **(** *optional*\_*parameter*\_*variadic*\_*list* **)**

*optional*\_*statement*\_*list* ::=

/\* *empty* \*/

| *statement*\_*list*

*statement*\_*list* ::=

*statement*

| *statement*\_*list* *statement*

*statement* ::=

*open*\_*statement*

| *closed*\_*statement*

*open*\_*statement* ::=

**if** **(** *expression* **)** *statement*

| **if** **(** *expression* **)** *closed*\_*statement* **else** *open*\_*statement*

| **switch** **(** *expression* **)** *open*\_*statement*

| **case** *const*\_*expression* **:** *open*\_*statement*

| **while** **(** *expression* **)** *open*\_*statement*

| **for** **(** *optional*\_*expression* **;** *optional*\_*expression* **;**

*optional*\_*expression* **)** *open*\_*statement*

| **identifier** **:** *open*\_*statement*

*closed*\_*statement* ::=

**if** **(** *expression* **)** *closed*\_*statement* **else** *closed*\_*statement*

| **switch** **(** *expression* **)** *closed*\_*statement*

| **while** **(** *expression* **)** *closed*\_*statement*

| **for** **(** *optional*\_*expression* **;** *optional*\_*expression* **;**

*optional*\_*expression* **)** *closed*\_*statement*

| *declaration*

| **case** *const*\_*expression* **:** *closed*\_*statement*

| **default** **:** *closed*\_*statement*

| *optional*\_*expression* **;**

| **(** *optional*\_*statement*\_*list* **)**

| **do** *statement* **while** **(** *expression* **)** **;**

| **goto** **identifier ;**

| **continue ;**

| **break ;**

| **return** *optional*\_*expression* **;**

| **load\_register** **(** **identifier ,** *expression* **) ;**

| **store\_register** **(** **identifier** **,** *expression* **) ;**

| **store\_flagbyte** **(** *expression* **) ;**

| **jump\_register** **(** **identifier** **) ;**

| **interupt** **(** *const*\_*expression* **) ;**

*optional*\_*expression* ::=

/\* *empty* \*/

| *expression*

*expression* ::=

*assignment*\_*expression*

| *expression* **,** *assignment*\_*expression*

*assignment*\_*expression* ::=

*conditional*\_*expression*

| *prefix*\_*expression* *assignment*\_*operator* *assignment*\_*expression*

*assignment*\_*operator* ::=

**=** | \***=** | /**=** | %**=** | +**=** | -**=** | <<**=** | <<**=** | &**=** | |**=** | ^**=**

*conditional*\_*expression* ::=

*logical*\_*or*\_*expression*

| *logical*\_*or*\_*expression*

**?** *expression* **:** *conditional*\_*expression*

*optional*\_*const*\_*expression* ::=

/\* *empty* \*/

| *const*\_*expression*

*const*\_*expression* ::=

*conditional*\_*expression*

*logical*\_*or*\_*expression* ::=

*logical*\_*and*\_*expression*

| *logical*\_*or*\_*expression* *||* *logical*\_*and*\_*expression*

*logical*\_*and*\_*expression* ::=

*bitwise\_or\_expression*

| *bitwise*\_*and*\_*expression* **&&** *bitwise\_or\_expression*

*bitwise\_or\_expression* ::=

*bitwise*\_*xor*\_*expression*

| *bitwise\_or\_expression* **|** *bitwise*\_*xor*\_*expression*

*bitwise*\_*xor*\_*expression* ::=

*bitwise*\_*and*\_*expression*

| *bitwise*\_*xor*\_*expression* **^** *bitwise*\_*and*\_*expression*

*bitwise*\_*and*\_*expression* ::=

*equality*\_*expression*

| *bitwise*\_*and*\_*expression* **&** *equality*\_*expression*

*equality*\_*expression* ::=

*relation*\_*expression*

| *equality*\_*expression* *equality*\_*operator* *relation*\_*expression*

*equality*\_*operator* ::=

**==** | **!=**

*relation*\_*expression* ::=

*shift*\_*expression*

| *relation*\_*expression* *relation*\_*operator* *shift*\_*expression*

*relation*\_*operator* ::=

**<** | **<=** | **>** | **>=**

*shift*\_*expression* ::=

*add*\_*expression*

| *shift*\_*expression* *shift*\_*operator* *add*\_*expression*

*shift*\_*operator* ::=

**<<** | **>>**

*add*\_*expression* ::=

*multiply*\_*expression*

| *add*\_*expression* *binary*\_*add*\_*operator* *multiply*\_*expression*

*binary*\_*add*\_*operator* ::=

**+** | **-**

*multiply*\_*expression* ::=

*type\_cast\_expression*

| *multiply*\_*expression* *multiply*\_*operator* *type\_cast\_expression*

*multiply*\_*operator* ::=

**\*** | **/** | **%**

*type\_cast\_expression* ::=

*prefix*\_*expression*

| **(** *type*\_*name* **)** *type\_cast\_expression*

*prefix*\_*expression* ::=

*postfix*\_*expression*

| *increment*\_*operator* *prefix*\_*expression*

| *prefix*\_*add*\_*operator* *type\_cast\_expression*

| **!** *type\_cast\_expression*

| **~** *type\_cast\_expression*

| **&** *type\_cast\_expression*

| **\*** *type\_cast\_expression*

| **sizeof** *prefix*\_*expression*

| **sizeof** **(** *type*\_*name* **)**

*prefix*\_*add*\_*operator* ::=

**+** | **-**

*increment*\_*operator* ::=

**++** | **--**

*postfix*\_*expression* ::=

*primary*\_*expression*

| *postfix*\_*expression* **[** *expression* **]**

| *postfix*\_*expression* **(** *optional*\_*argument*\_*expression*\_*list* **)**

| *postfix*\_*expression* **.** **identifier**

| *postfix*\_*expression* **->** **identifier**

| *postfix*\_*expression* *increment*\_*operator*

*primary*\_*expression* ::=

**identifier**

| **value**

| **(** *expression* **)**

*optional*\_*argument*\_*expression*\_*list* ::=

/\* *empty* \*/

| *argument*\_*expression*\_*list*

*argument*\_*expression*\_*list* ::=

*assignment*\_*expression*

| *argument*\_*expression*\_*list* , *assignment*\_*expression*

1. The Gardens Point Tools

In this chapter, we use the Gardens Point LEX (GPLEX) scanner generator and the Gardens Point Parser Generator (GPPG) to construct a parser and scanner for a simple demonstration programming language.

### The Language

A program is made up by a non-empty sequence of statements, each terminated by a semicolon. There are five kinds of statements:

Read. A non-empty sequence of variables is assigned values read from the input stream, with an optional prompt.

Write. The non-empty sequence of values is written to the output stream, with an optional text.

Assign. A variable is assigned the value of an expression.

The mathematical constants pi and e are stored in the variable map when the execution starts. An expression can be made up of the four rules of arithmetic, parentheses as well as the functions sin, cos, tan, log, exp, log10, exp10, and sqrt (square root). Division by zero, invalid function arguments, or invalid input generate error messages.

Below is an example:

// The area and circumference of a circle

read "Input the radius of a circle: ", radius;

assign area = pi \* radius \* radius;

assign circumference = 2 \* pi \* radius;

write "The area and circumference are: ", area, circumference;

newline;

// Fahrenheit to Celsius

read "Input a temperature in degrees Fahrenheit: ", fahrenheit;

assign celsius = (fahrenheit - 32) / 1.8;

write "In Celsius degrees: ", celsius;

newline;

// Pythagorean Theorem

read "Input the two cathetuses of a right-angled triangle: ", a, b;

assign c = sqrt(a \* a + b \* b);

write "The hypothenuse is: ", c;

When executed (the underlined text represent input values):

Input the radius of a circle: 10

The area and circumference are: 314.159265358979, 62.8318530717959

Input a temperature in degrees Fahrenheit: 212

In Celsius degrees: 100

Input the two cathetuses of a right-angled triangle: 3,4

The hypothenuse is: 5

### The Grammar

Every programming language has a syntax, which may be defined by a grammar. The grammar of our language is given below. A grammar is made up by a set of *rules*, among which one is the start rule (in this case statement\_list). The grammar rules for our language follows below, the vertical bar at the left of the rules can be read as “or”. The first rule statement\_list can be read as “a statement list is a statement or a statement followed by a statement list.”

statement\_list ::=

statement

| statement statement\_list

statement ::=

**assign** **identifier** EQUAL expression **;**

| **read** optional\_output identifier\_list **;**

| **write** optional\_output expression\_list **;**

| **newline** **;**

optional\_output ::=

/\* Empty. \*/

| **string** **,**

identifier\_list ::=

**identifier**

| **identifier** **,** identifier\_list

expression\_list ::=

expression

| expression **,** expression\_list

expression ::=

binary\_expression

| expression **+** binary\_expression

| expression **-** binary\_expression

binary\_expression ::=

unary\_expression

| binary\_expression **\*** unary\_expression

| binary\_expression **/** unary\_expression

unary\_expression ::=

primary\_expression

| **+** unary\_expression

| **-** unary\_expression

| **sin** unary\_expression

| **cos** unary\_expression

| **tan** unary\_expression

| **log** unary\_expression

| **exp** unary\_expression

| **log10** unary\_expression

| **exp10** unary\_expression

| **sqrt** unary\_expression

primary\_expression ::=

**(** expression **)**

| **identifier**

| **value**

The words and character in boldface is the end products of the grammar. The rules shall be applied in such a way that finally only tokens remain. The first line of the example in the previous section are made up of the following tokens:

**read string , identifier ;**

### GPPG

GPPG is a tool based on the classic parser generator Yacc. The name Yacc is an abbreviation for Yet Another Compiler-Compiler. However, it is pronounced as the animal yak. Even though it is possible to write a parser in pure C#, it is easier to use a tool like GPPG. Simply put, GPPG takes a grammar as input and generators C# code as output. However, it does not only check whether the input string comply with the grammar, each rule is also allowed to perform actions, defined by C# code.

The parser is divided into two parts: one part stating the rules and tokens of the grammar, and one part defining the grammar rules. The tokens written in boldface in the previous section are written in capital letters in GPPG.

Parser.gppg

%namespace Calculator

%partial

%{

// Empty.

%}

The following tokens do not hold attributes.

%token ASSIGN READ WRITE NEWLINE EQUAL PLUS MINUS TIMES DIVIDE SIN COS TAN

LOG EXP LOG10 EXP10 SQRT LEFT\_PAREN RIGHT\_PAREN COMMA SEMICOLON

The STRING, IDENTIFIER, and VALUE tokens have attributes.

%union {

public string name;

public string text;

public double value;

public List<string> nameList;

public List<double> valueList;

}

%token <text> TEXT

%token <name> NAME

%token <value> VALUE

%type <nameList> name\_list

%type <valueList> expression\_list

%type <value> expression binary\_expression unary\_expression primary\_expression

There are also the rules statement\_list, statement, and optional\_output in the rules section below. However, they do not return values and do not need to be defined in this section.

%%

statement\_list:

statement

| statement statement\_list;

The right-hand side are numbered from one, and the attribute values are accessible with the dollar-notation. In the rule below, the value of the identifier is stored in $2 and the value of the expression is stored in $4.

statement:

ASSIGN NAME EQUAL expression SEMICOLON {

MainX.VariableMap[$2] = $4;

}

| READ optional\_output name\_list SEMICOLON {

try {

string buffer = Console.In.ReadLine();

string[] textArray = buffer.Split(',');

if ($3.Count != textArray.Length) {

Console.Error.WriteLine("Invalid number of values.");

Environment.Exit(-1);

}

for (int index = 0; index < $3.Count; ++index) {

string name = $3[index], text = textArray[index];

double value = double.Parse(text);

MainX.VariableMap[name] = value;

}

}

catch (Exception exception) {

Console.Error.WriteLine("Invalid input: " + exception.ToString());

Environment.Exit(-1);

}

}

| WRITE optional\_output expression\_list SEMICOLON {

bool first = true;

foreach (double value in $3) {

Console.Out.Write((first ? "" : ", ") + value);

first = false;

}

Console.Out.WriteLine();

}

| NEWLINE SEMICOLON {

Console.Out.WriteLine();

};

optional\_output:

/\* Empty. \*/

| TEXT COMMA {

Console.Out.Write($1);

};

The value of the rule is stored in $$. However, the $$ value is always a reference to an Object. Therefore, we need to transform it into a list before we can call the add method.

name\_list:

NAME {

$$ = new List<string>();

$$.Add($1);

}

| name\_list COMMA NAME {

$$ = $1;

$$.Add($3);

};

expression\_list:

expression {

$$ = new List<double>();

$$.Add($1);

}

| expression\_list COMMA expression {

$$ = $1;

$$.Add($3);

};

expression:

binary\_expression {

$$ = $1;

}

| expression PLUS binary\_expression {

$$ = $1 + $3;

}

| expression MINUS binary\_expression {

$$ = $1 - $3;

};

binary\_expression:

unary\_expression {

$$ = $1;

}

| binary\_expression TIMES unary\_expression {

$$ = $1 \* $3;

}

| binary\_expression DIVIDE unary\_expression {

if ($3 == 0) {

Console.Error.WriteLine("Division by Zero.");

Environment.Exit(-1);

}

$$ = $1 / $3;

};

unary\_expression:

primary\_expression {

$$ = $1;

}

| PLUS unary\_expression {

$$ = $2;

}

| MINUS unary\_expression {

$$ = -$2;

}

| SIN unary\_expression {

$$ = Math.Sin($2);

}

| COS unary\_expression {

$$ = Math.Cos($2);

}

| TAN unary\_expression {

$$ = Math.Tan($2);

}

| LOG unary\_expression {

if ($2 <= 0) {

Console.Error.WriteLine("Logarithm of Non-Positive Value.");

Environment.Exit(-1);

}

$$ = Math.Log($2);

}

| EXP unary\_expression {

$$ = Math.Exp($2);

}

| LOG10 unary\_expression {

if ($2 <= 0) {

Console.Error.WriteLine("Logarithm of Non-Positive Value.");

Environment.Exit(-1);

}

$$ = Math.Log10($2);

}

| EXP10 unary\_expression {

$$ = Math.Pow(10, $2);

}

| SQRT unary\_expression {

if ($2 < 0) {

Console.Error.WriteLine("Square Root of Negativ Value.");

Environment.Exit(-1);

}

$$ = Math.Sqrt($2);

};

primary\_expression:

LEFT\_PAREN expression RIGHT\_PAREN {

$$ = $2;

}

| NAME {

if (MainX.VariableMap.ContainsKey($1)) {

$$ = MainX.VariableMap[$1];

}

else {

Console.Error.WriteLine("Unknown Name: \"" + $1 + "\".");

Environment.Exit(-1);

}

}

| VALUE {

$$ = $1;

};

%%

### JPlex

For the parser of the previous section to work properly it also needs a scanner that identifies the tokens. JPlex is a scanner-generator, which task is to identify and transform groups of characters into tokens.

When it comes to complicated tokens, JFlex applies regular expressions. Some characters have special meanings in accordance with the following table. However, their special status can be canceled by preceeding it with a backslash.

|  |  |
| --- | --- |
| **Character** | **Meaning** |
| ( and ) | Grouping subexpressions |
| [ and ] | Any character within the brackets |
| [^ and ] | Any character except the characters within the brackets |
| \* | Zero or more characters |
| + | One or moree characters |
| . | Any character except new line |

Scanner.jplex

%namespace Calculator

%{

// Empty.

%}

In the scanner code below, a line comment is two slashes followed by zero or more occurrences of any character except newline. Since both the slash has special meaning we must insert a backslash before it for it to mean just a slash.

LINE\_COMMENT \/\/.\*

A string is a double quotation mark, followed by zero or more occurrences of any character (except double quotation mark) followed by a double quotation mark.

TEXT \"[^\"]\*\"

An identifier is a capital or small letter, or a underline, followed by zero or more capital or small letters, or digits, or underlines.

NAME [a-zA-Z\_][a-zA-Z0-9\_]\*

A value is as least one digit, potentially followed by more digits, a dot, and even more digits.

VALUE [0-9]+|([0-9]+\.[0-9]+)

A white space is either a space, a tabulator, a return, a newline, or a form-feed.

WHITE\_SPACE [ \t\r\n\f]

%%

"assign" { return ((int) Tokens.ASSIGN); }

"read" { return ((int) Tokens.READ); }

"write" { return ((int) Tokens.WRITE); }

"newline" { return ((int) Tokens.NEWLINE); }

"sin" { return ((int) Tokens.SIN); }

"cos" { return ((int) Tokens.COS); }

"tan" { return ((int) Tokens.TAN); }

"log" { return ((int) Tokens.LOG); }

"exp" { return ((int) Tokens.EXP); }

"log10" { return ((int) Tokens.LOG10); }

"exp10" { return ((int) Tokens.EXP10); }

"sqrt" { return ((int) Tokens.SQRT); }

"=" { return ((int) Tokens.EQUAL); }

"+" { return ((int) Tokens.PLUS); }

"-" { return ((int) Tokens.MINUS); }

"\*" { return ((int) Tokens.TIMES); }

"/" { return ((int) Tokens.DIVIDE); }

"(" { return ((int) Tokens.LEFT\_PAREN); }

")" { return ((int) Tokens.RIGHT\_PAREN); }

"," { return ((int) Tokens.COMMA); }

";" { return ((int) Tokens.SEMICOLON); }

The yytext method return the scanned text, which is useful with identifiers and strings.

{NAME} {

yylval.name = yytext;

return ((int) Tokens.NAME);

}

{TEXT} {

yylval.text = yytext.Substring(1, yytext.Length - 2);

return ((int) Tokens.TEXT);

}

{VALUE} {

yylval.value = Double.Parse(yytext);

return ((int) Tokens.VALUE);

}

{LINE\_COMMENT} {

// Empty.

}

{WHITE\_SPACE} {

// Empty.

}

<<EOF>> {

return ((int) Tokens.EOF);

}

Finally, the last part of the scanner is the dot rule, which catch the case not caught by any of the previous tokens (equivalent to default in a switch statement). In that case, we just exit the execution with an error message.

. {

Console.Error.WriteLine("Unknown character: \'" + yytext + "\'.");

Environment.Exit(-1);

}

### Main

GPPG and JPLEX generates the Parser and Scanner classes. The main class creates a scanner object which is used by the parser object. The parse method is called with invoke the parsing. The input is read from BufferedReader, the output is written to OutputStream, error messages are written to ErrorStream, and VariableMap is used to keep track of the variables read and assigned by the program.

Main.cs

using System;

using System.IO;

using System.Collections.Generic;

namespace Calculator {

class MainX {

public static Dictionary<string,double> VariableMap =

new Dictionary<string,double>();

static void Main(string[] args) {

if (args.Length != 1) {

Console.Out.WriteLine("Usage: calculator inputfile");

}

{ System.Globalization.CultureInfo customCulture =

(System.Globalization.CultureInfo)

System.Threading.Thread.CurrentThread.CurrentCulture.Clone();

customCulture.NumberFormat.NumberDecimalSeparator = ".";

System.Threading.Thread.CurrentThread.CurrentCulture = customCulture;

}

VariableMap["pi"] = Math.PI;

VariableMap["e"] = Math.E;

string s = "abcd", t = "\"\"";

string u = s.Substring(1, s.Length - 2);

string v = s.Substring(1, t.Length - 2);

try {

FileStream file = new FileStream(args[0], FileMode.Open);

Scanner scanner = new Scanner(file);

Parser parser = new Parser(scanner);

parser.Parse();

}

catch (Exception exception) {

Console.Out.WriteLine(exception.ToString());

}

}

}

The Parser class is necessary to make the parser work with the scanner. It works as link between the scanner and the parser

public partial class Parser :

QUT.Gppg.ShiftReduceParser<ValueType, QUT.Gppg.LexLocation> {

public Parser(Scanner scanner)

:base(scanner) {

// Empty.

}

}

}

Finally, below follows a small script file that generates the C# source code for the parser and scanner.

Parser.bat

@C:

@cd C:\Users\Stefan\Documents\Calculator\_CS

@"C:\gppg-distro-1\_5\_2\binaries\Gppg" /gplex Parser.gppg > Parser.cs

@"C:\gppg-distro-1\_5\_2\binaries\Gplex" Scanner.gplex

1. Auxiliary Classes

C# comes with a large class library, with functionality for almost all requirements. However, there are some classes that are missing, which we need to write ourselves: a class for error handling, container classes, and a graph class.

## Error Handling

To begin with, we need error handling, a way to notify the programmer of errors and warnings. In the C Standard Library there is a macro called assert, and in this application, we imitate that macro by defining a class with the corresponding behavior. It has the two method sets error and warning, that writes a message. The difference between them is that error stops the execution.

Assert.cs

using System;

namespace CCompiler {

public class Assert {

public static void ErrorXXX(bool test) {

if (!test) {

Error(null, null);

}

}

public static void Error(string message) {

Error(message, null);

}

public static void Error(bool test, Message message) {

if (!test) {

Error(message, null);

}

}

public static void Error(object value, Message message) {

Error(false, value, message);

}

public static void Error(bool test, object value, Message message) {

if (!test) {

Error(message, value.ToString());

}

}

public static void Error(Message message) {

Error(message, null);

}

private static void Error(Message message, string text) {

Error(Enum.GetName(typeof(Message), message).

Replace("\_\_\_", ",").Replace("\_\_", "-").

Replace("\_", " "), text);

}

private static void Error(string message, string text) {

Message("Error", message, text);

Console.In.ReadLine();

System.Environment.Exit(-1);

}

private static void Message(string type, string message,

string text) {

string funcText;

if (SymbolTable.CurrentFunction != null) {

funcText = " in function " + SymbolTable.CurrentFunction.UniqueName;

}

else {

funcText = " in global space";

}

string extraText = (text != null) ? (": " + text) : "";

if ((message != null) &&

(CCompiler\_Main.Scanner.Path != null)) {

Console.Error.WriteLine(type + " at line " +

CCompiler\_Main.Scanner.Line + funcText +

" in file " + CCompiler\_Main.Scanner.Path.Name +

". " + message + extraText + ".");

}

else if ((message == null) &&

(CCompiler\_Main.Scanner.Path != null)) {

Console.Error.WriteLine(type + " at line " +

CCompiler\_Main.Scanner.Line + funcText +

" in file " + CCompiler\_Main.Scanner.Path.Name +

extraText + ".");

}

else if ((message != null) &&

(CCompiler\_Main.Scanner.Path == null)) {

Console.Error.WriteLine(type + ". " + message +

extraText + ".");

}

else if ((message == null) &&

(CCompiler\_Main.Scanner.Path == null)) {

Console.Error.WriteLine(type + extraText + ".");

}

}

}

}

Message.cs

namespace CCompiler {

public enum Message {

Keyword\_defined\_twice,

Invalid\_specifier\_sequence,

Invalid\_specifier\_sequence\_together\_with\_type,

Out\_of\_registers,

Unfinished\_block\_comment,

Invalid\_type\_cast,

Newline\_in\_string,

Unfinished\_string,

Parse\_error,

Invalid\_type,

Operator\_size,

Newline\_in\_character,

Unfinished\_character,

Invalid\_hexadecimal\_code,

Invalid\_slash\_sequence,

Double\_sharps\_at\_beginning\_of\_line,

Double\_sharps\_at\_end\_of\_line,

Two\_consecutive\_double\_sharps,

Only\_extern\_or\_static\_storage\_allowed\_for\_functions,,

Only\_auto\_or\_register\_storage\_allowed\_for\_struct\_or\_union\_member,

Function\_missing,

Not\_a\_function,

Function\_main\_must\_return\_void\_or\_integer,

Undefined\_parameter\_in\_old\_\_style\_function\_definitializerion,

Unmatched\_number\_of\_parameters\_in\_old\_\_style\_function\_definitializerion,

A\_function\_must\_be\_static\_or\_extern,

Invalid\_parameter\_list,

Undefined\_label,

Unreferenced\_label,

Missing\_goto\_address,

Tag\_already\_defined,

Tag\_not\_found,

Duplicate\_symbol,

Unknown\_enumeration,

Not\_an\_enumeration,

Different\_return\_type\_in\_function\_redeclaration,

Mixing\_variadic\_function\_parameter\_in\_redeclaration,

Different\_parameter\_lists\_in\_function\_redeclaration,

Functions\_cannot\_be\_initialized,

Typedef\_cannot\_be\_initialized,

Extern\_cannot\_be\_initialized,

Struct\_or\_union\_field\_cannot\_be\_initialized,

Auto\_or\_register\_storage\_in\_global\_scope,

Name\_already\_defined,

Different\_types\_in\_redeclaration,

Invalid\_tag\_redeclaration,

Case\_without\_switch,

Non\_\_constant\_case\_value,

Repeated\_case\_value,

Default\_without\_switch,

Repeted\_default,

Break\_without\_switch\_\_\_\_while\_\_\_\_do\_\_\_\_or\_\_\_\_for,

Continue\_without\_while\_\_\_\_do\_\_\_\_or\_\_\_\_for,

Unnamed\_function\_definitializerion,

New\_and\_old\_style\_mixed\_function\_definitializerion,

Non\_\_integral\_enum\_value,

Non\_\_constant\_enum\_value,

Bitfields\_only\_allowed\_in\_structs\_or\_unions,

Only\_auto\_or\_register\_storage\_allowed\_in\_struct\_or\_union,

Non\_\_integral\_bits\_expression,

Bits\_value\_out\_of\_range,

Non\_\_positive\_array\_size,

Non\_\_constant\_expression,

Not\_constant\_or\_static\_expression,

An\_variadic\_function\_must\_have\_at\_least\_one\_parameter,

A\_void\_parameter\_cannot\_be\_named,

An\_variadic\_function\_cannot\_have\_a\_void\_parameter,

Invalid\_void\_parameter,

Parameters\_must\_have\_auto\_or\_register\_storage,

If\_\_\_ifdef\_\_\_\_or\_ifndef\_directive\_without\_matching\_endif,

Array\_of\_incomplete\_type\_not\_allowed,

Array\_of\_function\_not\_allowed,

Function\_cannot\_return\_array,

Function\_cannot\_return\_function,

Duplicate\_name\_in\_parameter\_list,

Syntax\_error,

Invalid\_octal\_sequence,

Invalid\_char\_sequence,

Non\_\_integral\_expression,

Not\_assignable,

Invalid\_type\_in\_bitwise\_expression,

Invalid\_type\_in\_shift\_expression,

Pointer\_to\_void,

Invalid\_type\_in\_expression,

Invalid\_types\_in\_addition\_expression,

Invalid\_types\_in\_subtraction\_expression,

Non\_\_arithmetic\_expression,

Register\_storage\_not\_allowed\_in\_sizof\_expression,

Sizeof\_applied\_to\_function\_not\_allowed,

Sizeof\_applied\_to\_bitfield\_not\_allowed,

Not\_addressable,

Invalid\_address\_of\_register\_storage,

Invalid\_dereference\_of\_non\_\_pointer,

Not\_a\_pointer\_in\_arrow\_expression,

Not\_a\_pointer\_to\_a\_struct\_or\_union\_in\_arrow\_expression,

Unknown\_member\_in\_arrow\_expression,

Invalid\_type\_in\_increment\_expression,

Not\_a\_struct\_or\_union\_in\_dot\_expression,

Member\_access\_of\_uncomplete\_struct\_or\_union,

Unknown\_member\_in\_dot\_expression,

Invalid\_type\_in\_index\_expression,

Too\_few\_actual\_parameters\_in\_function\_call,

Too\_many\_parameters\_in\_function\_call,

Unknown\_name,

Unknown\_name\_\_\_\_assuming\_function\_returning\_int,

Too\_many\_initializers,

Duplicate\_global\_name,

Missing\_external\_function,

Reached\_the\_end\_of\_a\_non\_\_void\_function,

Floating\_stack\_overflow,

Unbalanced\_if\_and\_endif\_directive\_structure,

Invalid\_line\_number,

Invalid\_preprocessor\_directive,

Repeted\_include\_statement,

Defined\_twice,

Non\_\_void\_return\_from\_void\_function,

Void\_returned\_from\_non\_\_void\_function,

Invalid\_define\_directive,

Invalid\_macro\_definitializerion,

Repeated\_macro\_parameter,

Mixing\_signed\_and\_unsigned\_types,

Invalid\_macro\_redefinitializerion,

Invalid\_undef\_directive,

Macro\_not\_defined,

Preprocessor\_parser,

Elif\_directive\_without\_preceeding\_if\_\_\_\_ifdef\_\_\_\_or\_ifndef\_directive,

Elif\_directive\_following\_else\_directive,

Else\_directive\_without\_preceeding\_if\_\_\_\_ifdef\_\_\_\_or\_ifndef\_directive,

Else\_directive\_after\_else\_directive,

Endif\_directive\_without\_preceeding\_if\_\_\_\_ifdef\_\_\_\_or\_ifndef\_directive,

Invalid\_end\_of\_macro\_call,

Empty\_macro\_parameter,

Empty\_macro\_parameter\_list,

Invalid\_number\_of\_parameters\_in\_macro\_call,

Unknown\_character,

Function\_missing\_in\_linking,

Object\_missing\_in\_linking,

Non\_\_static\_initializer,

Unnamed\_parameter,

Unknown\_register,

Extern\_enumeration\_item\_cannot\_be\_initialized,

Only\_array\_struct\_or\_union\_can\_be\_initialized\_by\_a\_list,

Unmatched\_register\_size,

Value\_overflow,

Invalid\_expression,

String\_does\_not\_fit\_in\_array,

Only\_auto\_or\_register\_storage\_allowed\_in\_parameter\_declaration,

Only\_auto\_or\_register\_storage\_allowed\_for\_struct\_or\_union\_scope,

Only\_extern\_\_\_\_static\_\_\_\_or\_typedef\_storage\_allowed\_in\_global\_scope

};

}

## Container Classes

C# has a large class library holding many container classes. However, there are no classes for pairs or ifElseChains or.

### Ordered Pair

In the preprocessor (and in several other classes in the following chapters) we also need the auxiliary class Pair.

Pair.cs

namespace CCompiler {

public class Pair<FirstType,SecondType> {

private FirstType m\_first;

private SecondType m\_second;

public Pair(FirstType first, SecondType second) {

m\_first = first;

m\_second = second;

}

public FirstType First {

get { return m\_first; }

set { m\_first = value; }

}

public SecondType Second {

get { return m\_second; }

set { m\_second = value; }

}

public override int GetHashCode() {

return m\_first.GetHashCode() + m\_second.GetHashCode();

}

public override bool Equals(object obj) {

if (obj is Pair<FirstType,SecondType>) {

Pair<FirstType, SecondType> pair = (Pair<FirstType, SecondType>) obj;

return m\_first.Equals(pair.m\_first) && m\_second.Equals(m\_second);

}

return false;

}

}

}

## Graph

Graph implements a general unordered graph. It holds of set of vertices and a set of edges.

Graph.cs

using System.Collections.Generic;

namespace CCompiler {

public class Graph<VertexType> {

private ISet<VertexType> m\_vertexSet;

private ISet<Pair<VertexType,VertexType>> m\_edgeSet;

public Graph() {

m\_vertexSet = new HashSet<VertexType>();

m\_edgeSet = new HashSet<Pair<VertexType,VertexType>>();

}

public Graph(ISet<VertexType> vertexSet) {

m\_vertexSet = vertexSet;

m\_edgeSet = new HashSet<Pair<VertexType,VertexType>>();

}

public Graph(ISet<VertexType> vertexSet,

ISet<Pair<VertexType,VertexType>> edgeSet) {

m\_vertexSet = vertexSet;

m\_edgeSet = edgeSet;

}

public ISet<VertexType> VertexSet {

get { return m\_vertexSet; }

}

public ISet<Pair<VertexType,VertexType>> EdgeSet {

get { return m\_edgeSet; }

}

The neighbourSet method goes through all edges and add each found neighbor to the vertex.

public ISet<VertexType> GetNeighbourSet(VertexType vertex) {

ISet<VertexType> neighbourSet = new HashSet<VertexType>();

foreach (Pair<VertexType,VertexType> edge in m\_edgeSet) {

if (edge.First.Equals(vertex)) {

neighbourSet.Add(edge.Second);

}

if (edge.Second.Equals(vertex)) {

neighbourSet.Add(edge.First);

}

}

return neighbourSet;

}

### Addition and Removal of Vertices and Edges

public void AddVertex(VertexType vertex) {

m\_vertexSet.Add(vertex);

}

public void EraseVertex(VertexType vertex) {

ISet<Pair<VertexType,VertexType>> edgeSetCopy =

new HashSet<Pair<VertexType,VertexType>>(m\_edgeSet);

foreach (Pair<VertexType,VertexType> edge in edgeSetCopy) {

if ((vertex.Equals(edge.First)) || (vertex.Equals(edge.Second))) {

m\_edgeSet.Remove(edge);

}

}

m\_vertexSet.Remove(vertex);

}

public void AddEdge(VertexType vertex1, VertexType vertex2) {

Pair<VertexType,VertexType> edge =

new Pair<VertexType,VertexType>(vertex1, vertex2);

m\_edgeSet.Add(edge);

}

public void EraseEdge(VertexType vertex1, VertexType vertex2) {

Pair<VertexType,VertexType> edge =

new Pair<VertexType,VertexType>(vertex1, vertex2);

m\_edgeSet.Remove(edge);

}

### Graph Partition

The method partitionate divides the graph into free subgraphs; that is, subgraphs which vertices have no neighbors in any of the other free subgraphs. First we go through the vertices and perform a deep search to find all vertices reachable from the vertex. Then we generate a subgraph for each such vertex set.

public ISet<Graph<VertexType>> Split() {

ISet<ISet<VertexType>> subgraphSet = new HashSet<ISet<VertexType>>();

foreach (VertexType vertex in m\_vertexSet) {

ISet<VertexType> vertexSet = new HashSet<VertexType>();

DeepSearch(vertex, vertexSet);

subgraphSet.Add(vertexSet);

}

ISet<Graph<VertexType>> graphSet = new HashSet<Graph<VertexType>>();

foreach (ISet<VertexType> vertexSet in subgraphSet) {

graphSet.Add(InducedSubGraph(vertexSet));

}

return graphSet;

}

The DeepFirstSearch method search recursively through the graph in order to find all vertices reachable from the given vertex. To avoid cyclic search, the search is terminated if the vertex is already a member of the result set.

private void DeepSearch(VertexType vertex, ISet<VertexType> resultSet) {

if (!resultSet.Contains(vertex)) {

resultSet.Add(vertex);

ISet<VertexType> neighbourSet = GetNeighbourSet(vertex);

foreach (VertexType neighbour in neighbourSet) {

DeepSearch(neighbour, resultSet);

}

}

}

The generateSubGraph method goes through the edge set and add all edges which both end vertices are members of the vertex set.

private Graph<VertexType> InducedSubGraph(ISet<VertexType> vertexSet) {

ISet<Pair<VertexType,VertexType>> resultEdgeSet =

new HashSet<Pair<VertexType,VertexType>>();

foreach (Pair<VertexType,VertexType> edge in m\_edgeSet) {

if (vertexSet.Contains(edge.First) &&

vertexSet.Contains(edge.Second)) {

resultEdgeSet.Add(edge);

}

}

return (new Graph<VertexType>(vertexSet, resultEdgeSet));

}

}

}

1. The ASCII Table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | *nul \0* | 32 | *blank* | 64 | @ É | 96 | ` é |
| 1 | *soh* | 33 | ! | 65 | A | 97 | a |
| 2 | *stx* | 34 | " | 66 | B | 98 | b |
| 3 | *etx* | 35 | # | 67 | C | 99 | c |
| 4 | *eot* | 36 | $ | 68 | D | 100 | d |
| 5 | *enq* | 37 | % | 69 | E | 101 | e |
| 6 | *ack* | 38 | & | 70 | F | 102 | f |
| 7 | *bel \a* | 39 | ' | 71 | G | 103 | g |
| 8 | *bs \b* | 40 | ( | 72 | H | 104 | h |
| 9 | *ht \t* | 41 | ) | 73 | I | 105 | i |
| 10 | *lf \n* | 42 | \* | 74 | J | 106 | j |
| 11 | *vt \vt* | 43 | + | 75 | K | 107 | k |
| 12 | *ff \f* | 44 | , | 76 | L | 108 | l |
| 13 | *cr \r* | 45 | - | 77 | M | 109 | m |
| 14 | *soh* | 46 | . | 78 | N | 110 | n |
| 15 | *si* | 47 | / | 79 | O | 111 | o |
| 16 | *dle* | 48 | 0 | 80 | P | 112 | p |
| 17 | *dc1* | 49 | 1 | 81 | Q | 113 | q |
| 18 | *dc2* | 50 | 2 | 82 | R | 114 | r |
| 19 | *dc3* | 51 | 3 | 83 | S | 115 | s |
| 20 | *dc4* | 52 | 4 | 84 | T | 116 | t |
| 21 | *nak* | 53 | 5 | 85 | U | 117 | u |
| 22 | *syn* | 54 | 6 | 86 | V | 118 | v |
| 23 | *etb* | 55 | 7 | 87 | W | 119 | w |
| 24 | *can* | 56 | 8 | 88 | X | 120 | x |
| 25 | *em* | 57 | 9 | 89 | Y | 121 | y |
| 26 | *sub* | 58 | : | 90 | Z | 122 | z |
| 27 | *esc* | 59 | ; | 91 | [ Ä | 123 | { ä |
| 28 | *fs* | 60 | < | 92 | \ Ö | 124 | | ö |
| 29 | *gs* | 61 | = | 93 | ] Å | 125 | } å |
| 30 | *rs* | 62 | > | 94 | ^ Ü | 126 | ~ ü |
| 31 | *us* | 63 | ? | 95 | \_ | 127 | *delete* |

1. A Crash Course in C

This appendix gives a crash course in the C programming language. The story of C began in the late sixties with the Multics project, which was a predecessor to the UNIX operating system. The first two versions of Multics were written in the assembly language. However, since programming in assembly languages was rather cumbersome, an appropriate high-level language was called for.

The researchers Dennis Ritchie and Ken Thompon created the language B by simplifying the research language BCPL[[1]](#footnote-1) in 1969. The new versions, called New B, were introduced in 1972. The name did soon evolve into C. In 1978, Brian Kernighan and Dennis Ritchie wrote the book The C Programming Language, which introduced the informal K&R C standard, named after the authors. In 1983, a working group was formed by the American National Standard Institute (ANSI) in order to define a C standard, which was finally adopted in 1989. The ANSI C standard is described in an appendix of the second edition of The C Programming Languages.

## The Compiler and the Linker

The source code is the actual text the programmer writes. The compiler is the program that translates the source code into object code, which the computer can read. The linker puts several compiled files into an executable file

Let us say we have a C program in the source code file Program.c and a routine used by the program in Routine.c. Furthermore, the program calls a function in the standard library. In this case, the compiler translates the source code into object code and the linker joins the code into the executable file Program.exe.

Compiler

Program.c

Linker

Program.o

Routine.c

Routine.o

Compiler

Standard Library

Program.exe

If the compiler reports an error we refer to it as compile-time error, and if an error occurs during the execution of the program we call it a run-time error.

## The Hello-World Program

In the first edition of The C Programming Language the hello-world example was introduced; that is, a small program that prints the text “Hello, World!”[[2]](#footnote-2) on the screen. To write a hello-world programoften means to write a small program to test that the compiler and linker work properly. The execution of the program does always start with the function main.

#include <stdio.h>

void main(void) {

puts("Hello, World!");

}

When executing, the program above generates the following printout.



## Comments

In C, it is possible to insert comments to describe and clarify the meaning of the program. The comments are ignored by the compiler[[3]](#footnote-3). There are two types of comments: line comments and block comments. Line comments start with two slashes and ends at the end of the line.

printf("Hello, World!"); // Prints "Hello, World!".

Block comments begin with a slash and an asterisk and ends with an asterisk and a slash. A block comment may range over several lines.

/\* This is an example of a C program. It writes the text

"Hello, World!" at the screen. \*/

#include <stdio.h>

void main(void) {

printf("Hello, World!");

}

Block comments cannot be nested. The following example will result in a compile-time error.

/\* A block comment cannot be /\* nested inside \*/ another

block comment. \*/

A piece of advice is that you use the line comments for regular comments, and save the block comments for situations when you need to comment a whole block of code for debugging purposes. Actually, there is a third level of comments if you use conditional programming, see macros (Section 13.31) for details.

#if 0

This is an example of a C program. It writes the text

"Hello, World!" at the screen.

#endif

#include <stdio.h>

void main(void) {

printf("Hello, World!");

}

## Types and Variables

There are several types in C. They can be divided into two groups: simple and compunded. The simple types can be further classified into integral, floating, and logical types. The compunded types are arrays, structures, unions, and pointers. They all (directly or indirectly) composed by simple types. We can also define a type with our own integer values, called the enumeration type.

### Simple Types

There are four simple types intended for storing integers: char, short int, int, and long int. They are called the integral types. The types short int and long int may be abbreviated to short and long, respectively. As the names implies, they are designed for storing characters, small integers, normal integers, and large integers, respectively. The exact limits of the values possible to store varies between different compilers. Such property that is not defined in the standard, but is rather defined by the compiler, linker, or operating system, is said to implementation dependent.

Furthermore, the integral types may be signed or unsigned. An unsigned type must not have negative values. If the word signed or unsigned is left out, a short int, int, and long int will be signed. Weather a char will be signed or unsigned is implementation dependent. However, a char is always one byte long, which means that it always holds a single character, regardless whether it is unsigned or not.

The next category of simple types is the floating types, which are used to store decimal numbers. The types are float, double, and long double; where float stores the smallest value and long double the largest one. How large values each type can store is implementation dependent. A floating type cannot be unsigned.

In many languages, there is a logical type (named bool, boolean, or logical) that stores logical value true or false. In C, however, there is no such type[[4]](#footnote-4). Instead, int is used as a logical type, where zero represents false and all other values represent true.

### Variables

A variable can be regarded as a box in the memory. In almost every case, we do not need to know its exact memory address. A variable always has a name, a type, and a value. We define a variable by simple writing its type and name. If we want to, we can initialize the variable; that is, assign it a value. If we do not, the variable’s value is undefined[[5]](#footnote-5). In this chapter, that value is denoted by a question mark (?).

int i = 123, j;

double d = 3.14;

char c = 'a';

123

i

?

j

3.14

d

'a'

c

We can transform values between the types by stating the new type within parentheses. The use of stating the type is not strictly necessary. However, if omitted the compiler may give warnings[[6]](#footnote-6).

int i = 123;

double x = 1.23;

int j = (int) x;

double y = (double) i;

### Constants

As the name implies, a constant is a variable whose value cannot be altered since it has been initialized. Unlike variables, constants must always be initialized and are often written in capital letters.

const double PI = 3.14;

### Output

As noted in the first section of this chapter, the standard function for output is printf (formatted print) writes to the standard output device (normally a text window).

The first argument is the format strings. With it, we can write every kind of value. However, there are many format codes to keep track of. A format code begins with the per cent character (%) followed with optional digits and signs ending with a character. We can write all the types we have went through so far. The character '\n' represents a new line. Below follows a list of the most common format codes.

|  |  |
| --- | --- |
| Code | Type |
| d | decimal integer |
| o | octal integer |
| x | hexadecimal integer |
| f | double |
| lf | long double |
| c | char |
| s | char \* |
| p | Prints the address of the variable in hexadecimal form. |

In the program below, *i* represents an integer, *f[[7]](#footnote-7)* a double, and *c* a character. Note that it is the responsible of the programmer to match the values following the format string with values of the correct types. There is no built-in error handling and mistakes would likely result in a run-time error.

#include <stdio.h>

void main() {

char c = 'a';

int i = 123;

double d = 1.23;

printf("char %c, int %d, double %f\n", c, i, d);

}

There is also the puts function that prints a string followed by a new-line. The following two lines are equivalent.

printf("Hello, World!\n");

puts("Hello, world!");

### Input

The standard functions for input is and scanf (formatted scan), it reads from the standard input device (normally the keyboard).

Similar to printf, its first argument is the format string. The format codes mostly agree with those of printf, but not completely. See the table at the end of the chapter for a list of format codes. For instance, scanf expects lf to read a double value, in contrast to printf that expects f*[[8]](#footnote-8)*.

Another difference is that while printf wants the values corresponding to the format codes, scanf wants the addresses of the variables that are going to store the values. This is accomplished by using the ampersand (&), see the Section 10.1.1 on pointers. Below follows a list of the most common format codes.

|  |  |
| --- | --- |
| Code | Type |
| d | decimal integer |
| o | octal integer |
| x | hexadecimal integer |
| f | float |
| lf | double |
| ld | long double |
| c | Char |
| S | char \* |

In the program below, note that scanf expects lf to represent a double float, in contrast to printf that expects f.

#include <stdio.h>

void main() {

char c;

int i;

double d;

printf("char: ");

scanf("%c", &c);

printf("int: ");

scanf("%d", &i);

printf("double: ");

scanf("%lf", &d);

printf("char %c, int %d, double %f\n", c, i, d);

}

There is also the gets function that reads a string until it reaches a new-line, which differs from scanf that reads a string until it reaches a white-space character[[9]](#footnote-9). It the user prints “Hello World” followed by a new-line twice when executing the code below, gets will place “Hello world” is s, while scanf will place “Hello” in t. Note that we do not give the addresses of s or t, since they are pointer to characters in themselves.

char s[20], t[20];

gets(s);

scanf("%s", t);

### Enumerations

An enumeration is a way to create our own integral type. We can define which values a variable of the type can store. In practice, however, enumerations are essentially an easy way to define integer constants[[10]](#footnote-10).

enum Cars {FORD, VOLVO, TOYOTA, VOLKSWAGEN};

Unless we state otherwise, the constants are assigned to zero, one, two, three, and so on. In the example above, FORD is an integer constant with the value zero; VOLVO has the value one, TOYOTA three, and VOLKSWAGEN four.

We do not have to name the enumeration type. In the example above, Cars can be omitted. We can also assign an integer value to some (or all) of the constants. In the example below, TOYOTA is assigned the value 10. The constants without assigned values will be given the value of the preceding constant before, adding one. This implies that VOLKSWAGEN will be assigned the value 11.

enum {FORD, VOLVO, TOYOTA = 10, VOLKSWAGEN};

### Arrays

An array is a variable compiled by several values of the same type. The values are stored on consecutive locations in memory. The array size may be stated explicitly of implicitly by a list of values. In the following example, a is given the size three, b is given size two and c is given size four, even though only its first two values are defined, which may cause the compiler to emit a warning. Naturally, d is given the size three.

int a[3] = {11, 12, 13};

double b[] = {1.2, 3.4};

char c[4] = {'a', 'b'}, d[3];

d

11

a

b

c

12

13

1.2

3.4

'a'

'b'

?

?

?

?

?

0

1

2

0

1

0

1

2

3

0

1

2

A value of an array can be accessed by index (brackets) notation.

int i = a[2];

double x = b[0];

char t = c[1];

13

i

1.2

x

'b'

t

### Characters and Strings

As stated in Section 1.1.1, a char is a small integer type intended to store exactly one character. A string stores a (possible empty) sequence of characters. There is no built in type for describing a string. Instead, a string is defined by an array of characters. Note that characters are enclosed by single quotations while strings are enclosed by double quotations.

char c = 'a';

char s[] = "Hello, World!";

'a'

c

"Hello, World!"

s

When defining a string, the compiler adds a special zero character '\0' at the end. Its task is to mark the end of the string. If the code above, the character array s has a size of fourteen (thirteen regular characters, including space, comma, and exlamation mark as well as the finishing zero character). The two following definitions are equivalent, they both define character arrays of size four (three regular characters and the finishing zero character).

char s1[] = "abc";

char s2[] = {'a', 'b', 'c', '\0'};

### The ASCII Table

So far we have assigned characters to char variables. Technically, the compiler translates a character literal into an integer in accordance to the ASCII table (see Appendix Fel! Hittar inte referenskälla.), where character has a corresponding integer value. For instance, ‘a’ corresponds to 97, ‘b’ to 98, and so on. The line above could have been written as the following code instead. The possibility to write character and string literal is for convenience only.

char s2[] = {97, 98, 99, 0};

### Pointers

A pointer is a variable containing the address of value. For example, let us say that the integer i has the value 123 that is stored on the memory address 10,000, and that the size of each value is four bytes. As p is a pointer to i, it holds the value 10,000.

10000

p

123

10000

9992

9996

10004

In almost all cases, we do not really need to know the address of the value. Therefore, a simpler (and clearer) way to illustrate the same thing is to draw an arrow from the pointer to the value.

p

123

i

The following code gives rise to the diagram above[[11]](#footnote-11), where the ampersand (&) denotes the address of the variable.

int i = 123;

int \*p = &i;

If we want to access the value pointed at, we use the asterisk (\*), which dereferenceers the pointer (“follows the arrow”). As the address operator (&) gives the address of a variable (“against the arrow”), they address and dereferenceerring operators can be regarded as each other reverses. Note that the asterisk is used on two occasions, when we define a pointer variable and when we dereferenceer a pointer. The asterisk is in fact used on a third occasion, when multiplying two values. The ampersand is used on two occasions: address and bitwise and[[12]](#footnote-12).

int i = 999;

int \*p = &i; // Pointer definition.

int j = \*p; // Dereferenceereeing of a pointer.

int a = 1, b = 2;

int c = a \* b; // Multiplication.

### Pointers and Dynamic Memory

Pointers can also be used to allocate dynamic memory. There is a section of the memory called the heap that is used for dynamically allocated memory blocks. The standard functions malloc, calloc, realloc and free used to allocate, reallocate, and deallocate the memory, respectively. Memory not dynamically allocated is referred to static memory.

p

123

int \*p = malloc(sizeof(int));

\*p = 123;

free(p);

We can also allocate memory for a whole array. Even though p is a pointer in the example below, we can use the array index notation to access a value of the array in the allocated memory block.

p

123

124

125

0

1

2

int \*p = calloc(3, sizeof(int));

p[0] = 123;

p[1] = 124;

p[2] = 125;

free(p);

Once we have allocated a memory, we can allocate a block of different size by calling realloc, which copies to contents of the memory block to new address, is necessary.

int \*p = malloc(sizeof(int));

p[0] = 123;

p = realloc(2 \* sizeof(int));

p[1] = 456;

free(p);

p

123

456

0

1

The difference between malloc and calloc is that calloc resets the bytes of the memory blocks to zero, which malloc do not. Otherwise, the following two calls are equivalent.

int \*p = calloc(3, sizeof(int));

int \*q = malloc(3 \* sizeof(int));

The predefined constant NULL holds the pointer equivalence of the zero value. We say that the pointer is set to null. In a diagram, we can simple write null.

p

NULL

int \*p = NULL;

Sometimes, the electric ground symbol is used to symbolize a null pointer. For this reason, a null pointer is sometimes said to be a grounded pointer.

p

There is a special type void, which actually marks the absence of a type. We can define a pointer to void; we cannot, however, dereferenceer the pointer. It is useful in low-level application where we want to exam a specific location in memory.

void\* voidPtr = (void\*) 10000;

The void type is useful on one further occasion: to mark that a function does not return a value or misses parameters, see the function section later in this chapter.

In the example below, the memory block has been deallocated, but p has not been set to null. It has become a dangling pointer; it is not null and does not really points at anything. In spite of that, we try to access the value p points at. This is a dangerous operation and would most likely result in a run-time error.

p

dangling pointer

int \*p = malloc(sizeof(int));

\*p = 1;

free(p);

\*p = 2; // Undefined behavor

In the example below, we allocate memory for two pointers, p and q. Then we assign p to q, by doing so we have created a memory leak. There is no way we can access or deallocate the memory block that was pointed at by p. In fact, we deallocate the same memory block twice as both pointers by now points at the same memory block. This dangerous operation will most likely also result in a run-time error.

p

(a)

q

1

2

p

(b)

q

1

2

int \*p = malloc(sizoef(int)); // (a)

int \*q = malloc(sizeof(int));

\*p = 1;

\*q = 2;

p = q; // (b)

free(p); // Undefined behavor. Deallocates the same memory block twice,

free(q); // as p and q points at the same memory block. Also memory leak,

// since the memory block holding 1 cannot be accessed any more.

### Defining our own types

It is possible to define our own type with typedef, which is a great tool for increasing the readability of the code. However, too many defined types tend to make the code less readable. Therefore, a piece of advice would be to use typedef with care.

int i = 1;

typedef unsigned int unsigned\_int;

unsigned\_int u = 2;

typedef int\* int\_ptr;

int\_ptr ip = &i;

typedef unsigned\_int\* uint\_ptr;

uint\_ptr up = &u;

### The Type Size

The operator sizeof gives us the size of a type[[13]](#footnote-13), either by taking the type surrounded by parentheses or a value of the type. The size of a character is always one byte and the signed and unsigned forms of each integral type always have the same size. Otherwise, the sizes are decided by the compiler. In these cases, when a property is not specified by the standard by rather the compiler, we say the property is implementation dependant. The operator returns a value of the type size\_t, which exact definition is also implementation dependant. However, it is often an unsigned integer.

#include <stdio.h>

void main() {

int intSize1 = sizeof (int);

int intSize2 = sizeof intSize1;

int\* intPtr = &intSize1;

int ptrSize = sizeof intPtr;

int array[3] = {1, 2, 3};

int arraySize = sizeof array;

printf("Integer sizes: %d and %d.\n", intSize1, intSize2);

printf("Pointer size: %d\n", ptrSize);

printf("Array size: %d\n\n", arraySize);

}

## Expressions and Operators

The operations of C are divided into the arithmetic, relational, logical, and bitwise operators as well as simple and compound assignment. Moreover, there is the conditional operator.

In the figure below, + is an operator, 1 and 2 are operands and the whole term is an expression.



Operator

Operand

Operand

### Arithmetic Operators

The arithmetic operators are addition (+), subtraction (-), multiplication (\*), division (/), and modulo (%). The first four operators are equivalent to the four fundamental rules of arithmetic. The operators can take operands of integral and floating types. The last operator, modulo, gives the remainder of integer division. If we mix integral and floating types in the expression, the result will have floating type. The modulo operator, however, only works with integral operands. The last assignment in the following code may give rise to a compiler warning as the result of the division is a double and is converted into an integer.

int a = 10, b = 3, c;

c = a + b; // 13

c = a - b; // 7

c = a \* b; // 30

c = a / b; // 3, integer division

c = a % 3; // 1, remainder

double d = 3.0, e;

e = a / d; // 3.333, floating type

### Pointer Arithmetic

The addition and subtraction operators are also applicable on pointers. It is called pointer arithmetic. An integral value can be added to or subtracted from a pointer. The value of the pointer is then changed by the integral value times the size of the type the pointer points at. As the void type is not really a type, but rather the absence of a type, it has no size. Therefore, we cannot perform pointer arithmetic on void pointers.

In the code below, let us assume that number is stored at memory location 10,000 and that the integer type has the size of four bytes. Then the pointer numberPtr will assume the values 10,000, 10,004, 10,008, and 10,012, not the values 10,000, 10,001, 10,002, and 10,003, as pointer arithmetic always takes the size of the type into consideration.

numberPtr

102

100

101

103

10008

10000

10004

10012

int number = 100;

int\* numberPtr = &number;

numberPtr = numberPtr + 1;

\*numberPtr = number + 1;

numberPtr = numberPtr + 1;

\*numberPtr = number + 2;

numberPtr = numberPtr + 1;

\*numberPtr = number + 3;

There is also possible to subtract two pointers pointing at the same type. The result will be the difference in bytes between their two memory locations divided with the size of the type. Again, pointer arithmetic always considers the type size.

int array[] = {1, 2, 3};

int\* p1 = &array[0];

int\* p2 = &array[2];

int diff = p2 - p1; // 2

The index notation for arrays is equivalent to the dereferenceereeing of pointers together with pointer arithmetic. The second and third lines of the following code are by definition interchangeable.

int array[] = {1, 2, 3};

array[1] = array[2] + 1;

\*(array + 1) = \*(array + 2) + 1;

### Increment and decrement

There are two rather special operators: increment (++) and decrement (--)[[14]](#footnote-14). They add or subtract one from its operand. The operator can be placed before (prefix) or after (postfix) its operand.

int a = 1, b = 1;

++a; // 2, prefix increment

b++; // 2, postfix increment

However, there is a difference between prefix and postfix increment or decrement. In the prefix case below, the subtraction occurs first and the new value is returned; in the postfix case, after the subtraction the original value is returned.

int a = 1, b = 1, c, d;

c = --a; // c = 0, prefix decrement

d = b--; // d = 1, postfix decrement

### Relational Operators

There are six relational operators: equal to (==), not equal to (!=), less than (<), less than or equal to (<=), greater than (>), greater than or equal to (>=)[[15]](#footnote-15). Observe that the equal to operator is constituted by two equal signs rather than one. The operators give a logical value, true (an integer value not equal zero) or false (zero). The operands shall be of integral of floating type.

int i = 3;

double x = 1.2;

int b = i > 0; // true (an integer value not equal to zero, usually one)

int c = (x == 2); // false (always zero)

### Logical Operators

There are three logical operators: not (!), or (||), and and (&&). They take and return numerical values interpreted as logical values.

int i = 3;

int b, c, d, e;

b = (i == 3); // true (not zero)

c = !b; // false (zero)

d = b || c; // true (not zero)

e = b && c; // false (zero)

C applies lazy evaluation (also called short-circuit evaluation), which means that an expression will not be evaluated to a further extent than necessary to find its value. In the following example, the evaluation of the expression is completed when the left expression (i != 0) is evaluated to false. If the left expression is false, the whole expression must also be false because it needs both the left and right sub expressions to be true for the whole expression to be true. This gives that the right expression (1 / i == 1) will never be evaluated and the division with zero will never occur.

int i = 0;

int b = (i != 0) && (1 / i == 1); // false (zero);

### Bitwise Operators

An integer value can be viewed as a bit pattern. Our familiar decimal system has base 10; it can be marked with an index 10.



An integer value can also be viewed with the binary system, which has base 2. A single digit is called a bit, and the integer value is called a bit pattern. A bit may have the values one and zero.



There are four bitwise operations in C: inverse (~), and (&), or (|), and exclusive or (^). Exclusive or means that the result is one if one of its operand bits, but not both, is one. They all operate on integral values on bit level; that is, they examine each individual bit of an integer value.

101010102 101010102 101010102

& 100101102 | 100101102 ^ 100101102 ~ 100101102

----------- ------------ ------------ ------------

= 100000102 = 101111102 = 001111002 = 011010012

int a = 170; // 101010102

int b = 150; // 100101102

int c = a & b; // 100000102 = 13010

int d = a | b; // 101111102 = 19010

int e = a ^ b; // 001111002 = 6010

int f = ~b; // 011010012 = 10510

An integer value can also be shifted to the left (<<) or to the right (>>). Each left shift is equivalent with doubling the value, and each right shift is equivalent with (integer) dividing the value with 2. Overflowing bits are dropped for unsigned values; the behaviour of signed values is implementation dependent. Therefore, I recommend you only perform bitwise operations on unsigned values.

unsigned char a = 172; // 101011002

unsigned char b = a << 2; // 101100002 = 17610

unsigned char c = 166; // 101001102

unsigned char d = c >> 2; // 001010012 = 4110

### Assignment

There are two kinds of assignment operators: simple and compound. The simple variant is quite trivial, one or several variables are assigned the value of an expression. In the example below, a, b, and c are all assigned the value 123.

int a, b, c, d = 123;

a = d;

b = c = d;

The compound variant is more complicated. Let us start with the additional assignment operator. In the example below, the value of a is increased by the value of b; that is, a is given the value 4.

int a = 2, b = 4, c = 2;

a += c; // 4, equivalent to a = a + c.

b -= c; // 2, equivalent to a = a - c.

In a similar manner, there are operations -=, \*=, /=, %=, |=, &=, ^=, <<=, and >>=. Note the difference between a -= 2 and a =- 2. In the former case, a is decreased by 2. In the latter case, a is assigned -2.

### The Condition Operator

The condition operator resembles the if-else statement of the next section. It is the only C operator taking three operands. The first expression is evaluated. If it is true, the second expression is evaluated and its value is returned. If the first expression instead is false, the third expression is evaluated and its value is returned.

int a = 1, b = 2, max;

max = (a > b) ? a : b; // The maximal value of a and b.

Too frequent use of this operator tends to make the code compact and hard to read. A piece of advice is that you restrict your use of the operator to the trivial cases.

### Precedence and Associatively

Given the expression 1 + 2 \* 5, what is its value? It is 11 because we first multiply two with five and then add one. We say that multiplication has a higher precedence than addition.

What if we limit ourselves to one operator? Let us pick subtraction. What value has the expression 8 – 4 – 2? As we first subtract 4 from 8 and then subtract 2, the result is 2. As we evaluate the value from left to right, we say that subtraction is left associative.

Below follows a table over the prorities and associative of the C operators. The first operator in the table has the highest precedence.

|  |  |  |
| --- | --- | --- |
| Group | Operators | Associatively |
| Brackets and fields | () [] -> . | Left to Right |
| Unary operator | ! ~ ++ -- + - (type) sizeof | Right to Left |
| Arithmetic operators | \* / %  + - | Left to Right  Left to Right |
| Shift operators | << >> | Left to Right |
| Relation operators | < <= > >=  == != | Left to Right |
| Bitwise operators | &  ^  | | Left to Right |
| Logical operators | &&  || | Left to Right |
| Conditional operator | ?: | Right to Left |
| Assignment operators | = += -= \*/ /= %= &= ^= |= <<= >>= | Right to Left |
| Comma operator | , | Left to Right |

Note that unary +, -, &, and \* have higher precedence than their binary forms. Note that we always can change the evaluation order of an expression by inserting parantheses at appropriate posititions. The expression (1 + 2) \* 5 has the value 15.

## Statements

There are four kinds of statements in C: the selection, iteration, jump, and expression statements.

|  |  |
| --- | --- |
| Kind | Statements |
| Selection | if, if-else, switch |
| Iteration | for, while, do-while |
| Jump | break, continue, goto, return |
| Expression | expression ; |

### Selection statements

The if statement needs in its simplest form a logical expression to decide whether to execute the following statement. The example below means that the text will be output if i not equal zero.

if (i != 0) {

puts("i does not equal zero");

}

As all values not equal to zero are interpreted as true, the following statement is equivalent with the previous one.fexacltly

if (i) {

puts("i does not equal zero");

}

We can also attach an else part, which is executed if the expression is false.

if (i > 0) {

puts("i is greater than zero");

}

else { puts("i is not greater than zero");

}

Between the if and else part we can insert one or more else if parts.

if (i > 0) {

puts("i is greater than zero");

}

else if (i < 0) {

puts("i is less than zero");

}

else {

puts("i is equal to zero");

}

In the examples above, there is not strictly necessary to surround the output statements with brackets. However, in this book brackets are always used for sake of clarity. Moreover, it is always necessary in case of several statements. The brackets and the code in between are called a block.

if (i > 0) { int j = i + 1;

printf("j is %d\n", j);

}

A warning may be in order. In an if statement, it is perfectly legal to use one equal sign instead of two signs when comparing two values. As one equals sign is used for assignment, not comparison, the variable i in the following code will be assign the value one, and the expression will always be true.

if (i = 1) { // Always true.

// ...

}

One way to avoid the mistake above is to swap the variable and the value. As a value can be compared but not assigned, the compiler will issue an error message and the error will be recognized.

if (1 = i) { // Compile-time error.

// ...

}

The switch statement is simpler than the if statement, but not as powerful. It evaluates the switch expression and jumps to a case statement with the matching (constant) value. If no value matches, it jumps to the default statement, if present. The break statement is used to jump out of a switch or iteration statements. The switch expression must have an integral or pointer type and two case statements cannot hold the same value. We can only have one default statement, and it can be omitted. However, it needs not to be placed at the end of the switch statement, even though I recommend you to do so.

switch (i) { case 1:

puts("i is equal to 1");

break;

case 2:

puts("i is equal to 2");

break;

case 3:

puts("i is equal to 3");

break;

default:

puts("i is not equal to 1, 2, or 3.");

break;

}

It is important to remember the break statement. Otherwise, the execution would simple continue with the code attached to the next case statement[[16]](#footnote-16). However, in C we can use the fact that an omitted break statement makes the execution continue with the next statement to group several case statements together.

switch (i) { case 1:

case 2:

case 3:

printf("i is equal to 1, 2, or 3");

break;

// ...

}

### Iteration statements

Iteration statements, also called loops, iterate one or several statements as long as a certain condition is true; while is the simplest loop. It repeats one statement or a block of statements as long as the given expression is true. The example below uses the while statement to write the numbers one to ten.

int i = 1;

while (i <= 10) { printf("%d\n", i);

++i;

}

The same thing can be done with a do-while statement.

int i = 1;

do { printf("%d\n", i);

++i;

}

while (i <= 10);

However, the do-while statement is less powerful. If the expression is false to begin with, the while statement just skips the repetitions altogether, but the do-while statement must always execute the iteration statement at least once in order to reach the continuation expression at the end.

We can also use the for statement, which is a more compact variant of while. It takes three expressions, separated by semicolons. The first expression initializes the variable, the repetition continues as long as the second expression is true and the third expression is executed at the end of each repetition.

int i;

for (i = 1; i <= 10; ++i) { printf("%d\n", i);

}

Similar to the switch statement, an iteration statement can be interrupted by the break statement. Below is an infiniteloop (remember that every non-zero value is regarded as true) that is stopped by the break statement.

int i = 1;

while (1) { printf("%d\n", i);

++i;

if (i > 10) { break;

}

}

The same effect can be archived by omitting the second expression of the for statement.

int i;

for (i = 1; ; ++i) { printf("%d\n", i);

if (i == 10) { break;

}

}

An iteration statement can also include a continue statement. It skips the rest of the current repetition. The following example prints the numbers from 1 to 10 with the exception of 5.

int i;

for (i = 1; i <= 10; ++i) { if (i == 5) { continue;

}

printf("%d\n", i);

}

The following example, however, will not work as intended. Because the continue statement will skip the rest of the while block, i will never be updated and we will be stuck in an infiniteloop. Therefore, I recommend that you use continue with care, or avoid it altogether.

int i = 1;

while (i <= 10) { if (i == 5) { continue;

}

printf("%d\n", i);

++i;

}

### Jump statements

We can jump from one location to another inside the same function by marking the latter location with a label inside the block with the goto statement. The following code is yet another example of how to print the numbers from one to ten, inclusive. As you can, it is not as clear as the previous examples.

int i = 1;

label: printf("%d\n", i);

++i;

if (i <= 10) { goto label;

}

The goto statement is considered to give rise to unstructured code, so called “spagetti code.” I strongly recommended that you omit goto altogether. To put it bluntly: now when you seen it, forget it.

### Expression statements

An expression can form a statement.

a = b + 1; // Assignment operator.

puts("Hello, World!"); // Function call.

In the above examples, we are only interested in the side effects; that a is assigned a new value and that a text is written. We are allowed to write expression statements without side effects; even thought it has no meaning and is likely to be erased by the compiler.

a + b \* c;

### Volatile

A variable can be volatile, which has no functional effect but is rather a notification to the compiler that the variable shall not be subjected to optimization. Let us say we that the memory location 10000 is connected to an input port in a low-level application, and that we want to wait until we receive a signal through the port. If p is not marked as volatile, there is a risk that the compiler would replace the for-loop with another statement since the value of \*p does not seem to change inside the loop.

volatile int\* p = (void\*) 10000;

while (\*p != 0) {

// Wait.

}

## Functions

A function can be compared to a black box. We send in information (input) and we receive information (output). In C, the input values are called parameters and the output value is called the return value.

Input (Parameters)

Function

Output (Return Value)

To start with, let us try the function Square; it takes an integer and returns its square.

int Square(int n) {

return n \* n;

}

void main() {

int i = Square(2); // Square returns 4.

}

2

Square

4

In the example above, the parameter n in the Square definition is called a formal parameter, and the value 3 in the Square call in main is called an actual parameter or an argument.

Now, let us try a more complicated function: SquareRoot takes a value of double type and returns its square root. The idea is that the function iterates and calculates increasingly better root values by taking the mean value of the original value divided with the current root value and the previous root value. The process continues until the difference between two consecutive root values has reached an acceptable tolerance. Just like main[[17]](#footnote-17), a function can have local variables; root and oldRoot hold the current and previous value of the root, respectively.

#include <stdio.h>

double SquareRoot(double value) {

const double EPSILON = 1e-12;

double root = value, oldRoot = value;

while (1) { root = ((value / root) + root) / 2;

if ((oldRoot - root) <= EPSILON) { return root;

}

oldRoot = root;

}

}

void main() {

double input = 16;

printf("SquareRoot of %f: %f\n", input,

SquareRoot(input));

}

### Void Functions

A function does not have to return a value; in that case, we set void as the return type. As mentioned above, void is used to state the absence of a type rather than a type in itself. We can return from a void function by just stating return without a value.

void PrintSign(int value) {

if (value < 0) { puts("Negative.");

return;

}

if (value > 0) { puts("Positive.");

return;

}

puts("Zero");

}

There is not a problem if the execution of a void function reaches the end of the code, it just jump back to the calling function. However, a function not returning void shall always return a value before reaching the end of the function. Otherwise, the compiler often gives a warning.

### Local and global variables

There are two kinds of variables: local and global. A global variable is defined outside a function and a local variable is defined inside a function.

#include <stdio.h>

int global = 1;

void main() {

int local = 2;

printf("Global variable: %d, Local variable: %d.\n",

global, local);

}

A global and a local variable can have the same name. In that case, the name inside the function refers to the local variable. We cannot access the global variable in the inner scope[[18]](#footnote-18) when a local variable with same name is present.

#include <stdio.h>

int number = 1;

void main() {

int number = 2;

printf("Variable: %d.\n", number); // 2

}

One way to distinguish them is to precede the global variable name with ’g\_’.

#include <stdio.h>

int g\_number = 1;

void main() {

int number = 2;

printf("Global variable: %d, Local variable: %d.\n",

g\_number, number); // 1, 2

}

Global variables can only be initialized with constant values.

int g\_number1 = 1; // Right.

int g\_number2 = g\_number1 + 1; // Wrong.

### Inner Blocks with Local Variables

In C, you are allowed to use the bracket notation when initializerializing structures, unions, and arrays, but not when assigning[[19]](#footnote-19). Therefore, the following code will not work.

INT\_PAIR Block(int\* p) {

INT\_PAIR result;

if (p != NULL) {

result = {\*p - 1, \*p + 1}; // Wrong.

return result;

}

else {

result = {0, 0}; // Wrong.

return result;

}

}

One solution would be to a more clumsy code.

INT\_PAIR Block(int\* p) {

INT\_PAIR result;

if (p != NULL) {

result.a = \*p - 1;

result.b = \*p + 1;

return result;

}

else {

result.a = 0;

result.b = 0;

return result;

}

}

However, a more elegant solution would be to use inner blocks with local variables. Each block can define its own variable, which is especially valuable when initializerializing structures.

struct IntPair {

int a, b;

};

struct IntPair Block(int\* p) {

if (p != NULL) {

struct IntPair result = {\*p - 1, \*p + 1};

return result;

}

else {

struct IntPair empty = {0, 0};

return empty;

}

}

An alternative way is to use typedef.

typedef struct {

int a, b;

} INT\_PAIR;

INT\_PAIR Block(int\* p) {

if (p != NULL) {

INT\_PAIR result = {\*p - 1, \*p + 1};

return result;

}

else {

INT\_PAIR empty = {0, 0};

return empty;

}

}

### Call-by-Value and Call-by-Reference

Say we want to write a function for switching the values of two variables.

#include <stdio.h>

void Swap(int number1, int number2) {

int temp = number1; // (a)

number1 = number2; // (b)

number2 = temp; // (c)

}

void main() {

int num1 = 1, num2 = 2;

printf("Before: %d, %d\n", num1, num2); // 1, 2

Swap(num1, num2);

printf("After: %d, %d\n", num1, num2); // 1, 2

}

Unfortunately, it will not work; the variables keep their values. The explanation is that the values of firstNum and secondNum in main are copied into num1 and num2 in Swap. Then num1 and num2 exchange values with the help of temp. However, their values are not copied back into firstNum and secondNum in main.

1

2

main:

1

firstNum

2

Swap:

num2

1

temp

secondNum

num1

2

2

num2

1

temp

num1

2

1

num2

1

temp

num1

(a)

(b)

(c)

The problem can be solved with reference calls. Instead of sending the values of the actual parameters, we send theirs addresses by define the formal parameters as pointers to the type.

#include <stdio.h>

void Swap(int\* numberPtr1, int\* numberPtr2) {

int temp = \*numberPtr1; // (a)

\*numberPtr1 = \*numberPtr2; // (b)

\*numberPtr2 = temp; // (c)

}

void main() {

int num1 = 1, num2 = 2;

printf("Before: %d, %d\n", num1, num2); // 1, 2

Swap(&num1, &num2);

printf("After: %d, %d\n", num1, num2); // 2, 1

}

In this case, we do not send the values of firstNum and secondNum, but rather their addresses. Therefore, num1 and num2 in Swap does in fact contain the addresses of firstNum and secondNum on main. As in the reference section above, we illustrate this with dashed arrows. Therefore, when num1 and num2 exchanges values, in fact the values of firstNum and secondNum are exchanged.

1

2

firstNum

num2

1

temp

secondNum

num1

(a)

2

2

firstNum

num2

1

temp

num1

(b)

2

1

firstNum

num2

1

temp

num1

(c)

### Static, Extern, and Register Variables

In the function below, s\_count (the prefix ‘\_s’ is used for static variables) is a static local variable, which means that it is initialized when the execution of the program starts rather when the function is called. If s\_count was a regular local variable (without the keyword static), the function would, at every call, write that the function has been called once, as s\_count would be initialized to zero at every call.

void KeepCount() {

static int s\_count = 0;

s\_count++;

printf("This function has been called %d times.", s\_count);

}

If we define a global variable in one file, it will be accessable in another file if we declare[[20]](#footnote-20) it as extern. If we omit the extern keyword in the second file, the linker would complain about us defining two global variables with the same name.

File1.c

int g;

File2.c

extern int g;

Finally, a variable can also be marked with the keyword register, which is a notification to the compiler that the variable is suitable to place in a process register rather them the memory. The only function difference is that we cannot get the address of a register variable, since it might be located in a register.

register int i;

int\* p = &i; // Wrong

### Recursion

A function may call itself; it is called recursion. In the following example, the mathematical function factorial (n!) is implemented. It can be defined in two ways. The first definition is rather straightforward. The result of the function applied to a positive integer n is the product of all positive integers up to and including n.



int Factorial(int n) {

int result = 1, i;

for (i = 1; i <= n; ++ i) { result \*= i;

}

return result;

}

An equivalent definition involves a recursive call that is easier to implement.



int Factorial(int n) {

if (n == 1) {

return 1;

}

else { return n \* Factorial(n - 1);

}

}

### Definition and Declaration

It important to distinguish between definition and declaration. For a function, its definition generates code while the declaration is merely an item of information to the compiler. A function declaration is also called a prototype.

When it comes to mutual recursion (two functions calling each other), at least the second of them must have a prototype to avoid compiler warnings. I recommend that you put prototypes for all functions at the beginning of the file (or in a header file, see Section 13.29). In the following example, we use two functions to decide whether a given non-negative integer is even or odd according to the formulas below.





Note that it is not necessary to name the parameters in a function prototype. If names after all are given, they will be ignored by the compiler, which means that the parameters do not need to have the same names in the function declaration and definition. The only restriction is that two parameters cannot have the same name.

int IsEven(int);

int IsOdd(int n);

As C does not include a logical type, we use int to represent true (1) or false (0).

int IsEven(int num) {

if (num == 0) {

return 1;

}

else { return IsOdd(num - 1);

}

}

int IsOdd(int num) {

if (num == 0) {

return 0;

}

else {

return IsEven(num - 1);

}

}

One peculiar thing about prototypes in C is that they can have an unspecified parameter list. If the parameter list is completely omitted in a function prototype, every parameter list is allowed in the call. The following code is will not result in any compile-time errors. However, there is likely to be run-time errors.

void Print();

void main(void) {

Print();

Print(1);

Print(1, 2);

Print(1, 2, 3);

}

The obvious way to avoid the run-time errors is to always state the parameter list in function prototypes. If the function does not have any parameters, it can be stated by the void type. In that case, the function can only be called without parameters.

void Print(void);

### Higher-Order Functions

A function that takes another function as a parameter is called a higher-order function. Technically, it does not take the function itself as a parameter, but rather a pointer to the function. However, the pointer marker (\*) may be omitted. The parameter function is also called a callback function. The following example takes an array of the given size and applies the given function to each integer in the array. ApplyArray a higher order function and Apply is a callback function.

#include <stdio.h>

void ApplyArray(int intArray[], int size, int Apply(int)) {

int index;

for (index = 0; index < size; ++index) { intArray[index] = Apply(intArray[index]);

}

}

int Double(int number) {

return 2 \* number;

}

int Square(int number) {

return number \* number;

}

void PrintArray(int intArray[], int size) {

int index;

for (index = 0; index < size; ++index) { printf("%d ", intArray[index]);

}

printf("\n");

}

void main() {

int numberArray[] = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10};

int arraySize = sizeof numberArray / sizeof numberArray[0];

PrintArray(numberArray, arraySize);

// Doubles every value in the array.

ApplyArray(numberArray, arraySize, Double);

PrintArray(numberArray, arraySize);

// Squares every value in the array.

ApplyArray(numberArray, arraySize, Square);

PrintArray(numberArray, arraySize);

}

An alternative way is to define the type of the callback function. In the following code segment, the APPLY\_FUNC type is a pointer to a function that takes an integer parameter and returns an integer value.

typedef int (\*APPLY\_FUNC)(int);

void ApplyArray(int intArray[], int size,

APPLY\_FUNC pApplyFunc) {

int index;

for (index = 0; index < size; ++index) { intArray[index] = pApplyFunc(intArray[index]);

}

}

One additional point of the example above is how to of find the size of an array: we divide the size of the array with the size of its first value. This function does only work on static arrays, not on dynamically allocated arrays or array given as parameters to functions. A parameter array is in fact converted to a pointer to the array type. The following two function definitions are (by definition) equivalent. This gives that the function must have an extra integer parameter stating the array size.

void PrintArray(int intArray[], int arraySize) {

// ...

}

void PrintArray(int\* intArray, int arraySize) {

// ...

}

## Structures and Linked Lists

A struct is a compound type. Like arrays, it holds several values, called fields. However, in contrast to arrays, the values can hold different types.

A pointer may point at a struct value as well as a simple value, and a struct may have a pointer to another struct value as a field, which in turn points at another struct value, and so one. In this way, a linked list can be constructed. The list must end eventually, so the last pointer points at null. A pointer to the next cell in the list is called a link.

struct Cell {

int value;

struct Cell\* nextPtr;

};

struct Cell cell3 = {3, NULL};

struct Cell cell2 = {2, &cell3};

struct Cell cell1 = {1, &cell2};

1

value

nextPtr

2

value

nextPtr

3

value

nextPtr

NULL

### Stacks and Linked Lists

A stack is very valuable in a number of applications and it can be implemented with a linked list. We can add a value on top of the stack, we can inspect or remove the topmost value, and we can check whether the stack is empty. However, we cannot do anything to the values that are not on top. The function adding a new value on top of the stack is called push and the function removing it is called pop. Let us say that we push our stack three times with the values 1, 2, and 3. Then we can only access the topmost value, 3, and not the two below, 1 or 2.

1

2

3

push 1

push 2

push 3

Cell.h

struct Cell { int m\_value;

struct Cell\* m\_nextPtr;

};

typedef struct Cell CELL;

void CellInitializer(CELL\* cellPtr, int value, CELL\* nextPtr);

void CellSetValue(CELL\* cellPtr, int value);

int CellGetValue(CELL\* cellPtr);

void CellSetNext(CELL\* cellPtr, CELL\* nextPtr);

CELL\* CellGetNext(CELL\* cellPtr);

Cell.c

#include "cell.h"

void CellInitializer(CELL\* cellPtr, int value, CELL\* nextPtr) {

cellPtr->m\_value = value;

cellPtr->m\_nextPtr = nextPtr;

}

void CellSetValue(CELL\* cellPtr, int value) {

cellPtr->m\_value = value;

}

int CellGetValue(CELL\* cellPtr) {

return cellPtr->m\_value;

}

void CellSetNext(CELL\* cellPtr, CELL\* nextPtr) {

cellPtr->m\_nextPtr = nextPtr;

}

CELL\* CellGetNext(CELL\* cellPtr) {

return cellPtr->m\_nextPtr;

}

Main.c

#include <stdlib.h>

#include "cell.h"

void main() {

CELL\* cellPtr1 = malloc(sizeof(CELL));

CELL\* cellPtr2 = malloc(sizeof(CELL));

CELL\* cellPtr3 = malloc(sizeof(CELL));

CellInitializer(cellPtr3, 3, NULL);

CellInitializer(cellPtr2, 2, cellPtr3);

CellInitializer(cellPtr1, 1, cellPtr2);

}

The following struct is created by main above.

1

m\_value

m\_nextPtr

2

m\_value

m\_nextPtr

3

m\_value

m\_nextPtr

NULL

cellPtr1

cellPtr2

cellPtr3

However, there is one more thing to think about. What happens if we run out of dynamic memory or try to access the topmost value of an empty stack? We can deal with the problem in some different ways, everything from ignoring it to abort the execution. In this book, I have limited the error handling of memory shortage to the use of assert, which is a macro (see Section 13.31) that takes a logical value and aborts the execution if the value is false, adding an error message with information about the file name and the code line number. To keep things simple, let us use that function in the stack structure too.

Stack.h

struct Stack { CELL\* m\_firstCellPtr;

};

typedef struct Stack STACK;

void StackInitializer(STACK\* stackPtr);

void StackClear(STACK\* stackPtr);

void StackPush(STACK\* stackPtr, int value);

void StackPop(STACK\* stackPtr);

int StackTop(STACK\* stackPtr);

int StackIsEmpty(STACK\* stackPtr);

Stack.c

#include <stdlib.h>

#include <assert.h>

#include "cell.h"

#include "stack.h"

void StackInitializer(STACK\* stackPtr) {

stackPtr->m\_firstCellPtr = NULL;

}

void StackClear(STACK\* stackPtr) {

CELL\* cellPtr = stackPtr->m\_firstCellPtr;

while (cellPtr != NULL) { CELL\* tempCellPtr = cellPtr;

cellPtr = CellGetNext(cellPtr);

free(tempCellPtr);

}

}

void StackPush(STACK\* stackPtr, int value) {

CELL\* newCellPtr= malloc(sizeof(CELL));

assert(newCellPtr!= NULL);

CellInitializer(newCellPtr, value, stackPtr->m\_firstCellPtr);

stackPtr->m\_firstCellPtr = newCellPtr;

}

void StackPop(STACK\* stackPtr) {

CELL\* tempCellPtr;

assert(stackPtr->m\_firstCellPtr != NULL);

tempCellPtr = stackPtr->m\_firstCellPtr;

stackPtr->m\_firstCellPtr = CellGetNext(stackPtr->m\_firstCellPtr);

free(tempCellPtr);

}

int StackTop(STACK\* stackPtr) {

assert(stackPtr->m\_firstCellPtr != NULL);

return CellGetValue(stackPtr->m\_firstCellPtr);

}

int StackIsEmpty(STACK\* stackPtr) {

return (stackPtr->m\_firstCellPtr == NULL);

}

Main.c

#include <stdlib.h>

#include "cell.h"

#include "stack.h"

void main() {

STACK stack;

StackInitializer(&stack);

StackPush(&stack, 1);

StackPush(&stack, 2);

StackPush(&stack, 3);

StackClear(&stack);

}

The stack in main above gives rise to the following struct. However, when StackClear is called, it deallocates the allocated memory. All memory allocated with malloc or calloc must (or at least should) be deallocated with free.

m\_nextPtr

3

2

m\_value

m\_nextPtr

1

m\_value

m\_nextPtr

NULL

m\_value

m\_firstCellPtr

### Unions

A union is a compound type like struct. The difference is that while a struct places each of its members on different memory addresses in order for them to not intervene with each other, a union locates each of its members on the same address. This implies that when one member is assigned a value, it overwrites the other values. Unions are mostly used in low-level applications. In the code below, the two structures overlap each other.

struct \_Bits8Registers { // A character always has the size of 1 byte (8 bits).

unsigned char ah, al, bh, bl, ch, cl, dh, dl;

};

struct \_Bits16Registers { // Assuming short has a size of 2 bytes (16 bits).

unsigned short ax, bx, cx, dx;

};

union \_Registers {

struct \_Bits8Registers bits8Registers;

struct \_Bits16Registers bits16Registers;

};

### Bitfields

A bitfield is a struct with the possibility to define the size of the members in bits, which is useful in low-level applications. Only integral types can have bits, and the total number of bits must not exceed the number of bits an integer (eight times sizeof (int)). In the code below, the whole struct shares one byte (eight bits). The second field does not have a value; it only marks that the second bit shall be ignored.

struct Communicate {

int alert : 1;

int :1;

int read : 3;

int write : 3;

};

## Structures with function pointers

A struct can hold pointers to functions as well as values. However, in order to obtain the specific struct value, we have to add its address a pointer parameter[[21]](#footnote-21) in the function call.

#include <stdio.h>

struct Person { char name[100];

int age;

void (\*Print)(struct Person\*);

};

void PrintNameAndAge(struct Person\* personPtr) {

printf("Name: %s\n", personPtr->name);

printf("Age: %d\n\n", personPtr->age);

}

void PrintOnlyName(struct Person\* personPtr) {

printf("Name: %s\n\n", personPtr->name);

}

void PrintOnlyAge(struct Person\* personPtr) {

printf("Age: %d\n\n", personPtr->age);

}

void main() {

struct Person adam = {"Adam", 10, &PrintNameAndAge};

struct Person bertil = {"Bertil", 20, &PrintOnlyName};

struct Person ceasar = {"Ceasar", 30, &PrintOnlyAge};

adam.Print(&adam); // Prints the name and age.

bertil.Print(&bertil); // Prints only the name.

ceasar.Print(&ceasar); // Prints only the age.

}

## The Preprocessor

The preprocessor is a tool that precedes the compiler in interpreting the code. The #include directive is one of its parts. It opens the file and includes its text. So far, we have only included system header files, whose names are surrounded by arrow brackets (for instance, <stdio.h>). Later on, we will include our own header files. Then we will use parentheses instead of arrow brackets. The difference is that the pre-processor looks for the system header files in a special system directory while it looks for our header files in the local directory (usually the source code directory).

Another part of the pre-processor is the macros. They acts like functions with the difference that they do not perform any type checking, they just replace text. We have already used the assert macro. A macro is introduced with the #define directive and is often written with capitals.

#define ADD(a, b) ((a) + (b))

printf("%d\n", ADD(1 + 2, 3 \* 4)); // 15

Unlike regular C code, if we add a page-break we have to mark it by a backslash.

#define ADD(a, b) ((a) + \

(b))

It is also possible to perform conditional programming by checking the value of macros. In the following example, we define a system integer according to the underlying operational system.

#ifdef WINDOWS

#define SYSINT int

#endif

#ifdef LINUX

#define SYSINT unsigned int

#endif

#ifdef MAC

#define SYSINT long int

#endif

SYSINT iOpData = 0;

Condition programming can also be used as a third level of comments (above line and block comments, see Section 13.31). In the following code the whole segment between the #if and #endif directives will be omitted, regardless of the comments, as zero always is interpreted as false.

#if 0

int Square(int value)

{ return value \* value; // Square.

}

#endif

## The Standard Library

The C standard library hold around 200 functions and some macros divided into several sections.

### File Management

We can open, write to, read from, and close files with the help of file pointers. A file pointer is a pointer to a value of the FILE type; it can be considered a connection to the file. The program below reads a series of integers from the text file Input.txt and writes their squares to the file Output.txt. The function feof returns zero when there is no more value to be read from the file. Finally, we must not forget to close the file.

#include <stdlib.h>

#include <assert.h>

#include <stdio.h>

void main(void) {

FILE\* inFile = fopen("Input.txt", "r");

FILE\* outFile = fopen("Output.txt", "w");

assert(inFile != NULL);

assert(outFile != NULL);

while (!feof(inFile)) { double value;

fscanf(inFile, "%lf", &value);

fprintf(outFile, "%f\n", value \* value);

}

fclose(inFile);

fclose(outFile);

}

The text files are written in plain texts and can be viewed by the editor.

Input.txt

1

2

3

4

5

Output.txt

1

4

9

16

25

We can also read and write binary data. The program below writes the number one to ten to the file Numbers.bin and then reads the same series of values from the file. The functions write and read takes the address of the value to be read or written and the size of the value in bytes. They return the number of bytes actually read or written. When reading, we can check whether we have reached the end of the file by counting the number of read bytes; if it is zero, we have reached the end.

BinaryFile.c

#include <stdlib.h>

#include <assert.h>

#include <stdio.h>

void main(void) {

int index, value;

FILE \*outFile = fopen("Number.bin", "w"), \*inFile;

assert(outFile != NULL);

for (index = 1; index <= 10; ++index) { fwrite(&index, 1, sizeof index, outFile);

}

fclose(outFile);

inFile = fopen("Number.bin", "r");

assert(inFile != NULL);

while (fread(&value, 1, sizeof value, inFile) > 0) { printf("%d\n", value);

}

fclose(inFile);

}

The values are stored in compressed form in the binary file Numbers.bin, why they are not readable in the editor. Here is a screen dump of the file.



### Program Parameters

It possible to start the program execution with parameters. In main, the argc parameter holds the number of input strings, and argv holds the string themselves.

#include <stdio.h>

void main(int argc, char\* argv[]) {

int index;

printf("argc: %d\n", argc);

for (index = 0; index < argc; ++index) {

printf("argv[%d]: %s\n", index, argv[index]);

}

}

The parameters are given when the program executes. Note that the value of argv[0] always is the program path.



### Environment Functions

There is a set of functions that communicates with the surrounding environment: exit quits the execution and sends an integer value to the environment; atexit states a function that is called when exit is called; getenv reads the value of an environment variable, and system executes a system command. There are two macros [EXIT\_SUCCESS](http://www.cplusplus.com/reference/clibrary/cstdlib/EXIT_SUCCESS/) and [EXIT\_FAILURE](http://www.cplusplus.com/reference/clibrary/cstdlib/EXIT_FAILURE/) that can used with the exit function.

#include <stdio.h>

#include <stdlib.h>

void ExitFunction(void) {

printf("The Program is exiting.\n");

}

void main() {

char\* path = getenv("PATH");

system("dir \*.c");

atexit(&ExitFunction);

if (path != NULL) {

printf("Path: %s.\n", path);

exit(EXIT\_SUCCESS);

}

else {

printf("Path not found.\n");

exit(EXIT\_FAILURE);

}

}

Another way to send an exit message to the surrounding environment is to define int as main’s return type. It has the same effect as calling exit in main (except that the exit function will not be called).

#include <stdio.h>

#include <stdlib.h>

int main() {

char\* path = getenv("PATH");

if (path != NULL) {

printf("Path: %s.\n", path);

return EXIT\_SUCCESS;

}

else {

printf("Path not found.\n");

return EXIT\_FAILURE;

}

}

### Searching and Sorting

The bsearch function performs a binary search through a sorted list of values. We need to add a pointer function parameter specifying a function comparing two values.

#include <stdlib.h>

int CompareIntegers(const int\* intPtr1, const int\* intPtr2) {

return (\*intPtr1 < \*intPtr2) ? -1 : (((\*intPtr1 > \*intPtr2) ? 1 : 0));

}

void main() {

int intArray[] = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10},

arraySize = sizeof intArray / sizeof intArray[0],

value = 6;

int\* resultPtr = bsearch(&value, intArray, arraySize,

sizeof (int), &CompareIntegers);

if (resultPtr != NULL) {

int index = ((int\*) resultPtr) - intArray;

printf("Index %d.\n", index);

}

else {

printf("Not found.\n");

}

}

There is also the qsort function that performs a fast sorting.

#include <stdlib.h>

void main() {

int index, intArray[] = {7, 2, 5, 9, 3, 1, 10, 8, 4, 6},

arraySize = sizeof intArray / sizeof intArray[0],

key = 5;

qsort(intArray, arraySize, sizeof (int), &CompareIntegers);

for (index = 0; index < arraySize; ++index) {

printf("%d\n", intArray[index]);

}

}

### Functions with Variable Number of Parameters

In C, there is possible to define functions with a variable number of parameters (such as printf and scanf). One condition is that the function has at least one regular parameter in order for the va\_start macro to hook up to. The va\_arg macro reads the arguments and the va\_end macro ends the reading. Note that the macros do not give any information about the number or types of the arguments. In the following code, we use the first argument count to count the number of arguments thereafter.

#include <stdio.h>

#include <stdarg.h>

void Print(int count, ...) {

va\_list arg\_list;

va\_start(arg\_list, count);

while (count > 0) { int value = va\_arg(arg\_list, int);

printf("%d ", value);

--count;

}

printf("\n");

va\_end(arg\_list);

}

void main(void) {

Print(1, 1);

Print(2, 1, 2);

Print(3, 1, 2, 3);

}

### String Management

There are several functions dealing with strings: strcpy copies a string; strcat adds a string; strlen gives the length of a string; strcmp compares two strings; strchr and strrchr gives the first and last occurrence of a character in a string. All functions look for the finishing null character (‘\0’), strcmp returns a value less than zero if the first string is smaller than second string, a value greater than zero if the second string is larger, and zero if they are equals. Note that the comparison continues until the finishing zero character has been found, even though the character arrays may be longer. As strchr and strrchr return a pointer to the found character (or null if it was not found), we can calculate the array index by subtracting the pointers.

int FirstIndexOf(char s[], char c) {

char\* p = strchr(s, c);

return (p != NULL) ? (p - s) : -1;

}

int LastIndexOf(char s[], char c) {

char\* p = strrchr(s, c);

return (p != NULL) ? (p - s) : -1;

}

void main(void) {

char s1[] = "Hello", s2[] = "World";

char t1[20], t2[20];

strcpy(t1, s1);

strcat(t1, s2);

sprintf(t2, "%s, %s!", s1, s2);

printf("t1: \"%s\", length of t1: %d, t2: \"%s\"\n",

t1, strlen(t1), t2);

printf("Comparing t1 and t2: %d\n", strcmp(t1, t2));

printf("First Index of 'l' in t2: %d, last index of 'l' in t2: %d.\n",

FirstIndexOf(t2, 'l'), LastIndexOf(t2, 'l'));

}

It is also possible to write to a string with sprintf and read from a string with sscanf. These function come in handy when to cast between strings and numerical values.

void main() {

int i;

double x;

char s1[20], s2[20];

sprintf(s1, "%d", 123);

sprintf(s2, "%f", 123.456e-3);

printf("s1 = \"%s\", s2 = \"%s\"\n", s1, s2);

sscanf(s1, "%d", &i);

sscanf(s2, "%lf", &x);

printf("i = %d, x = %f\n", i, x);

}

The functions return the number of values written or read, which gives that we can check that the conversion worked.

if (sscanf(s2, "%lf", &x) == 1) {

printf("x = %f\n", x);

}

else {

printf("Could not cast the string \"%s\".\n", s2);

}

Moreover, we can also count the number of read character and decide whether to whole string has been read.

sscanf(s2, "%lf%n", &x, &charCount);

if (charCount == strlen(s2)) {

printf("x = %f\n", x);

}

else {

printf("Could not cast the string \"%s\".\n", s2);

}

### Memory Management

There is also a set of function for general memory management: memset sets each byte of a memory block to a given value; memcmp compares the memory blocks; memcpy and memmove copies a memory block. The difference between them is that memcpy copies the bytes of the block directly when memmove uses a buffer, which gives that memcpy is faster and memmove is safer when copying overlapping blocks.

These functions work similar to their string counterparts. However, note that they do not look for a finishing character. Instead, the sizes of the function blocks are given as a parameter.

#include <stdlib.h>

void main() {

int a1[3], a2[3];

memset(a1, 1, sizeof a1);

memcpy(a2, a1, sizeof a2);

printf("Comparing a1 and a2: %d\n", memcmp(a1, a2, sizeof a2));

}

### Mathematical Functions

C standard library includes the trigonometric cos, sin, tan, acos, asin, and atan, and atan2; the hyperbolic functions cosh, sinh, and tanh; ceil and floor that rounds a floating value to the closest higher and lower value, respectively; exp returning the natural exponent; log returning the natural logarithm; log10 returning the logarithm of ten; sqrt returning the square root, and pow returning the power of two values. All of these functions takes and returns a double value (except atan2 and pow that takes two values). The trigonometric functions deal with angles in radians.

When converting a floating value to an integral value, it will always be truncated; that is, rounded to the integer value closest to zero. One way to find the closest integer value is to add 0.5 to the double value before converting it.

printf("double %f, rounded to the closest integer value: %d\n",

x, ((int) (x + 0.5)));

### Long Jumps

Regular jumps with goto can only occur in the same function body; long jumps with the standard function setjmp and longjump can, on the other hand, occur throw call sequences; setjmp saves the state of the program execution and longjmp uses that information to restore that state, which gives the illusion of a backwards jump.

In the program below, the user inputs a number, if it does not equal zero its inverse is return. If the number does equal zero, a long jump occurs. The call to setjump returns zero when the jump is set and a non-zero value when a long jump occurs[[22]](#footnote-22).

#include <stdio.h>

#include <setjmp.h>

jmp\_buf env;

double Divide(double numerator, double denominator) {

if (denominator == 0) { longjmp(env, -1); // -1 to represents an error state.

}

return numerator / denominator;

}

double Inverse(double number) {

return Divide(1.0, number);

}

void main() {

if (setjmp(env) == 0) { double value;

printf("Value: ");

scanf("%lf", &value);

printf("Inverse: %f\n", Inverse(value));

}

else {

puts("Division by zero.");

}

}

### Time

The time function gives the number of seconds since January 1, 1970. That value can then be used to fill a tm struct with values regarding local date and time by calling the localtime function. If we call instead call gmtime, the [Greenwich Mean Time](http://sv.wikipedia.org/wiki/Greenwich_Mean_Time) (Coordinated Universal Time) will be given. Note that the tm\_year field gives the year starting from 1900.

#include <time.h>

void main() {

time\_t t = time(NULL);

struct tm time;

char \*months[] = {"Jan", "Feb", "Mar", "Apr",

"May", "Jun", "Jul", "Aug",

"Sep", "Oct", "Nov","Dec"};

char \*days[] = {"Sun", "Mon", "Tue", "Wed",

"Thu", "Fri", "Sat"};

localtime\_s(&time, &t);

printf("%02d:%02d:%02d\n", time.tm\_hour, time.tm\_min, time.tm\_sec);

printf("%s %s %d, %d\n", days[time.tm\_wday], months[time.tm\_mon],

time.tm\_mday, 1900 + time.tm\_year);

printf("Day of year: %d.\n", time.tm\_yday);

printf("Daylight saving time: %s.\n", time.tm\_isdst ? "Yes" : "No");

}

It is also possible to measure the period between two events.

#include <stdio.h>

#include <time.h>

void main() {

time\_t t1 = time(NULL);

int index;

for (index = 0; index < 100000; ++index) {

printf("%d\n", index);

}

{ time\_t t2 = time(NULL);

printf("The loop took %f seconds.\n", difftime(t2, t1));

}

}

### Random Numbers

It is possible to generate a sequence of pseudo-random numbers with the rand function. In order to initialize the sequence we need to sow a random seed with the srand function. One way to generate different seeds is to use the time function. The [RAND\_MAX](http://www.cplusplus.com/reference/clibrary/cstdlib/RAND_MAX/) macro gives the largest possible random numbers; however, usually we want random numbers in a certain interval. The following code generates five sequences of ten numbers between 1 and 100.

#include <stdlib.h>

void main() {

int iIndex1, iIndex2;

for (iIndex1 = 0; iIndex1 < 5; ++iIndex1) {

srand((unsigned int) time(NULL));

for (iIndex2 = 0; iIndex2 < 10; ++iIndex2) {

int randomNumber = rand() % 100 + 1;

printf("%d\n", randomNumber);

}

printf("\n");

}

}

### Limits of Integral and Floating Types

Since the type sizes are implementation dependent, there are a set of macros defining the minimum and maximum values of the integrated and floating types. The minimum values of unsigned integral types are always zero.

#include <stdio.h>

#include <limits.h> // The integral type limit constants.

#include <float.h> // The floating type limit constants.

void main() {

printf("Minimum signed char: %d\n", SCHAR\_MIN);

printf("Maximum signed char: %d\n", SCHAR\_MAX);

printf("Minimum signed short int: %d\n", SHRT\_MIN);

printf("Maximum signed short int: %d\n", SHRT\_MAX);

printf("Minimum signed int: %d\n", INT\_MIN);

printf("Maximum signed int: %d\n", INT\_MAX);

printf("Minimum signed long int: %ld\n", LONG\_MIN);

printf("Maximum signed long int: %ld\n\n", LONG\_MAX);

printf("Maximum unsigned char: %u\n", UCHAR\_MAX);

printf("Maximum unsigned short int: %u\n", USHRT\_MAX);

printf("Maximum unsigned int: %u\n", UINT\_MAX);

printf("Maximum unsigned long int: %lu\n\n", ULONG\_MAX);

printf("Minimum float: %e\n", FLT\_MIN);

printf("Maximum float: %e\n", FLT\_MAX);

printf("Minimum double: %e\n", DBL\_MIN);

printf("Maximum double: %e\n", DBL\_MAX);

printf("Minimum long double: %le\n", LDBL\_MIN);

printf("Maximum long double: %le\n", LDBL\_MAX);

}

### Character Functions

There are a set of functions for classifying a character.

if (isdigit(c)) {

printf("%c is a digit.", c);

}

Below follows a table of the character functions. The ASCII codes are stated within brackets.

|  |  |
| --- | --- |
| Function | Checks |
| **isalpha** | Letter: a-z and A-Z. |
| **islower** | Lower case letter: a-z. |
| **isupper** | Upper case letter: A-Z. |
| **isdigit** | Digit: 0-9. |
| **Isxdigit** | Hexadecimal digit: a-f, a-f, and 0-9. |
| **Isalnum** | Alphanumeric character: letter or digit. |
| **Isspace** | White-space character: space (0x20), horizontal tab (0x09), vertical tab (0x0B), new-line (0x0A), form-feed (0x0D), or carriage return (0x0C). |
| **Iscntrl** | Control character: characters with ASCII codes 0x00 to 0x0F, inclusive, and delete (0x1F). |
| **Isprint** | Printable character: any character that is not a control character. |
| **Isgraph** | Graphical character: any printable character that is not a white-space character. |
| **ispunct** | Punctuation character: any graphical character that is not an alphanumeric character. |

There is also the tolower and toupper functions that returns the character in lower and upper case, respectively.

printf("%c in lower case: %c, and in upper case: %c.",

c, tolower(c), toupper(c));

## Further Reading

I recommend *The C Programming Language* by Brian Kernighan and Dennis Ritchie (2nd Edition, Prentice Hall, 1988). It describes ANSI C and the standard library in a short and consist way, and holds an appendix about the ANSI C standard.

If you want to learn more about C programming, I recommend *C How to Program* by P. J. [Deitel](http://www.amazon.com/s/ref=ntt_athr_dp_sr_1?_encoding=UTF8&sort=relevancerank&search-alias=books&ie=UTF8&field-author=Paul%20J.%20Deitel) (6th Edition, Prentice Hall, 2009), which which discusses in depth theory as well as practice of programming in C, with many clear and comprehensive examples. C in a Nutshell by [Peter Prinz, Tony Crawford](http://shop.oreilly.com/product/9780596006976.do#tab_04) (O'Reilly Media, 2005) and Practical C Programming (3th edition, O'Reilly Media, 1997) are also good choices.

When you have mastered the basic of C programming, I recommend Expert C Programming by P. Van Der Linden (Prentice Hall, 1994). Do not worry; you do not have to be an expert to read it.

# Foreword

The compiler is made up of several components:

#### The Type System

Every programming language has a type system, which determines the properties of its types.

#### The Symbol Table

In many languages, the source code can hold a definition of a variables that is later referred to. In that case the, the variable (its name, type, and potential value) needs to be stored in a symbol table.

#### Scanning

The scanner takes a stream of characters and put them together into a sequence of *tokens*; that is, the least meaningful parts of the source code. For instance, the characters ‘i' and ‘f’ is put together into the keyword **if**, and the characters ‘!’ and ‘=’ is put together into the operator **not\_equal**.

#### Parsing and Middle Code Generation

The syntax of a programming language is often defined by a grammar, and the parser checks whether the token sequence generated by the scanner is accepted by the grammar. Moreover, the parser does also generate middle code.

#### Middle Code Optimization

The middle code generated by the parser often holds unnecessary instructions that needs to be removed.

#### Target Code Generation

When the middle code has been generated and optimized, it is time to generate the target code. In this book, we actually generate target code for two formats: Intel x86 assembly source code and executable code in the 16-bits COM-format.

#### Register Allocation

When the target code has been generated, the registers need to be allocated. We use a graph coloring algorithm to do that.

#### Linking

When each source file has been completely compiled, they need to be linked together into a executable file format.

**The Preprocessor**

The C programming language comes with a preprocessor that needs to be.

**The Standard Library**

The C programming language also comes with a standard library.

3. I have another book proposal for you, this time about compiler technology. The book will describe the design and implementation of a C++ compiler (including the preprocessor, linker, and standard library), with all code given.

So far, I have written a C compiler and I plan to extend it to a C++ compiler in the near future. Even though C++ is a much larger language than C, it not too much extra job. Most of the job has already been done with the C compiler.

The idea of the book is basically the same as the Windows book: I explain the details and provide all source code of the compiler. The source code written in Java with CUP and JLex and generates both assembly code and executable code.

The books below are the latest editions of three classic books. They describe in detail the basic features of compiler construction and to some extent advanced features. The structure of these books is similar to mine, one can say that I have followed these books. However, while these books give many small examples, my book describes in each chapter the features necessary to be included in a C++ compiler.

Alfred V. Aho, Lam, Monica S. Lam, Ravi Sethi, Jeffrey D. Ullman. Compilers – Principles, Techniques, and Tools, 2nd edition. Prentice Hall, 2006.

Keith Cooper, Linda Torczon. Engineering a Compiler, 2nd edition. Morgan Kaufman, 2011.

Charles N. Fischer, Ron K. Cytron, Richard J. LeBlanc. Crafting a Compiler. Pearson Education, 2009.

The following books describe compiler construction in Java, C, and ML. They describe a compiler for a smaller language with excerpt from the code given in the book (the complete code is downloadable). However, these books are briefer than the books above, and describe the compiler features in less detail.

Andrew W Appel. Modern Compiler Implementation in Java, 2nd edition. Cambridge University Press, 2002.

Andrew W Appel. Modern Compiler Implementation in C, 2nd edition. Cambridge University Press, 2004.

Andrew W Appel. Modern Compiler Implementation in ML, 2nd edition. Cambridge University Press, 2004.

The following book could be said to be closely related to mine, as it also described the code for a C compiler. However, this book is more or less unreadable. It presents the code, which (in my option) is unstructured, and it does not include much text describing the code. Hopefully, my book will describe the topic in a clearer way. I also plan to describe theory of a compiler, not just the code to implement it. Moreover, this book does not include the preprocessor, linker, or standard library, and the code is written in C, not Java.

Christopher W. Fraser, David R. Hanson. A Retargetable C Compiler : Design and Implementation. Benjamin Cummings, 1995.

This books deals with the more advanced parts of compiler construction; that is, compiler optimization. Even though I will include some optimization at the end of my book, it does not compete with this book.

Steven S. Muchnick. Advanced Compiler Design & Implementation. Morgan Kaufmann, 2003.

Here is a suggestion for the chapters:

1. Introduction. Introduces the compiler phases described in the following chapters by demonstrating a compiler for a small toy language generating MIPS-code executable in the SPIM simulator.

2. The Scanner. The scanner is a relatively small part of the compiler. Its task is to put together characters into tokens (the smallest significant parts of the source code). Examples are as key words, operators, and numerical values. The scanner is written in JLex, a lexical generator for Java, based on Lex for C.

3. The Parser. The parser, on the other hand, is a large part of the compiler. Its task is to confirm that the given tokens (generated by the scanner) agree with the syntax of the programming language, which is represented by a set of grammatical rules. The parser is defined in CUP, a syntax generator for Java, based on Yacc for C. Each rule can also be equipped with code dealing with type checking and target code generation. However, I will have try to omit as much as possible of the code in this chapter. Most of the code is made up of calls to methods defined in later chapters.

4. Declarations and the Symbol Table. C++ has a rather complicated declaration system with aggregated types such as classes, structures and arrays with a corresponding complicated syntax. All defined variables and functions are stored in the symbol table, which is a hierarchical structure matching the program structure.

5. Type Checking. C++ has a rather large set of operators with complicated rules that need to be checked.

6. Intermediate Code Generation. When the types of expressions and statements are checked, three-address-code are generated, which is a simple intermediate language used to represent the code internally. Type conversation is also included in this chapter.

7. Static and Dynamic Initialization. In C++, variables can be initialized. If the variable is static, the data shall be generated and placed in the static area of the final target code. If it is dynamic, the initialization will result in a series of assignments. One thing that complicates the issue is that it is possible to initialize hierarchical structures made up by structures and arrays with one flat list.

8. Intermediate Code Optimization. The intermediate code can be optimized. For instance, code that is never reached and assignment of variables that are never used shall be removed.

9. Target Code Generation. The target code of the C++ compiler of this book is Intel x86, which is harder to deal with than the MIPS code the first chapter. It holds a few registers and registers of different sizes overlap. Therefore, the register allocation process needs to closely keep track on which variable values that are currently stored in the registers.

10. The Preprocessor. Before the actually compilation starts, the source code has been traversed by the preprocessor that replaces macros with text, includes header files, and provides conditional programming.

11. The Linker. When the target code has finally been generated, it becomes stored into an object file. As the source code can be distributed over several files, the target code need to be merged into one executable file. This is the task of the linker, it merges the code and data area, resolve the static and extern references and generate the executable file.

12. The C Standard Library, made up by functions and macros, which will be included by the linker in the final executable file.

13. The C++ Standard Library, made up by classes, to a certain extent is based on C standard library.

The compiler is made up of several phases:

Preprocessor

Source Code

Parser

Processed Code

Syntax Tree Optimister

Syntax Tree

Middle Code Generator

Optimized Syntax Tree

Middle Code Optimizer

Middle Code List

Optimized Middle Code List

Scanner

Symbol Table

Preprocessor

Source Code

Parser

Processed Code

Syntax Tree Optimister

Syntax Tree

Middle Code Generator

Optimized Syntax Tree

Middle Code Optimizer

Middle Code List

Optimized Middle Code List

Scanner

Symbol Table

1. BCPL stands for Basic Command Programming Language and was developed at join effort at London University and Cambridge University. Later on, BCPL has jokely been said to be an acronym for the Before C Programming Language. [↑](#footnote-ref-1)
2. Actually, their program wrote "hello, world\n" without a capital letter or an exclamation mark. Their example was inherited from the internal 1974 memorandum Programming in C: A Tutorial by [Brian Kernighan](http://en.wikipedia.org/wiki/Brian_Kernighan). The oldest known instance of the usage of the words ”hello” and ”world” together occurred in Kernighan’s 1972 tutorial *Introduction to the Language* [B](http://en.wikipedia.org/wiki/B_(programming_language)), even though the syntax was different from C. [↑](#footnote-ref-2)
3. Technically, the compiler replaces every block comment with a single space character, which makes it unwise to place a comment inside a name. [↑](#footnote-ref-3)
4. Usually, the problem is solved by the following macros (see Section 14.29):

   #define BOOL int

   #define TRUE 1

   #define FALSE 0 [↑](#footnote-ref-4)
5. Technically, it is given the value that happened to be stored on the memory address. [↑](#footnote-ref-5)
6. Depending of the compiler and its warning level settings. [↑](#footnote-ref-6)
7. Actually, f represent both float or double. The reason is that a float value is always converted to double when used as input to printf (or any other function). [↑](#footnote-ref-7)
8. As we inputs pointers to the variables rather than the variables themselves when calling scanf, no conversation from float to double occur. That is why f represents float and lf represents double. That lf (long float) represents double in scanf is an emergency solution since d represents an integer in decimal notation in both printf and scanf. [↑](#footnote-ref-8)
9. A white-space character is a space (‘ ‘), horizontal tabulator (‘\t’), vertical tabulator (‘\v’), new-line (‘\n’), form-feed (‘\f’), or carriage return (‘\r’). [↑](#footnote-ref-9)
10. Technically, each enumeration value holds the type unsigned int. [↑](#footnote-ref-10)
11. However, we have no knowledge about the actual address; it may be 10,000 or anything else. [↑](#footnote-ref-11)
12. In C++, the ampersand is also used to define references variables, which are not included in C. [↑](#footnote-ref-12)
13. More accurately, the number of bytes needed to hold a value of the type, [↑](#footnote-ref-13)
14. They were added to C in order to match two specific assembler instructions. Since then, they have stuck. More recent languages, such as C++, Java, and C#, do also support increment and decrement. [↑](#footnote-ref-14)
15. Some programmers prefer to denote equal to and not equal to as comparison operators. [↑](#footnote-ref-15)
16. In C#, the break statement is mandatory. [↑](#footnote-ref-16)
17. Technically, main is a regular function. The only difference, compared to other functions, is that the linker generates code that starts the program execution by calling main. Therefore, the program must include exactly one **main** function. [↑](#footnote-ref-17)
18. In C++, it is possible to access the global variable by prefixing colons, like ::number. [↑](#footnote-ref-18)
19. There is no rational explanation, just accept it. [↑](#footnote-ref-19)
20. Note the distinction of variable definition and declaration: a definition reserves memory for the variable while a declaration is simple a notification to the compiler. [↑](#footnote-ref-20)
21. In object-oriented languages the pointer is hidden and referred to as the this pointer. [↑](#footnote-ref-21)
22. This construction performs the same task as exception handling does in object-oriented languages. [↑](#footnote-ref-22)