x86 Disassembly

Exploring the relationship between C, x86 Assembly, and Machine Code

Contents

0.1 Wikibooks:Collections Preface

This book was created by volunteers at Wikibooks (http://en.wikibooks.org).

0.1.1 What is Wikibooks?

Started in 2003 as an offshoot of the popular Wikipedia project, Wikibooks is a free, collaborative wiki website dedicated to creating high-quality textbooks and other educational books for students around the world. In addition to English, Wikibooks is available in over 130 languages, a complete listing of which can be found at http://www.wikibooks.org. Wikibooks is a "wiki", which means anybody can edit the content there at any time. If you find an error or omission in this book, you can log on to Wikibooks to make corrections and additions as necessary. All of your changes go live on the website immediately, so your effort can be enjoyed and utilized by other readers and editors without delay.

Books at Wikibooks are written by volunteers, and can be accessed and printed for free from the website. Wikibooks is operated entirely by donations, and a certain portion of proceeds from sales is returned to the Wikimedia Foundation to help keep Wikibooks running smoothly. Because of the low overhead, we are able to produce and sell books for much cheaper then proprietary textbook publishers can. **This book can be edited by anybody at any time, including you.** We don't make you wait two years to get a new edition, and we don't stop selling old versions when a new one comes out.

Note that Wikibooks is not a publisher of books, and is not responsible for the contributions of its volunteer editors. PediaPress.com is a print-on-demand publisher that is also not responsible for the content that it prints. Please see our disclaimer for more information: http://en.wikibooks.org/wiki/Wikibooks:General_disclaimer .

0.1.2 What is this book?

This book was generated by the volunteers at Wikibooks, a team of people from around the world with varying backgrounds. The people who wrote this book may not be experts in the field. Some may not even have a passing familiarity with it. The result of this is that some information in this book may be incorrect, out of place, or misleading. For this reason, you should never rely on a community-edited Wikibook when dealing in matters of medical, legal, financial, or other importance. Please see our disclaimer for more details on this.

Despite the warning of the last paragraph, however, books at Wikibooks are continuously edited and improved. If errors are found they can be corrected immediately. If you find a problem in one of our books, we ask that you **be bold** in fixing it. You don't need anybody's permission to help or to make our books better.

Wikibooks runs off the assumption that many eyes can find many errors, and many able hands can fix them. Over time, with enough community involvement, the books at Wikibooks will become very high-quality indeed. **You are invited to participate at Wikibooks to help make our books better**. As you find problems in your book don't just complain about them: Log on and fix them! This is a kind of proactive and interactive reading experience that you probably aren't familiar with yet, so log on to http://en.wikibooks.org and take a look around at all the possibilities. We promise that we won't bite!

0.1.3 Who are the authors?

The volunteers at Wikibooks come from around the world and have a wide range of educational and professional backgrounds. They come to Wikibooks for different reasons, and perform different tasks. Some Wikibookians are prolific authors, some are perceptive editors, some fancy illustrators, others diligent organizers. Some Wikibookians find and remove spam, vandalism, and other nonsense as it appears. Most wikibookians perform a combination of these jobs.

It's difficult to say who are the authors for any particular book, because so many hands have touched it and so many changes have been made over time. It's not unheard of for a book to have been edited thousands of times by hundreds of authors and editors. *You could be one of them too*, if you're interested in helping out.

0.1.4 Wikibooks in Class

Books at Wikibooks are free, and with the proper editing and preparation they can be used as cost-effective textbooks in the classroom or for independent learners. In addition to using a Wikibook as a traditional read-only learning aide, it can also become an interactive class

project. Several classes have come to Wikibooks to write new books and improve old books as part of their normal course work. In some cases, the books written by students one year are used to teach students in the same class next year. Books written can also be used in classes around the world by students who might not be able to afford traditional textbooks.

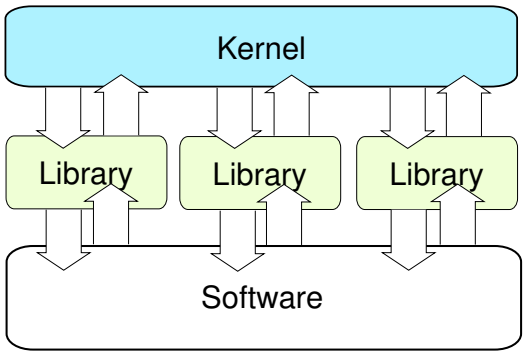
0.1.5 Happy Reading!

We at Wikibooks have put a lot of effort into these books, and we hope that you enjoy reading and learning from them. We want you to keep in mind that what you are holding is not a finished product but instead a work in progress. These books are never "finished" in the traditional sense, but they are ever-changing and evolving to meet the needs of readers and learners everywhere. Despite this constant change, we feel our books can be reliable and high-quality learning tools at a great price, and we hope you agree. Never hesitate to stop in at Wikibooks and make some edits of your own. We hope to see you there one day. **Happy reading!**

0.2 Cover

The Wikibook of

x86 Disassembly
Using C and Assembly Language



From Wikibooks: The Free Library

0.3 x86 Disassembly/Introduction

0.3.1 What Is This Book About?

This book is about the disassembly of x86 machine code into human-readable assembly, and the decompilation of x86 assembly code into human-readable C or C++ source code. Some topics covered will be common to all computer architectures, not just x86-compatible machines.

0.3.2 What Will This Book Cover?

This book is going to look in-depth at the disassembly and decompilation of x86 machine code and assembly code. We are going to look at the way programs are made using assemblers and compilers, and examine the way that assembly code is made from C or C++ source code. Using this knowledge, we will try to reverse the process. By examining common structures, such as data and control structures, we can find patterns that enable us to disassemble and decompile programs quickly.

0.3.3 Who Is This Book For?

This book is for readers at the undergraduate level with experience programming in x86 Assembly and C or C++. This book is not designed to teach assembly language programming, C or C++ programming, or compiler/assembler theory.

0.3.4 What Are The Prerequisites?

The reader should have a thorough understanding of x86 Assembly, C Programming, and possibly C++ Programming. This book is intended to increase the reader's understanding of the relationship between x86 machine code, x86 Assembly Language, and the C Programming Language. If you are not too familar with these topics, you may want to reread some of the above-mentioned books before continuing.

0.3.5 What is Disassembly?

Computer programs are written originally in a human readable code form, such as assembly language or a high-level language. These programs are then compiled into a binary format called **machine code**. This binary format is not directly readable or understandable by humans. Many programs -- such as malware, proprietary commercial programs, or very old legacy programs -- may not have the source code available to you.

Programs frequently perform tasks that need to be duplicated, or need to be made to interact with other programs. Without the source code and without adequate documentation, these tasks can be difficult to accomplish. This book outlines tools and techniques for attempting to convert the raw machine code of an executable file into equivalent code in assembly language and the high-level languages C and C++. With the high-level code to perform a particular task, several things become possible:

1. Programs can be ported to new computer platforms, by compiling the source code in a different environment.
2. The algorithm used by a program can be determined. This allows other programs to make use of the same algorithm, or for updated versions of a program to be rewritten without needing to track down old copies of the source code.
3. Security holes and vulnerabilities can be identified and patched by users without needing access to the original source code.
4. New interfaces can be implemented for old programs. New components can be built on top of old components to speed development time and reduce the need to rewrite large volumes of code.
5. We can figure out what a piece of malware does. We hope this leads us to figuring out how to block its harmful effects. Unfortunately, some malware writers use self-modifying code techniques (polymorphic camouflage, XOR encryption, scrambling)[1], apparently to make it difficult to even detect that malware, much less disassemble it.

Disassembling code has a large number of practical uses. One of the positive side effects of it is that the reader will gain a better understanding of the relation between machine code, assembly language, and high-level languages. Having a good knowledge of these topics will help programmers to produce code that is more efficient and more secure.

[1] "How does a crypter for bypass antivirus detection work?"

Chapter 1

Tools

1.1 x86 Disassembly/Assemblers and Compilers

1.1.1 Assemblers

Assemblers are significantly simpler than compilers, and are often implemented to simply translate the assembly code to binary machine code via one-to-one correspondence. Assemblers rarely optimize beyond choosing the shortest form of an instruction or filling delay slots.

Because assembly is such a simple process, disassembly can often be just as simple. Assembly instructions and machine code words have a one-to-one correspondence, so each machine code word will exactly map to one assembly instruction. However, disassembly has some other difficulties which cannot be accounted for using simple code-word lookups. We will introduce assemblers here, and talk about disassembly later.

1.1.2 Assembler Concepts

Assemblers, on a most basic level, translate assembly instructions into machine code with a one to one correspondence. They can also translate named variables into hard-coded memory addresses and labels into their relative code addresses.

Assemblers, in general, do not perform code optimization. The machine code that comes out of an assembler is equivalent to the assembly instructions that go into the assembler. Some assemblers have high-level capabilities in the form of *Macros.*

Some information about the program is lost during the assembly process. First and foremost, program data is stored in the same raw binary format as the machine code instructions. This means that it can be difficult to determine which parts of the program are actually instructions. Notice that you can disassemble raw data, but the resultant assembly code will be nonsensical. Second, textual information from the assembly source code file, such as variable names, label names, and code comments are all destroyed during assembly. When you disassemble the code, the instructions will be the same, but all the other helpful information will be lost. The code will be accurate, but more difficult to read.

Compilers, as we will see later, cause even more information to be lost, and decompiling is often so difficult and convoluted as to become nearly impossible to do accurately.

1.1.3 Intel Syntax Assemblers

Because of the pervasiveness of Intel-based IA-32 microprocessors in the home PC market, the majority of assembly work done (and the majority of assembly work considered in this wikibook) is x86-based. Many of these assemblers (or new versions of them) can handle amd64/x86_64/EMT64 code as well, although this wikibook will focus primarily on 32 bit (x86/IA-32) code examples.

MASM

MASM is Microsoft's assembler, an abbreviation for "Macro Assembler." However, many people use it as an acronym for "Microsoft Assembler," and the difference isn't a problem at all. MASM has a powerful macro feature, and is capable of writing very low-level syntax, and pseudo-high-level code with its macro feature. MASM 6.15 is currently available as a free-download from Microsoft, and MASM 7.xx is currently available as part of the Microsoft platform DDK.

- MASM uses Intel Syntax.
- MASM is used by Microsoft to implement some low-level portions of its Windows Operating systems.
- MASM, contrary to popular belief, has been in constant development since 1980, and is upgraded on a needs-basis.
- MASM has always been made compatible by Microsoft to the current platform, and executable file types.

- MASM currently supports all Intel instruction sets, including SSE2.

Many users love MASM, but many more still dislike the fact that it isn't portable to other systems.

TASM

TASM, Borland's "Turbo Assembler," is a functional assembler from Borland that integrates seamlessly with Borland's other software development tools. Current release version is version 5.0. TASM syntax is very similar to MASM, although it has an "IDEAL" mode that many users prefer. TASM is not free.

NASM

NASM, the "Netwide Assembler," is a free, portable, and retargetable assembler that works on both Windows and Linux. It supports a variety of Windows and Linux executable file formats, and even outputs pure binary. NASM is not as "mature" as either MASM or TASM, but is:

- more portable than MASM
- cheaper than TASM
- strives to be very user-friendly

NASM comes with its own disassembler ndisasm, and supports 64-bit (x86-64/x64/AMD64/Intel 64) CPUs.

NASM is released under the LGPL.

FASM

FASM, the "Flat Assembler" is an open source assembler that supports x86, and IA-64 Intel architectures.

1.1.4 (x86) AT&T Syntax Assemblers

AT&T syntax for x86 microprocessor assembly code is not as common as Intel-syntax, but the GNU Assembler (GAS) uses it, and it is the *de facto* assembly standard on Unix and Unix-like operating systems.

GAS

The GNU Assembler (GAS) is the default back-end to the GNU Compiler Collection (GCC) suite. As such, GAS is as portable and retargetable as GCC is. However, GAS uses the AT&T syntax for its instructions as default, which some users find to be less readable than Intel syntax. Newer versions of gas can be switched to Intel syntax with the directive ".intel_syntax noprefix".

GAS is developed specifically to be used as the GCC backend. Because GCC always feeds it syntactically correct code, GAS often has minimal error checking.

GAS is available as a part of either the GCC package or the GNU binutils package.

1.1.5 Other Assemblers

HLA

HLA, short for "High Level Assembler" is a project spearheaded by Randall Hyde to create an assembler with high-level syntax. HLA works as a front-end to other assemblers such as FASM (the default), MASM, NASM, and GAS. HLA supports "common" assembly language instructions, but also implements a series of higher-level constructs such as loops, if-then-else branching, and functions. HLA comes complete with a comprehensive standard library.

Since HLA works as a front-end to another assembler, the programmer must have another assembler installed to assemble programs with HLA. HLA code output therefore, is as good as the underlying assembler, but the code is much easier to write for the developer. The high-level components of HLA may make programs less efficient, but that cost is often far outweighed by the ease of writing the code. HLA high-level syntax is very similar in many respects to Pascal, which in turn is itself similar in many respects to C, so many high-level programmers will immediately pick up many of the aspects of HLA.

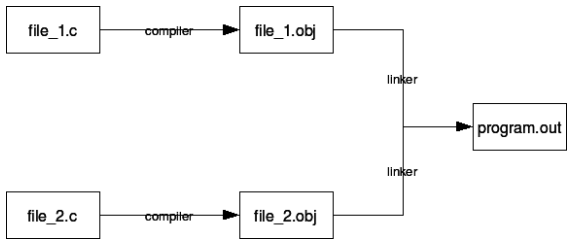
Here is an example of some HLA code:

mov(src, dest); // C++ style comments pop(eax); push(ebp); for(mov(0, ecx); ecx < 10; inc(ecx)) do mul(ecx); endfor;

Some disassemblers and debuggers can disassemble binary code into HLA-format, although none can faithfully recreate the HLA macros.

1.1.6 Compilers

A compiler is a program that converts instructions from one language into equivalent instructions in another language. There is a common misconception that a compiler always directly converts a high level language into machine language, but this isn't always the case. Many compilers convert code into assembly language, and a few even convert code from one high level language into another. Common examples of compiled languages are: C/C++, Fortran, Ada, and Visual Basic. The figure below shows the common compile-time steps to building a program using the C programming language. The compiler produces object files which are linked to form the final executable:



For the purposes of this book, we will only be considering the case of a compiler that converts C or C++ into assembly code or machine language. Some compilers, such as the Microsoft C compiler, compile C and C++ source code directly into machine code. GCC on the other hand compiles C and C++ into assembly language, and an assembler is used to convert that into the appropriate machine code. From the standpoint of a disassembler, it does not matter exactly how the original program was created. Notice also that it is not possible to exactly reproduce the C or C++ code used originally to create an executable. It is, however, possible to create code that compiles identically, or code that performs the same task.

C language statements do not share a one to one relationship with assembly language. Consider that the following C statements will typically all compile into the same assembly language code:

*arrayA = arrayB[x++]; *arrayA = arrayB[x]; x++; arrayA[0] = arrayB[x++]; arrayA[0] = arrayB[x]; x++;

Also, consider how the following loop constructs perform identical tasks, and are likely to produce similar or even identical assembly language code:

for(;;) { ... } while(1) { ... } do { ... } while(1)

1.1.7 Common C/C++ Compilers

The purpose of this section is to list some of the most common C and C++ compilers in use for developing *production-level* software. There are many many C compilers in the world, but the reverser doesn't need to consider all cases, especially when looking at professional software. This page will discuss each compiler's strengths and weaknesses, its availability (download sites or cost information), and it will also discuss how to generate an assembly listing file from each compiler.

Microsoft C Compiler

The Microsoft C compiler is available from Microsoft for free as part of the Windows Server 2003 SDK. It is the same compiler and library as is used in MS Visual Studio, but doesn't come with the fancy IDE. The MS C Compiler has a very good optimizing engine. It compiles C and C++, and has the option to compile C++ code into MSIL (the .NET bytecode).

Microsoft's compiler only supports Windows systems, and Intel-compatible 16/32/64 bit architectures.

The Microsoft C compiler is **cl.exe** and the linker is **link.exe**

Listing Files In this wikibook, cl.exe is frequently used to produce assembly listing files of C source code. To produce an assembly listing file yourself, use the syntax:

cl.exe /Fa<assembly file name> <C source file>

The "/Fa" switch is the command-line option that tells the compiler to produce an assembly listing file.

For example, the following command line:

cl.exe /FaTest.asm Test.c

would produce an assembly listing file named "Test.asm" from the C source file "Test.c". Notice that there is no space between the /Fa switch and the name of the output file.

GNU C Compiler

The GNU C compiler is part of the GNU Compiler Collection (GCC) suite. This compiler is available for most systems and it is free software. Many people use it exclusively so that they can support many platforms with just one compiler to deal with. The GNU GCC Compiler is the *de facto* standard compiler for Linux and Unix systems. It is retargetable, allowing for many input languages (C, C++, Obj-C, Ada, Fortran, etc...), and supporting multiple target OSes and architectures. It optimizes well, but has a non-aggressive IA-32 code generation engine.

The GCC frontend program is "gcc" ("gcc.exe" on Windows) and the associated linker is "ld" ("ld.exe" on Windows). Windows cmd searches for the programs with ".exe" extensions automatically, so you don't need to type the filename extension.

Listing Files To produce an assembly listing file in GCC, use the following command line syntax:

gcc -S /path/to/sourcefile.c

For example, the following commandline:

gcc -S test.c

will produce an assembly listing file named "test.s". Assembly listing files generated by GCC will be in GAS format. On x86 you can select the syntax with -masm=intel or -masm=att. GCC listing files are frequently not as well commented and laid-out as are the listing files for cl.exe.

You may add `-g3` flags to enable source-code-level de-

bugging symbols so you can see the line numbers in the listing. The -fno-asynchronous-unwind-tables flag can help eliminate some macros in the listing.

Intel C Compiler

This compiler is used only for x86, x86-64, and IA-64 code. It is available for both Windows and Linux. The Intel C compiler was written by the people who invented the original x86 architecture: Intel. Intel's development tools generate code that is tuned to run on Intel microprocessors, and is intended to squeeze every last ounce of speed from an application. AMD IA-32 compatible processors are not guaranteed to get the same speed boosts because they have different internal architectures.

Metrowerks CodeWarrior

This compiler is commonly used for classic MacOS and for embedded systems. If you try to reverse-engineer a piece of consumer electronics, you may encounter code generated by Metrowerks CodeWarrior.

Green Hills Software Compiler

This compiler is commonly used for embedded systems. If you try to reverse-engineer a piece of consumer electronics, you may encounter code generated by Green Hills C/C++.

1.2 x86 Disassembly/Disassemblers and Decompilers

1.2.1 What is a Disassembler?

In essence, a **disassembler** is the exact opposite of an assembler. Where an assembler converts code written in an assembly language into binary machine code, a disassembler reverses the process and attempts to recreate the assembly code from the binary machine code.

Since most assembly languages have a one-to-one correspondence with underlying machine instructions, the process of disassembly is relatively straight-forward, and a basic disassembler can often be implemented simply by reading in bytes, and performing a table lookup. Of course, disassembly has its own problems and pitfalls, and they are covered later in this chapter.

Many disassemblers have the option to output assembly language instructions in Intel, AT&T, or (occasionally) HLA syntax. Examples in this book will use Intel and AT&T syntax interchangeably. We will typically not use HLA syntax for code examples, but that may change in the future.

1.2.2 x86 Disassemblers

Here we are going to list some commonly available disassembler tools. Notice that there are professional disassemblers (which cost money for a license) and there are freeware/shareware disassemblers. Each disassembler will have different features, so it is up to you as the reader to determine which tools you prefer to use.

Online Disassemblers

ODA is a free, web-based disassembler for a wide variety of architectures. You can use "Live View" to see how code is disassembled in real time, one byte at a time, or upload a file. The site is currently in beta release but will hopefully only get better with time.

http://www.onlinedisassembler.com

Commercial Windows Disassemblers

IDA Pro is a professional disassembler that is expensive, extremely powerful, and has a whole slew of features. The downside to IDA Pro is that it costs $515 US for the standard single-user edition. As such this wikibook will not consider IDA Pro specifically because the price tag is exclusionary. Freeware versions do exist; see below.

- (version 6.x) http://www.hex-rays.com/idapro/

Hopper Disassembler is a reverse engineering tool for the Mac, that lets you disassemble, decompile and debug 32/64bits Intel Mac executables. It can also disassemble and decompile Windows executables.

http://www.hopperapp.com

OBJ2ASM is an object file disassembler for 16 and 32 bit x86 object files in Intel OMF, Microsoft COFF format, Linux ELF or Mac OS X Mach-O format.

http://www.digitalmars.com/ctg/obj2asm.html

PE Explorer is a disassembler that "focuses on ease of use, clarity and navigation." It isn't as feature-filled as IDA Pro and carries a smaller price tag to offset the missing functionality: $130

http://www.heaventools.com/PE_Explorer_disassembler.htm

W32DASM W32DASM was an excellent 16/32 bit disassembler for Windows, it seems it is no longer developed. the latest version available is from 2003. the website went down and no replacement went up.

http://www.softpedia.com/get/Programming/Debuggers-Decompilers-Dissasemblers/WDASM.shtml

Commercial Freeware/Shareware Windows Disassemblers

OllyDbg OllyDbg is one of the most popular disassemblers recently. It has a large community and a wide variety of plugins available. It emphasizes binary code analysis. Supports x86 instructions only (no x86_64 support for now, although it is on the way).

http://www.ollydbg.de/ (official website)

http://www.openrce.org/downloads/browse/OllyDbg_Plugins (plugins)

http://www.ollydbg.de/odbg64.html (64 bit version)

Free Windows Disassemblers

Capstone Capstone is an open source disassembly framework for multi-arch (including support for x86, x86_64) & multi-platform with advanced features.

http://www.capstone-engine.org/

Objconv A command line disassembler supporting 16, 32, and 64 bit x86 code. Latest instruction set (SSE4, AVX, XOP, FMA, etc.), several object file formats, several assembly syntax dialects. Windows, Linux, BSD, Mac. Intelligent analysis.

- http://www.agner.org/optimize/#objconv

IDA 3.7 A DOS GUI tool that behaves very much like IDA Pro, but is considerably more limited. It can disassemble code for the Z80, 6502, Intel 8051, Intel i860, and PDP-11 processors, as well as x86 instructions up to the 486.

- http://www.simtel.net/product.php (search for **ida37fw**)

IDA Pro Freeware Behaves almost exactly like IDA Pro, but disassembles only Intel x86 opcodes and is Windows-only. It can disassemble instructions for those processors available as of 2003. Free for non-commercial use.

- (version 4.1) http://www.themel.com/idafree.zip
- (version 4.3) http://www.datarescue.be/idafreeware/freeida43.exe
- (version 5.0) http://www.hex-rays.com/idapro/idadownfreeware.htm

BORG Disassembler BORG is an excellent Win32 Disassembler with GUI.

http://www.caesum.com/

HT Editor An analyzing disassembler for Intel x86 instructions. The latest version runs as a console GUI program on Windows, but there are versions compiled for Linux as well.

http://hte.sourceforge.net/

diStorm64 diStorm is an open source highly optimized stream disassembler library for 80x86 and AMD64.

http://ragestorm.net/distorm/

crudasm crudasm is an open source disassembler with a variety of options. It is a work in progress and is bundled with a partial decompiler.

http://code.google.com/p/vm64dec/downloads/list

BeaEngine BeaEngine is a complete disassembler library for IA-32 and intel64 architectures (coded in C and usable in various languages : C, Python, Delphi, PureBasic, WinDev, masm, fasm, nasm, GoAsm).

http://www.beaengine.org

Visual DuxDebugger is a 64-bit debugger disassembler for Windows.

http://www.duxcore.com/products.html

BugDbg is a 64-bit user-land debugger designed to debug native 64-bit applications on Windows.

http://www.pespin.com/

DSMHELP Disassemble Help Library is a disassembler library with single line Epimorphic assembler. Supported instruction sets - Basic,System,SSE,SSE2,SSE3,SSSE3,SSE4,SSE4A,MMX,FPU,3DNO

http://dsmhelp.narod.ru/ (in Russian)

Unix Disassemblers

Many of the Unix disassemblers, especially the open source ones, have been ported to other platforms, like Windows (mostly using MinGW or Cygwin). Some Disassemblers like otool (OS X) is distro-specific.

Capstone Capstone is an open source disassembly framework for multi-arch (including support for x86, x86_64) & multi-platform (including Mac OSX, Linux, *BSD, Android, iOS, Solaris) with advanced features.

http://www.capstone-engine.org/

Bastard Disassembler The Bastard disassembler is a powerful, scriptable disassembler for Linux and FreeBSD.

http://bastard.sourceforge.net/

ndisasm NASM's disassembler for x86 and x86-64. Works on DOS, Windows, Linux, Mac OS X and various other systems.

udis86 Disassembler Library for x86 and x86-64

http://udis86.sourceforge.net/

Verteron Disassembler Engine (VDE) Fast and lightweight x86/x86-64 disassembler library.

https://github.com/flobernd/verteron-disassembler-engine/

Objconv See above.

ciasdis The official name of ciasdis is *computer_intelligence_assembler_disassembler*. This Forth-based tool allows to incrementally and interactively build knowledge about a code body. It is unique that all disassembled code can be re-assembled to the exact same code. Processors are 8080, 6809, 8086, 80386, Pentium I en DEC Alpha. A scripting facility aids in analyzing Elf and MSDOS headers and makes this tool extendable. The Pentium I ciasdis is available as a binary image, others are in source form, loadable onto lina Forth, available from the same site.

http://home.hccnet.nl/a.w.m.van.der.horst/ciasdis.html

objdump comes standard, and is typically used for general inspection of binaries. Pay attention to the relocation option and the dynamic symbol table option.

gdb comes standard, as a debugger, but is very often used for disassembly. If you have loose hex dump data that you wish to disassemble, simply enter it (interactively) over top of something else or compile it into a program as a string like so: char foo[] = {0x90, 0xcd, 0x80, 0x90, 0xcc, 0xf1, 0x90};

lida linux interactive disassembler an interactive disassembler with some special functions like a crypto analyzer. Displays string data references, does code flow analysis, and does not rely on objdump. Utilizes the Bastard disassembly library for decoding single opcodes. The project was started in 2004 and remains dormant to this day.

http://lida.sourceforge.net

dissy This program is a interactive disassembler that uses objdump.

http://code.google.com/p/dissy/

EmilPRO replacement for the deprecated dissy disassembler.

http://github.com/SimonKagstrom/emilpro

x86dis This program can be used to display binary streams such as the boot sector or other unstructured binary files.

ldasm LDasm (Linux Disassembler) is a Perl/Tk-based GUI for objdump/binutils that tries to imitate the 'look and feel' of W32Dasm. It searches for cross-references (e.g. strings), converts the code from GAS to a MASM-like style, traces programs and much more. Comes along with PTrace, a process-flow-logger.

http://www.feedface.com/projects/ldasm.html

llvm LLVM has two interfaces to its disassembler:

llvm-objdump Mimics GNU objdump.

llvm-mc See the LLVM blog. Example usage:

```
$ echo '1 2' | llvm-mc -disassemble -
triple=x86_64-apple-darwin9
addl %eax, (%rdx)
$ echo '0x0f 0x1 0x9' | llvm-mc -disassemble
-triple=x86_64-apple-darwin9
sidt (%rcx)
$ echo '0x0f 0xa2' | llvm-mc -disassemble
-triple=x86_64-apple-darwin9
cpuid
$ echo '0xd9 0xff' | llvm-mc -disassemble
-triple=i386-apple-darwin9
fcos
```

otool OS X's object file displaying tool.

1.2.3 Disassembler Issues

As we have alluded to before, there are a number of issues and difficulties associated with the disassembly process. The two most important difficulties are the division between code and data, and the loss of text information.

Separating Code from Data

Since data and instructions are all stored in an executable as binary data, the obvious question arises: how can a disassembler tell code from data? Is any given byte a variable, or part of an instruction?

The problem wouldn't be as difficult if data were limited to the .data section (segment) of an executable (explained in a later chapter) and if executable code were limited to the .code section of an executable, but this is often not the case. Data may be inserted directly into the code section (e.g. jump address tables, constant strings), and executable code may be stored in the data section (although new systems are working to prevent this for security reasons). AI programs, LISP or Forth compilers may not contain .text and .data sections to help decide, and have code and data interspersed in a single section that is readable, writable and executable, Boot code may even require substantial effort to identify sections. A technique that is often used is to identify the entry point of an executable, and find all code reachable from there, recursively. This is known as “code crawling”.

Many interactive disassemblers will give the user the option to render segments of code as either code or data, but non-interactive disassemblers will make the separation automatically. Disassemblers often will provide the instruction AND the corresponding hex data on the same line, shifting the burden for decisions about the nature of the code to the user. Some disassemblers (e.g. ciasdis) will allow you to specify rules about whether to disassemble as data or code and invent label names, based on the content of the object under scrutiny. Scripting your own “crawler” in this way is more efficient; for large programs interactive disassembling may be impractical to the point of being unfeasible.

The general problem of separating code from data in arbitrary executable programs is equivalent to the halting problem. As a consequence, it is not possible to write a disassembler that will correctly separate code and data for all possible input programs. Reverse engineering is full of such theoretical limitations, although by Rice's theorem all interesting questions about program properties are undecidable (so compilers and many other tools that deal with programs in any form run into such limits as well). In practice a combination of interactive and automatic analysis and perseverance can handle all but programs specifically designed to thwart reverse engineering, like using encryption and decrypting code just prior to use, and moving code around in memory.

Lost Information

User defined textual identifiers, such as variable names, label names, and macros are removed by the assembly process. They may still be present in generated object files, for use by tools like debuggers and relocating linkers, but the direct connection is lost and re-establishing that connection requires more than a mere disassembler. Especially small constants may have more than one possible name. Operating system calls (like dll's in MS-Windows, or syscalls in Unices) may be reconstructed, as their names appear in a separate segment or are known beforehand. Many disassemblers allow the user to attach a name to a label or constant based on his understanding of the code. These identifiers, in addition to comments in the source file, help to make the code more readable to a human, and can also shed some clues on the purpose of the code. Without these comments and identifiers, it is harder to understand the purpose of the source code, and it can be difficult to determine the algorithm being used by that code. When you combine this problem with the possibility that the code you are trying to read may, in reality, be data (as outlined above), then it can be ever harder to determine what is going on.

1.2.4 Decompilers

Akin to Disassembly, **Decompilers** take the process a step further and actually try to reproduce the code in a high level language. Frequently, this high level language is C, because C is simple and primitive enough to facilitate the decompilation process. Decompilation does have its drawbacks, because lots of data and readability constructs are lost during the original compilation process, and they cannot be reproduced. Since the science of decompilation is still young, and results are “good” but not “great”, this page will limit itself to a listing of decompilers, and a general (but brief) discussion of the possibilities of decompilation.

Decompilation: Is It Possible?

In the face of optimizing compilers, it is not uncommon to be asked “Is decompilation even possible?" To some degree, it usually is. Make no mistake, however: an optimizing compiler results in the irretrievable loss of information. An example is in-lining, a subroutine call is removed and the actual code is put in its place. A further optimization will combine that code with its surroundings, such that the places where the original subroutine is called are not even similar. An optimizer that reverses that process is comparable to an artificial intelligence program that recreates a poem in a different language. So perfectly operational decompilers are a long way off. At most, current Decompilers can be used as simply an aid for the reverse engineering process leaving lots of arduous work.

Common Decompilers

Hex-Rays Decompiler Hex-Rays is a commercial decompiler. It is made as an extension to popular IDA-Pro disassembler. It is currently the only viable commercially available decompiler which produces usable results. It supports both x86 and ARM architecture.

http://www.hex-rays.com/products/decompiler/index.shtml

DCC Decompiler Dcc is an excellent theoretical look at de-compilation, but currently it only supports small files.

http://www.itee.uq.edu.au/~{}cristina/dcc.html{[}deadlink]

mirrors: ;

Boomerang Decompiler Project Boomerang Decompiler is an attempt to make a powerful, retargetable decompiler. So far, it only decompiles into C with moderate success.

http://boomerang.sourceforge.net/

Reverse Engineering Compiler (REC) REC is a powerful "decompiler" that decompiles native assembly code into a *C-like* code representation. The code is half-way between assembly and C, but it is much more readable than the pure assembly is. Unfortunately the program appears to be rather unstable.

http://www.backerstreet.com/rec/rec.htm

ExeToC ExeToC decompiler is an interactive decompiler that boasts pretty good results.

http://sourceforge.net/projects/exetoc

Decompile-It Decompile-It was a web-based decompiler for 32-bit Linux x86 executables compiled with -g

http://decompile-it.com

C4Decompiler C4Decompiler is an interactive, static decompiler under development (Alpha in 2013). It performs global analysis of the binary and presents the resulting C source in a Windows GUI. Context menus support navigation, properties, cross references, C/Asm mixed view and manipulation of the decompile context (function ABI).

http://www.c4decompiler.com

1.2.5 Disassembly of 8 bit CPU code

Most CPUs are 8-bit CPUs.[1]

Normally when a subroutine is finished, it returns to executing the next address immediately following the call instruction.

However, assembly-language programmers occasionally use several different techniques that adjust the return address, making disassembly more difficult:

- jump tables,
- calculated jumps, and
- a parameter after the call instruction.

jump tables and other calculated jumps

On 8-bit CPUs, calculated jumps are often implemented by pushing a calculated "return" address to the stack, then jumping to that address using the "return" instruction. For example, the RTS Trick uses this technique to implement jump tables (w:branch table).

parameters after the call instruction

Instead of picking up their parameters off the stack or out of some fixed global address, some subroutines provide parameters in the addresses of memory that follow the instruction that called that subroutine. Subroutines that use this technique adjust the return address to skip over all the constant parameter data, then return to an address many bytes after the "call" instruction. One of the more famous programs that used this technique is the "Sweet 16" virtual machine.

The technique may make disassembly more difficult.

A simple example of this is the write() procedure implemented as follows:

; assume ds = cs, e.g like in boot sector code start: call write ; push message's address on top of stack db "Hello, world",0dh,0ah,00h ; return point ret ; back to DOS write proc near pop si ; get string address mov ah,0eh ; BIOS: write teletype w_loop: lodsb ; read char at [ds:si] and increment si or al,al ; is it 00h? jz short w_exit int 10h ; write the character jmp w_loop ; continue writing w_exit: jmp si write endp end start

A macro-assembler like TASM will then use a macro like this one:

_write macro message call write db message db 0 _write endm

From a human disassembler's point of view, this is a nightmare, although this is straightforward to read in the

original Assembly source code, as there is no way to decide if the db should be interpreted or not from the binary form, and this may contain various jumps to real executable code area, triggering analysis of code that should never be analysed, and interfering with the analysis of the real code (e.g. disassembling the above code from 0000h or 0001h won't give the same results at all).

However a half-decent tool with possibilities to specifiy rules, and heuristic means to identify texts will have little trouble.

1.2.6 Disassembly of 32 bit CPU code

Most 32-bit CPUs use the ARM instruction set.[1][2][3]

Typical ARM assembly code is a series of subroutines, with literal constants scattered between subroutines. The standard prolog and epilog for subroutines is pretty easy to recognize.

1.2.7 A brief list of disassemblers

- ciasdis "an assembler where the elements opcode, operands and modifiers are all objects, that are reusable for disassembly." For 8080 8086 80386 Alpha 6809 and should be usable for Pentium 68000 6502 8051.
- radare, the reverse engineering framework includes open-source tools to disassemble code for many processors including x86, ARM, PowerPC, m68k, etc. several virtual machines including java, msil, etc., and for many platforms including Linux, BSD, OSX, Windows, iPhoneOS, etc.
- IDA, the Interactive Disassembler (IDA Pro) can disassemble code for a huge number of processors, including ARM Architecture (including Thumb and Thumb-2), ATMEL AVR, INTEL 8051, INTEL 80x86, MOS Technologies 6502, MC6809, MC6811, M68H12C, MSP430, PIC 12XX, PIC 14XX, PIC 18XX, PIC 16XXX, Zilog Z80, etc.
- Disassemblers at DMOZ lists a huge number of disassemblers
- Program transformation wiki: disassembly lists many highly recommended disassemblers
- Wikipedia: objdump, part of the GNU binutils, can disassemble code for several processors and platforms.
- search for "disassemble" at SourceForge shows many disassemblers for a variety of CPUs.
- Hopper is a disassembler that runs on OS-X and disassembles 32/64-bit OS-X and windows binaries.
- The University of Queensland Binary Translator (UQBT) is a reusable, component-based binary-translation framework that supports CISC, RISC, and stack-based processors.

1.2.8 Further reading

[1] Jim Turley. "The Two Percent Solution". 2002.

[2] Mark Hachman. "ARM Cores Climb Into 3G Territory". 2002. "Although Intel and AMD receive the bulk of attention in the computing world, ARM's embedded 32-bit architecture, ... has outsold all others."

[3] Tom Krazit. "ARMed for the living room". "ARM licensed 1.6 billion cores [in 2005]". 2006.

- http://www.crackmes.de/ : reverse engineering challenges
- "A Challengers Handbook" by Caesum has some tips on reverse engineering programs in JavaScript, Flash Actionscript (SWF), Java, etc.
- the Open Source Institute occasionally has reverse engineering challenges among its other brainteasers.
- The Program Transformation wiki has a Reverse engineering and Re-engineering Roadmap, and discusses disassemblers, decompilers, and tools for translating programs from one high-level language to another high-level language.
- Other disassemblers with multi-platform support

1.3 x86 Disassembly/Disassembly Examples

1.3.1 Example: Hello World Listing

Write a simple "Hello World" program using C or C++ and your favorite compiler. Generate a listing file from the compiler. Does the code look the way you expect it to? Do you understand what the assembly code means?

Here are examples of C and C++ "Hello World!" programs.

#include <stdio.h> int main() { printf("Hello World!\n"); return 0; }
#include <iostream> int main() { std::cout << "Hello World!\n"; return 0; }

1.3.2 Example: Basic Disassembly

Write a basic “Hello World!" program (see the example above). Compile the program into an executable with your favorite compiler, then disassemble it. How big is the disassembled code file? How does it compare to the code from the listing file you generated? Can you explain why the file is this size?

1.4 x86 Disassembly/Analysis Tools

1.4.1 Debuggers

Debuggers are programs that allow the user to execute a compiled program one step at a time. You can see what instructions are executed in which order, and which sections of the program are treated as code and which are treated as data. Debuggers allow you to analyze the program while it is running, to help you get a better picture of what it is doing.

Advanced debuggers often contain at least a rudimentary disassembler, often times hex editing and reassembly features. Debuggers often allow the user to set *breakpoints* on instructions, function calls, and even memory locations.

A breakpoint is an instruction to the debugger that allows program execution to be halted when a certain condition is met. For instance, when a program accesses a certain variable, or calls a certain API function, the debugger can pause program execution.

Windows Debuggers

SoftICE A *de facto* standard for Windows debugging. SoftICE can be used for *local kernel debugging*, which is a feature that is very rare, and very valuable. SoftICE was taken off the market in April 2006.

WinDbg WinDbg is a free piece of software from Microsoft that can be used for local user-mode debugging, or even remote kernel-mode debugging. WinDbg is not the same as the better-known Visual Studio Debugger, but comes with a nifty GUI nonetheless. Available in 32 and 64-bit versions.

http://www.microsoft.com/whdc/devtools/debugging/installx86.mspx

IDA Pro The multi-processor, multi-OS, interactive disassembler by DataRescue.

http://www.hex-rays.com/idapro/

OllyDbg OllyDbg is a free and powerful Windows debugger with a built-in disassembly and assembly engine. Very useful for patching, disassembling, and debugging.

http://www.ollydbg.de/

Immunity Debugger Immunity Debugger is a branch of OllyDbg v1.10, with built-in support for Python scripting and much more.

http://immunityinc.com/products-immdbg.shtml

Linux Debuggers

Many of the open source debuggers on Linux, again, are cross-platform. They may be available on some other Unix(-like) systems, or even Windows. Some of the debuggers may give you better experience than the old and native ones on your system.

gdb The GNU debugger, comes with any normal Linux install. It is quite powerful and even somewhat programmable, though the raw user interface is harsh.

lldb LLVM’s debugger.

emacs The GNU editor, can be used as a front-end to gdb. This provides a powerful hex editor and allows full scripting in a LISP-like language.

ddd The Data Display Debugger. It’s another front-end to gdb. This provides graphical representations of data structures. For example, a linked list will look just like a textbook illustration.

strace, ltrace, and xtrace Lets you run a program while watching the actions it performs. With strace, you get a log of all the system calls being made. With ltrace, you get a log of all the library calls being made. With xtrace, you get a log of some of the funtion calls being made.

valgrind Executes a program under emulation, performing analysis according to one of the many plug-in modules as desired. You can write your own plug-in module as desired. Newer versions of valgrind also support OS X.

NLKD A kernel debugger.

http://forge.novell.com/modules/xfmod/project/?nlkd

edb A fully featured plugin-based debugger inspired by the famous OllyDbg. Project page

KDbg A gdb front-end for KDE. http://kdbg.org

RR0D A Ring-0 Debugger for Linux. RR0D Project Page

Winedbg Wine's debugger. Debugs Windows executables using wine.

Debuggers for Other Systems

dbx The standard Unix debugger on systems derived from AT&T Unix. It is often part of an optional development toolkit package which comes at an extra price. It uses an interactive command line interface.

ladebug An enhanced debugger on Tru64 Unix systems from HP (originally Digital Equipment Corporation) that handles advanced functionality like threads better than dbx.

DTrace An advanced tool on Solaris that provides functions like profiling and many others on the entire system, including the kernel.

mdb The Modular Debugger (MDB) is a new general purpose debugging tool for the Solaris Operating Environment. Its primary feature is its extensibility. The Solaris Modular Debugger Guide describes how to use MDB to debug complex software systems, with a particular emphasis on the facilities available for debugging the Solaris kernel and associated device drivers and modules. It also includes a complete reference for and discussion of the MDB language syntax, debugger features, and MDB Module Programming API.

Debugger Techniques

Setting Breakpoints As previously mentioned in the section on disassemblers, a 6-line C program doing something as simple as outputting “Hello, World!" turns into massive amounts of assembly code. Most people don't want to sift through the entire mess to find out the information they want. It can be time consuming just to find the information one desires by just looking through the code. As an alternative, one can choose to set breakpoints to halt the program once it has reached a given point within the program’s code.

For instance, let’s say that in your program you consistantly experience crashes after one particular event: immediately after closing a message box. You set breakpoints on all calls to *MessageBoxA*. You run your program with the breakpoints set, and it stops, ready to call *MessageBoxA*. Executing each line one-by-one thereafter (referred to as *stepping*) through the code, and watching the program stack, you see that a buffer overflow occurs soon after the call.

1.4.2 Hex Editors

Hex editors are able to directly view and edit the binary of a source file, and are very useful for investigating the structure of proprietary closed-format data files. There are many hex editors in existence. This section will attempt to list some of the best, some of the most popular, or some of the most powerful.

Hexinator (For Windows and Linux) lets you edit files of unlimited size (overwrite, insert, delete), displays text with dozens of text encodings, shows variables in little and big endian byte order.

https://hexinator.com

wxHexEditor (For Windows and Linux, Free & Open Source) A fast hex editor specially for HUGE files and disk devices, allows up to hexabyte, allow size changes (inject and deletes) without creating temp file, could view files with multiple panes, has built-in disassembler, supports tags for (reverse) engineering big binaries or file systems, could view files thrug XOR encryption.

http://wxhexeditor.sourceforge.net/

HxD (Freeware) For Windows. A fast and powerful free hex, disk and RAM editor

http://mh-nexus.de/hxd/

Freeware Hex Editor XVI32 For Windows. A freeware hex editor.

http://www.chmaas.handshake.de/delphi/freeware/xvi32/xvi32.htm

HHD Software Hex Editor Neo For Windows. A fast file, disk, and memory editor with built-in disassembler and file structure viewer.

http://www.hhdsoftware.com/Family/hex-editor.html

Catch22 HexEdit For Windows. his is a powerful hex editor with a slew of features. Has an excellent data structure viewer.

http://www.catch22.net/software/hexedit.asp

BreakPoint Hex Workshop For Windows. An excellent and powerful hex-editor, its usefulness is restricted by the fact that it is not free like some of the other options.

http://www.bpsoft.com/

Tiny Hexer Free and does statistics. For Windows.

http://www.mirkes.de/files/

frhed - free hex editor For Windows. Free and open-source.

http://www.kibria.de/frhed.html

Cygnus Hex Editor For Windows. A very fast and easy-to-use hex editor, available in a 'Free Edition'.

http://www.softcircuits.com/cygnus/fe/

Hexprobe Hex Editor For Windows. A professional hex editor designed to include all the power to deal with hex data, particularly helpful in the areas of hex-byte editing and byte-pattern analysis.

http://www.hexprobe.com/hexprobe/index.htm

UltraEdit32 For Windows. A hex editor/text editor, won "Application of the Year" at 2005 Shareware Industry Awards Conference.

http://www.ultraedit.com/

ICY Hexplorer For Windows. A lightweight free and open source hex file editor with some nifty features, such as pixel view, structures, and disassembling.

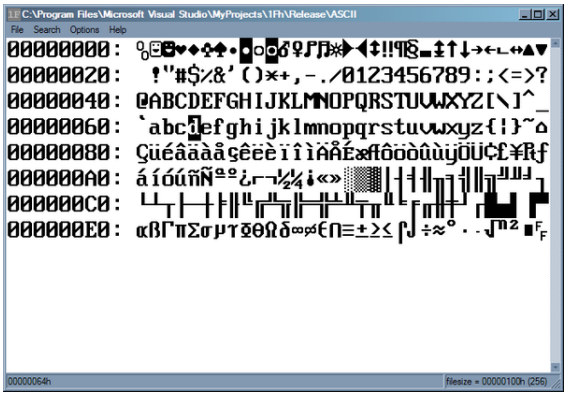
http://hexplorer.sourceforge.net/

WinHex For Windows. A powerful hex file and disk editor with advanced abilities for computer forensics and data recovery (used by governments and military).

http://www.x-ways.net/index-m.html

010 Editor For Windows. A very powerful and fast hex editor with extensive support for data structures and scripting. Can be used to edit drives and processes.

http://www.sweetscape.com/010editor/



A view of a small binary file in a 1Fh hex editor.

1Fh For Windows. A free binary/hex editor which is very fast, even while working with large files. It's the only Windows hex editor that allows you to view files in byte code (all 256-characters).

http://www.4neurons.com/1Fh/

HexEdit For Windows (Open source) and shareware versions. Powerful and easy to use binary file and disk editor.

http://www.hexedit.com/

HexToolkit For Windows. A free hex viewer specifically designed for reverse engineering file formats. Allows data to be viewed in various formats and includes an expression evaluator as well as a binary file comparison tool.

http://www.binaryearth.net/HexToolkit

FlexHex For Windows. It Provides full support for NTFS files which are based on a more complex model than FAT32 files. Specifically, FlexHex supports Sparse files and Alternate data streams of files on any NTFS volume. Can be used to edit OLE compound files, flash cards, and other types of physical drives.

http://www.heaventools.com/flexhex-hex-editor.htm

HT Editor For Windows. A file editor/viewer/analyzer for executables. Its goal is to combine the low-level functionality of a debugger and the usability of IDEs.

http://hte.sourceforge.net/

HexEdit For MacOS. A simple but reliable hex editor wher you to change highlight colours. There is also a port for Apple Classic users.

http://hexedit.sourceforge.net/

Hex Fiend For MacOS. A very simple hex editor, but incredibly powerful nonetheless. It's only 346 KB to download and takes files as big as 116 GB.

http://ridiculousfish.com/hexfiend/

Linux Hex Editors only

bvi A typical three-pane hex editor, with a vi-like interface.

emacs Along with everything else, emacs also includes a hex editor.

joe Joe's own editor now also supports hex editing.

bless A very capable gtk based hex editor.

xxd and any text editor Produce a hex dump with xxd, freely edit it in your favorite text editor, and then convert it back to a binary file with your changes included.

GHex Hex editor for GNOME.

http://directory.fsf.org/All_Packages_in_Directory/ghex.html

Okteta The well-integrated hexeditor from KDE since 4.1. Offers the traditional two-columns layout, one with numeric values (binary, octal, decicmal, hexdecimal) and one with characters (lots of charsets supported). Editing can be done in both columns, with unlimited undo/redo. Small set of tools (searching/replacing, strings, binary filter, and more).

http://utils.kde.org/projects/okteta

BEYE A viewer of binary files with built-in editor in binary, hexadecimal and disassembler modes. It uses native Intel syntax for disassembly. Highlight AVR/Java/Athlon64/Pentium 4/K7-Athlon disassembler, Russian codepages converter, full preview of formats - MZ, NE, PE, NLM, coff32, elf partial - a.out, LE, LX, PharLap; code navigator and more over. (

http://beye.sourceforge.net/en/beye.html

BIEW A viewer of binary files with built-in editor in binary, hexadecimal and disassembler modes. It uses native Intel syntax for disassembly. Highlight AVR/Java/Athlon64/Pentium 4/K7-Athlon disassembler, Russian codepages converter, full preview of formats - MZ, NE, PE, NLM, coff32, elf partial - a.out, LE, LX, PharLap; code navigator and more over. (PROJECT RENAMED, see BEYE)

http://biew.sourceforge.net/en/biew.html

hview A curses based hex editor designed to work with large (600+MB) files with as quickly, and with little overhead, as possible.

http://web.archive.org/web/20010306001713/http://tdistortion.esmartdesign.com/Zips/hview.tgz

HexCurse An ncurses-based hex editor written in C that currently supports hex and decimal address output, jumping to specified file locations, searching, ASCII and EBCDIC output, bolded modifications, an undo command, quick keyboard shortcuts, etc.

http://www.jewfish.net/description.php?title=HexCurse

hexedit View and edit files in hexadecimal or in ASCII.

http://rigaux.org/hexedit.html

Data Workshop An editor to view and modify binary data; provides different views which can be used to edit, analyze and export the binary data.

http://www.dataworkshop.de/

VCHE A hex editor which lets you see all 256 characters as found in video ROM, even control and extended ASCII, it uses the /dev/vcsa* devices to do it. It also could edit non-regular files, like hard disks, floppies, CDROMs, ZIPs, RAM, and almost any device. It comes with a ncurses and a raw version for people who work under X or remotely.

http://www.grigna.com/diego/linux/vche/

DHEX DHEX is just another Hexeditor with a Diff-mode for ncurses. It makes heavy use of colors and is themeable.

http://www.dettus.net/dhex/

1.4.3 Other Tools for Windows

Resource Monitors

SysInternals Freeware This page has a large number of excellent utilities, many of which are very useful to security experts, network administrators, and (most importantly to us) reversers. Specifically, check out **Process Monitor**, **FileMon**, **RegMon**, **TCPView**, and **Process Explorer**.

http://technet.microsoft.com/sysinternals/default.aspx

API Monitors

SpyStudio Freeware The Spy Studio software is a tool to hook into windows processes, log windows API call to DLLs, insert breakpoints and change parameters.

http://www.nektra.com/products/spystudio/

rohitab.com API Monitor API Monitor is a free software that lets you monitor and control API calls made by applications and services. Features include detailed parameter information, structures, unions, enumerated/flag data types, call stack, call tree, breakpoints, custom DLL's, memory editor, call filtering, COM monitoring, 64-bit. Includes definitions for over 13,000 API's and 1,300+ COM interfaces.

http://www.rohitab.com/apimonitor

PE File Header dumpers

Dumpbin Dumpbin is a program that previously used to be shipped with MS Visual Studio, but recently the functionality of Dumpbin has been incorporated into the Microsoft Linker, link.exe. to access dumpbin, pass /dump as the first parameter to link.exe:

link.exe /dump [options]

It is frequently useful to simply create a batch file that handles this conversion:

::dumpbin.bat link.exe /dump %*

All examples in this wikibook that use dumpbin will call it in this manner.

Here is a list of useful features of dumpbin :

dumpbin /EXPORTS displays a list of functions exported from a library dumpbin /IMPORTS displays a list of functions imported from other libraries dumpbin /HEADERS displays PE header information for the executable

http://msdn.microsoft.com/library/default.asp?url=/library/en-us/vccore/html/_core_dumpbin_reference.asp

Depends Dependency Walker is a GUI tool which will allow you to see exports and imports of binaries. It ships with many Microsoft tools including MS Visual Studio.

1.4.4 GNU Tools

The GNU packages have been ported to many platforms including Windows.

GNU BinUtils The GNU BinUtils package contains several small utilities that are very useful in dealing with binary files. The most important programs in the list are the GNU objdump, readelf, GAS assembler, and the GNU linker, although the reverser might find more use in addr2line, c++filt, nm, and readelf.

http://www.gnu.org/software/binutils/

objdump Dumps out information about an executable including symbols and assembly. It comes standard. It can be made to support non-native binary formats.

objdump -p displays a list of functions imported from other libraries, exported to and miscellaneous file header information

It's useful to check dll dependencies from command line

readelf Like *objdump* but more specialized for ELF executables.

size Lists the sizes of the segments.

nm Lists the symbols in an ELF file.

strings Prints the strings from a file.

file Tells you what type of file it is.

fold Folds the results of *strings* into something pageable.

kill Can be used to halt a program with the sig_stop signal.

strace Trace system calls and signals.

1.4.5 Other Tools for Linux

oprofile Can be used the find out what functions and data segments are used

subterfugue A tool for playing odd tricks on an executable as it runs. The tool is scriptable in python. The user can write scripts to take action on events that occur, such as changing the arguments to system calls.

http://subterfugue.org/

lizard Lets you run a program *backwards*.

http://lizard.sourceforge.net/

dprobes Lets you work with both kernel and user code.

biew Both a hex editor and a disassembler.

ltrace Displays runtime library call information for dynamically linked executables.

asmDIFF Searches for functions, instructions and memory pointers in different versions of same binary by using code metrics. Supports x86, x86_64 code in PE and ELF files.

http://duschkumpane.org/index.php/asmdiff

1.4.6 XCode Tools

XCode contains some extra tools to be used under OS X with the Mach-O format. You can see more of them under /Applications/Xcode.app/Contents/Developer/usr/bin/.

lipo Manages fat binaries with multiple architectures.

otool *Object file displaying tool*, works somehow like objdump and readelf.

XCode also packs a lot of Unix tools, with many of them sharing the names (and functions) of the GNU tools. Other tools like nasm/ndisasm, lldb and GNU as can also be found.

Chapter 2

Platforms

2.1 x86 Disassembly/Microsoft Windows

2.1.1 Microsoft Windows

The **Windows operating system** is a popular reverse engineering target for one simple reason: the OS itself (market share, known weaknesses), and most applications for it, are not Open Source or free. Most software on a Windows machine doesn't come bundled with its source code, and most pieces have inadequate, or non-existent documentation. Occasionally, the only way to know precisely what a piece of software does (or for that matter, to determine whether a given piece of software is malicious or legitimate) is to reverse it, and examine the results.

2.1.2 Windows Versions

Windows operating systems can be easily divided into 2 categories: Win9x, and WinNT.

Windows 9x

The Win9x kernel was originally written to span the 16bit - 32bit divide. Operating Systems based on the 9x kernel are: Windows 95, Windows 98, and Windows ME. Win9x Series operating systems are known to be prone to bugs and system instability. The actual OS itself was a 32 bit extension of MS-DOS, its predecessor. An important issue with the 9x line is that they were all based around using the ASCII format for storing strings, rather than Unicode.

Development on the Win9x kernel ended with the release of Windows ME.

Windows NT

The WinNT kernel series was originally written as enterprise-level server and network software. WinNT stresses stability and security far more than Win9x kernels did (although it can be debated whether that stress was good enough). It also handles all string operations internally in Unicode, giving more flexibility when using different languages. Operating Systems based on the WinNT kernel are: Windows NT (versions 3.1, 3.5, 3.51 and 4.0), Windows 2000 (NT 5.0), Windows XP (NT 5.1), Windows Server 2003 (NT 5.2), Windows Vista (NT 6.0), and Windows 7 (NT 6.1).

The Microsoft XBOX and and XBOX 360 also run a variant of NT, forked from Windows 2000. Most future Microsoft operating system products are based on NT in some shape or form.

2.1.3 Virtual Memory

32 bit WinNT allows for a maximum of 4Gb of virtual memory space, divided into “pages” that are 4096 bytes by default. Pages not in current use by the system or any of the applications may be written to a special section on the harddisk known as the “paging file.” Use of the paging file may increase performance on some systems, although high latency for I/O to the HDD can actually reduce performance in some instances.

2.1.4 System Architecture

The Windows architecture is heavily layered. Function calls that a programmer makes may be redirected 3 times or more before any action is actually performed. There is an unignorable penalty for calling Win32 functions from a user-mode application. However, the upside is equally unignorable: code written in higher levels of the windows system is much easier to write. Complex operations that involve initializing multiple data structures and calling multiple sub-functions can be performed by calling only a single higher-level function.

The Win32 API comprises 3 modules: KERNEL, USER, and GDI. KERNEL is layered on top of NTDLL, and most calls to KERNEL functions are simply redirected into NTDLL function calls. USER and GDI are both based on WIN32K (a kernel-mode module, responsible for the Windows “look and feel”), although USER also

makes many calls to the more-primitive functions in GDI. This and NTDLL both provide an interface to the Windows NT kernel, NTOSKRNL (see further below).

NTOSKRNL is also partially layered on HAL (Hardware Abstraction Layer), but this interaction will not be considered much in this book. The purpose of this layering is to allow processor variant issues (such as location of resources) to be made separate from the actual kernel itself. A slightly different system configuration thus requires just a different HAL module, rather than a completely different kernel module.

2.1.5 System calls and interrupts

After filtering through different layers of subroutines, most API calls require interaction with part of the operating system. Services are provided via 'software interrupts', traditionally through the "int 0x2e" instruction. This switches control of execution to the NT executive / kernel, where the request is handled. It should be pointed out here that the stack used in kernel mode is different from the user mode stack. This provides an added layer of protection between kernel and user. Once the function completes, control is returned back to the user application.

Both Intel and AMD provide an extra set of instructions to allow faster system calls, the "SYSENTER" instruction from Intel and the SYSCALL instruction from AMD.

2.1.6 Win32 API

Both WinNT and Win9x systems utilize the Win32 API. However, the WinNT version of the API has more functionality and security constructs, as well as Unicode support. Most of the Win32 API can be broken down into 3 separate components, each performing a separate task.

kernel32.dll

Kernel32.dll, home of the KERNEL subsystem, is where non-graphical functions are implemented. Some of the APIs located in KERNEL are: The Heap API, the Virtual Memory API, File I/O API, the Thread API, the System Object Manager, and other similar system services. Most of the functionality of kernel32.dll is implemented in ntdll.dll, but in undocumented functions. Microsoft prefers to publish documentation for kernel32 and guarantee that these APIs will remain unchanged, and then put most of the work in other libraries, which are then not documented.

gdi32.dll

gdi32.dll is the library that implements the GDI subsystem, where primitive graphical operations are performed. GDI diverts most of its calls into WIN32K, but it does contain a manager for GDI objects, such as pens, brushes and device contexts. The GDI object manager and the KERNEL object manager are completely separate.

user32.dll

The USER subsystem is located in the user32.dll library file. This subsystem controls the creation and manipulation of USER objects, which are common screen items such as windows, menus, cursors, etc... USER will set up the objects to be drawn, but will perform the actual drawing by calling on GDI (which in turn will make many calls to WIN32K) or sometimes even calling WIN32K directly. USER utilizes the GDI Object Manager.

2.1.7 Native API

The native API, hereby referred to as the NTDLL subsystem, is a series of undocumented API function calls that handle most of the work performed by KERNEL. Microsoft also does not guarantee that the native API will remain the same between different versions, as Windows developers modify the software. This gives the risk of native API calls being removed or changed without warning, breaking software that utilizes it.

ntdll.dll

The NTDLL subsystem is located in ntdll.dll. This library contains many API function calls, that all follow a particular naming scheme. Each function has a prefix: Ldr, Nt, Zw, Csr, Dbg, etc... and all the functions that have a particular prefix all follow particular rules.

The "official" native API is usually limited only to functions whose prefix is Nt or Zw. These calls are in fact the same in user-mode: the relevant Export entries map to the same address in memory. However, in kernel-mode, the Zw* system call stubs set the *previous mode* to kernel-mode, ensuring that certain parameter validation routines are *not* performed. The origin of the prefix "Zw" is unknown; it is rumored that this prefix was chosen due to its having no significance at all.

In actual implementation, the system call stubs merely load two registers with values required to describe a native API call, and then execute a software interrupt (or the sysenter instruction).

Most of the other prefixes are obscure, but the known ones are:

- Rtl stands for "Run Time Library", calls which help functionality at runtime (such as RtlAllocateHeap)
- Csr is for "Client Server Runtime", which represents the interface to the win32 subsystem located in csrss.exe

- Dbg functions are present to enable debugging routines and operations
- Ldr provides the ability to load, manipulate and retrieve data from DLLs and other module resources

User Mode Versus Kernel Mode

Many functions, especially Run-time Library routines, are shared between ntdll.dll and ntoskrnl.exe. Most Native API functions, as well as other kernel-mode only functions exported from the kernel are useful for driver writers. As such, Microsoft provides documentation on many of the native API functions with the Microsoft Server 2003 Platform DDK. The DDK (Driver Development Kit) is available as a free download.

2.1.8 ntoskrnl.exe

This module is the Windows NT "'Executive'", providing all the functionality required by the native API, as well as the kernel itself, which is responsible for maintaining the machine state. By default, all interrupts and kernel calls are channeled through ntoskrnl in some way, making it the single most important program in Windows itself. Many of its functions are exported (all of which with various prefixes, a la NTDLL) for use by device drivers.

2.1.9 Win32K.sys

This module is the "Win32 Kernel" that sits on top of the lower-level, more primitive NTOSKRNL. WIN32K is responsible for the "look and feel" of windows, and many portions of this code have remained largely unchanged since the Win9x versions. This module provides many of the specific instructions that cause USER and GDI to act the way they do. It's responsible for translating the API calls from the USER and GDI libraries into the pictures you see on the monitor.

2.1.10 Win64 API

With the advent of 64-bit processors, 64-bit software is a necessity. As a result, the Win64 API was created to utilize the new hardware. It is important to note that the format of many of the function calls are identical in Win32 and Win64, except for the size of pointers, and other data types that are specific to 64-bit address space.

Differences

2.1.11 Windows Vista

Microsoft has released a new version of its Windows operation system, named "Windows Vista." Windows Vista may be better known by its development code-name "Longhorn." Microsoft claims that Vista has been written largely from the ground up, and therefore it can be assumed that there are fundamental differences between the Vista API and system architecture, and the APIs and architectures of previous Windows versions. Windows Vista was released January 30th, 2007.

2.1.12 Windows CE/Mobile, and other versions

Windows CE is the Microsoft offering on small devices. It largely uses the same Win32 API as the desktop systems, although it has a slightly different architecture. Some examples in this book may consider WinCE.

2.1.13 "Non-Executable Memory"

Recent windows service packs have attempted to implement a system known as "Non-executable memory" where certain pages can be marked as being "non-executable". The purpose of this system is to prevent some of the most common security holes by not allowing control to pass to code inserted into a memory buffer by an attacker. For instance, a shellcode loaded into an overflowed text buffer cannot be executed, stopping the attack in its tracks. The effectiveness of this mechanism is yet to be seen, however.

2.1.14 COM and Related Technologies

COM, and a whole slew of technologies that are either related to COM or are actually COM with a fancy name, is another factor to consider when reversing Windows binaries. COM, DCOM, COM+, ActiveX, OLE, MTS, and Windows DNA are all names for the same subject, or subjects, so similar that they may all be considered under the same heading. In short, COM is a method to export Object-Oriented Classes in a uniform, *cross-platform* and *cross-language* manner. In essence, COM is .NET, version 0 beta. Using COM, components written in many languages can export, import, instantiate, modify, and destroy objects defined in another file, most often a DLL. Although COM provides cross-platform (to some extent) and cross-language facilities, each COM object is compiled to a native binary, rather than an intermediate format such as Java or .NET. As a result, COM does not require a virtual machine to execute such objects.

This book will attempt to show some examples of COM files, and the reversing challenges associated with them, although the subject is very broad, and may elude the scope of this book (or at least the early sections of it). The discussion may be part of an "Advanced Topic" found in the later sections of this book.

Due to the way that COM works, a lot of the methods and

data structures exported by a COM component are difficult to perceive by simply inspecting the executable file. Matters are made worse if the creating programmer has used a library such as ATL to simplify their programming experience. Unfortunately for a reverse engineer, this reduces the contents of an executable into a "Sea of bits", with pointers and data structures everywhere.

2.1.15 Remote Procedure Calls (RPC)

RPC is a generic term referring to techniques that allow a program running on one machine to make calls that actually execute on another machine. Typically, this is done by *marshalling* all of the data needed for the procedure including any state information stored on the first machine, and building it into a single data structure, which is then transmitted over some communications method to a second machine. This second machine then performs the requested action, and returns a data packet containing any results and potentially changed state information to the originating machine.

In Windows NT, RPC is typically handled by having two libraries that are similarly named, one which generates RPC requests and accepts RPC returns, as requested by a user-mode program, and one which responds to RPC requests and returns results via RPC. A classic example is the print spooler, which consists of two pieces: the RPC stub spoolss.dll, and the spooler proper and RPC service provider, spoolsv.exe. In most machines, which are stand-alone, it would seem that the use of two modules communicating by means of RPC is overkill; why not simply roll them into a single routine? In networked printing, though, this makes sense, as the RPC service provider can be resident physically on a distant machine, with the remote printer, and the local machine can control the printer on the remote machine in exactly the same way that it controls printers on the local machine.

2.2 x86 Disassembly/Windows Executable Files

2.2.1 MS-DOS COM Files

COM files are loaded into RAM exactly as they appear; no change is made at all from the harddisk image to RAM. This is possible due to the 20-bit memory model of the early x86 line. Two 16-bit registers would have to be set, one dividing the 1MB+64K memory space into 65536 'segments' and one specifying an offset from that. The segment register would be set by DOS and the COM file would be expected to respect this setting and not ever change the segment registers. The offset registers, however, were free game and served (for COM files) the same purpose as a modern 32-bit register. The downside was that the offset registers were only 16-bit and, therefore, since COM files could not change the segment registers, COM files were limited to using 64K of RAM. The good thing about this approach, however, was that no extra work was needed by DOS to load and run a COM file: just load the file, set the segment register, and jump to it. (The programs could perform 'near' jumps by just giving an offset to jump to.)

COM files are loaded into RAM at offset $100. The space before that would be used for passing data to and from DOS (for example, the contents of the command line used to invoke the program).

Note that COM files, by definition, cannot be 32-bit. Windows provides support for COM files via a special CPU mode.

2.2.2 MS-DOS EXE Files

One way MS-DOS compilers got around the 64k memory limitation was with the introduction of **memory models**. The basic concept is to cleverly set different segment registers in the x86 CPU (CS, DS, ES, SS) to point to the same or different segments, thus allowing varying degrees of access to memory. Typical memory models were:

tiny All memory access are 16-bit (never reload any segment register). Produces a .COM file instead of an .EXE file.

small All memory access are 16-bit (never reload any segment register).

compact accesses to the data segment reload the DS, ES, SS register, allowing 32-bit of data. Code accesses don't reload the CS register, allowing 16-bit of code.

medium accesses to the code segment reload the CS register, allowing 32-bit of code. Data accesses don't reload the DS, ES, SS registers, allowing 16-bit of data.

large both code and data accesses use the segment registers (CS for code, DS, ES, SS for data), allowing 32-bit of code and 32-bit of data.

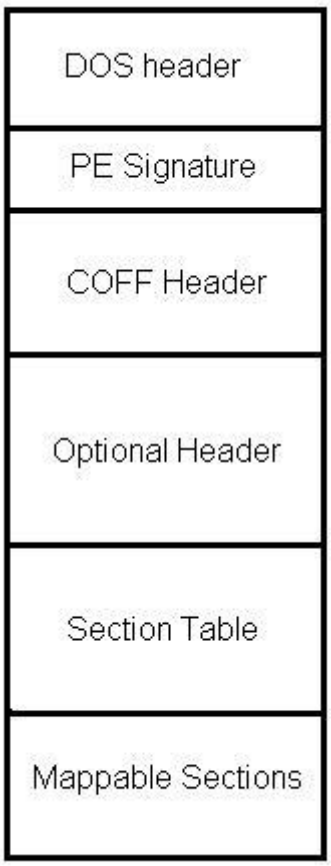
huge same as the large model, with additional arithmetic being generated by the compiler to allow access to arrays larger than 64k.

When looking at a COM file, one has to decide which memory model was used to build that file.

2.2.3 PE Files

A **Portable Executable** (**PE**) file is the standard binary file format for an Executable or DLL under Windows NT,

Windows 95, and Win32. The Win32 SDK contains a file, *winnt.h*, which declares various structs and variables used in the PE files. Some functions for manipulating PE files are also included in *imagehlp.dll*. PE files are broken down into various sections which can be examined.



The basic format of a Microsoft PE file

2.2.4 Relative Virtual Addressing (RVA)

In a Windows environment, executable modules can be loaded at any point in memory, and are expected to run without problem. To allow multiple programs to be loaded at seemingly random locations in memory, PE files have adopted a tool called RVA: Relative Virtual Addresses. RVA's assume that the "base address" of where a module is loaded into memory is not known at compile time. So, PE files describe the location of data in memory as an *offset from the base address*, wherever that may be in memory.

Some processor instructions require the code itself to directly identify where in memory some data is. This is not possible when the location of the module in memory is not known at compile time. The solution to this problem is described in the section on "Relocations".

It is important to remember that the addresses obtained from a disassembly of a module will not always match up to the addresses seen in a debugger as the program is running.

2.2.5 File Format

The PE portable executable file format includes a number of informational headers, and is arranged in the following format:

MS-DOS header

Open any Win32 binary executable in a hex editor, and note what you see: The first 2 letters are **always** the letters "MZ". To some people, the first few bytes in a file that determine the type of file are called the "magic number," although this book will not use that term, because there is no rule that states that the "magic number" needs to be a single number. Instead, we will use the term "File ID Tag", or simply, File ID. Sometimes this is also known as File Signature.

After the File ID, the hex editor will show several bytes of either random-looking symbols, or whitespace, before the human-readable string "This program cannot be run in DOS mode".

What is this?

```
00000000  4D 5A 90 00 03 00 00 00 04 00 00 00 FF
00000010  B8 00 00 00 00 00 00 00 40 00 00 00 00
00000020  00 00 00 00 00 00 00 00 00 00 00 00 00
00000030  00 00 00 00 00 00 00 00 00 00 00 00 D8
00000040  0E 1F BA 0E 00 B4 09 CD 21 B8 01 4C CD
00000050  69 73 20 70 72 6F 67 72 61 6D 20 63 61
00000060  74 20 62 65 20 72 75 6E 20 69 6E 20 44
00000070  6D 6F 64 65 2E 0D 0D 0A 24 00 00 00 00
00000080  26 C3 8E 85 62 A2 E0 D6 62 A2 E0 D6 62
00000090  E1 BE EE D6 6F A2 E0 D6 62 A2 E0 D6 6D
```

Hex Listing of an MS-DOS file header

What you are looking at is the MS-DOS header of the Win32 PE file. To ensure either a) backwards compatibility, or b) graceful decline of new file types, Microsoft has engineered a series of DOS instructions into the head of each PE file. When a 32-bit Windows file is run in a 16-bit DOS environment, the program will terminate immediately with the error message: "This program cannot be run in DOS mode".

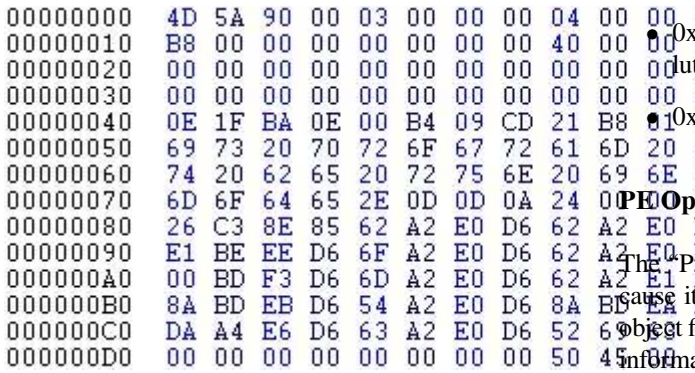
The DOS header is also known by some as the EXE header. Here is the DOS header presented as a C data structure:

struct DOS_Header { char signature[2] = "MZ"; short lastsize; short nblocks; short nreloc; short hdrsize; short minalloc; short maxalloc; void *ss; void *sp; short checksum; void *ip; void *cs; short relocpos; short noverlay; short reserved1[4]; short oem_id; short oem_info; short reserved2[10]; long e_lfanew; }

Immediately following the DOS Header will be the classic error message "This program cannot be run in DOS mode".

PE Header

At offset 60 (0x3C) from the beginning of the DOS header is a pointer to the Portable Executable (PE) File header (e_lfanew in MZ structure). DOS will print the error message and terminate, but Windows will follow this pointer to the next batch of information.



Hex Listing of a PE signature, and the pointer to it

The PE header consists only of a File ID signature, with the value "PE\0\0" where each '\0' character is an ASCII NUL character. This signature shows that a) this file is a legitimate PE file, and b) the byte order of the file. Byte order will not be considered in this chapter, and all PE files are assumed to be in "little endian" format.

The first big chunk of information lies in the COFF header, directly after the PE signature.

COFF Header

The COFF header is present in both COFF object files (before they are linked) and in PE files where it is known as the "File header". The COFF header has some information that is useful to an executable, and some information that is more useful to an object file.

Here is the COFF header, presented as a C data structure:

struct COFFHeader { short Machine; short NumberOfSections; long TimeDateStamp; long PointerToSymbolTable; long NumberOfSymbols; short SizeOfOptionalHeader; short Characteristics; }

Machine This field determines what machine the file was compiled for. A hex value of 0x14C (332 in decimal) is the code for an Intel 80386.

Here's a list of possible values it can have.

NumberOfSections The number of sections that are described at the end of the PE headers.

TimeDateStamp 32 bit time at which this header was generated: is used in the process of "Binding", see below.

SizeOfOptionalHeader this field shows how long the "PE Optional Header" is that follows the COFF header.

Characteristics This is a field of bit flags, that show some characteristics of the file.

- 0x02 = Executable file
- 0x200 = file is non-relocatable (addresses are absolute, not RVA).
- 0x2000 = File is a DLL Library, not an EXE.

PE Optional Header

The "PE Optional Header" is not "optional" per se, because it is required in Executable files, but not in COFF object files. The Optional header includes lots and lots of information that can be used to pick apart the file structure, and obtain some useful information about it.

The PE Optional Header occurs directly after the COFF header, and some sources even show the two headers as being part of the same structure. This wikibook separates them out for convenience.

Here is the PE Optional Header presented as a C data structure:

struct PEOptHeader { short signature; //decimal number 267 for 32 bit, and 523 for 64 bit. char MajorLinkerVersion; char MinorLinkerVersion; long SizeOfCode; long

SizeOfInitializedData; long SizeOfUninitializedData; long AddressOfEntryPoint; //The RVA of the code entry point long BaseOfCode; long BaseOfData; long ImageBase; long SectionAlignment; long FileAlignment; short MajorOSVersion; short MinorOSVersion; short MajorImageVersion; short MinorImageVersion; short MajorSubsystemVersion; short MinorSubsystemVersion; long Reserved; long SizeOfImage; long SizeOfHeaders; long Checksum; short Subsystem; short DLLCharacteristics; long SizeOfStackReserve; long SizeOfStackCommit; long SizeOfHeapReserve; long SizeOfHeapCommit; long LoaderFlags; long NumberOfRvaAndSizes; data_directory DataDirectory[16]; //Can have any number of elements, matching the number in NumberOfRvaAndSizes. } //However, it is always 16 in PE files.
struct data_directory { long VirtualAddress; long Size; }

Some of the important pieces of information:

MajorLinkerVersion, MinorLinkerVersion The version, in x.y format of the linker used to create the PE.

AddressOfEntryPoint The RVA of the entry point to the executable. Very important to know.

SizeOfCode Size of the .text (.code) section

SizeOfInitializedData Size of .data section

BaseOfCode RVA of the .text section

BaseOfData RVA of .data section

ImageBase Preferred location in memory for the module to be based at

Checksum Checksum of the file, only used to verify validity of modules being loaded into kernel space. The formula used to calculate PE file checksums is proprietary, although Microsoft provides API calls that can calculate the checksum for you.

Subsystem the Windows subsystem that will be invoked to run the executable

- 1 = native
- 2 = Windows/GUI
- 3 = Windows non-GUI
- 5 = OS/2
- 7 = POSIX

DataDirectory Possibly the most interesting member of this structure. Provides RVAs and sizes which locate various data structures, which are used for setting up the execution environment of a module. The details of what these structures do exist in other sections of this page, but the most interesting entries in DataDirectory are below:

- IMAGE_DIRECTORY_ENTRY_EXPORT (0) : Location of the export directory
- IMAGE_DIRECTORY_ENTRY_IMPORT (1) : Location of the import directory
- IMAGE_DIRECTORY_ENTRY_RESOURCE (2) : Location of the resource directory
- IMAGE_DIRECTORY_ENTRY_BOUND_IMPORT (11) : Location of alternate import-binding directory

2.2.6 Code Sections

The PE Header defines the number of sections in the executable file. Each section definition is 40 bytes in length. Below is an example hex from a program I am writing:

2E746578_74000000_00100000_00100000_A8050000
.text 00040000_00000000_00000000_00000000_20000000
2E646174_61000000_00100000_00200000_86050000
.data 000A0000_00000000_00000000_00000000_40000000
2E627373_00000000_00200000_00300000_00000000
.bss 00000000_00000000_00000000_00000000_80000000

The structure of the section descriptor is as follows:

Offset Length Purpose ------ ------- ------------------------ ---------- 0x00 8 bytes Section Name - in the above example the names are .text .data .bss 0x08 4 bytes Size of the section once it is loaded to memory 0x0C 4 bytes RVA (location) of section once it is loaded to memory 0x10 4 bytes Physical size of section on disk 0x14 4 bytes Physical location of section on disk (from start of disk image) 0x18 12 bytes Reserved (usually zero) (used in object formats) 0x24 4 bytes Section flags

A PE loader will place the sections of the executable image at the locations specified by these section descriptors (relative to the base address) and usually the alignment is 0x1000, which matches the size of pages on the x86.

Common sections are:

1. .text/.code/CODE/TEXT - Contains executable code (machine instructions)
2. .testbss/TEXTBSS - Present if Incremental Linking is enabled
3. .data/.idata/DATA/IDATA - Contains initialised data
4. .bss/BSS - Contains uninitialised data

Section Flags

The section flags is a 32-bit bit field (each bit in the value represents a certain thing). Here are the constants defined in the WINNT.H file for the meaning of the flags:

#define IMAGE_SCN_TYPE_NO_PAD 0x00000008 // Reserved. #define IMAGE_SCN_CNT_CODE 0x00000020 // Section contains code. #define IMAGE_SCN_CNT_INITIALIZED_DATA 0x00000040 // Section contains initialized data. #define IMAGE_SCN_CNT_UNINITIALIZED_DATA 0x00000080 // Section contains uninitialized data. #define IMAGE_SCN_LNK_OTHER 0x00000100 // Reserved. #define IMAGE_SCN_LNK_INFO 0x00000200 // Section contains comments or some other type of information. #define IMAGE_SCN_LNK_REMOVE 0x00000800 // Section contents will not become part of image. #define IMAGE_SCN_LNK_COMDAT 0x00001000 // Section contents comdat. #define IMAGE_SCN_NO_DEFER_SPEC_EXC 0x00004000 // Reset speculative exceptions handling bits in the TLB entries for this section. #define IMAGE_SCN_GPREL 0x00008000 // Section content can be accessed relative to GP #define IMAGE_SCN_MEM_FARDATA 0x00008000 #define IMAGE_SCN_MEM_PURGEABLE 0x00020000 #define IMAGE_SCN_MEM_16BIT 0x00020000 #define IMAGE_SCN_MEM_LOCKED 0x00040000 #define IMAGE_SCN_MEM_PRELOAD 0x00080000 #define IMAGE_SCN_ALIGN_1BYTES 0x00100000 // #define IMAGE_SCN_ALIGN_2BYTES 0x00200000 // #define IMAGE_SCN_ALIGN_4BYTES 0x00300000 // #define IMAGE_SCN_ALIGN_8BYTES 0x00400000 // #define IMAGE_SCN_ALIGN_16BYTES 0x00500000 // Default alignment if no others are specified. #define IMAGE_SCN_ALIGN_32BYTES 0x00600000 // #define IMAGE_SCN_ALIGN_64BYTES 0x00700000 // #define IMAGE_SCN_ALIGN_128BYTES 0x00800000 // #define IMAGE_SCN_ALIGN_256BYTES 0x00900000 // #define IMAGE_SCN_ALIGN_512BYTES 0x00A00000 // #define IMAGE_SCN_ALIGN_1024BYTES 0x00B00000 // #define IMAGE_SCN_ALIGN_2048BYTES 0x00C00000 // #define IMAGE_SCN_ALIGN_4096BYTES 0x00D00000 // #define IMAGE_SCN_ALIGN_8192BYTES 0x00E00000 // #define IMAGE_SCN_ALIGN_MASK 0x00F00000 #define IMAGE_SCN_LNK_NRELOC_OVFL 0x01000000 // Section contains extended relocations. #define IMAGE_SCN_MEM_DISCARDABLE 0x02000000 // Section can be discarded. #define IMAGE_SCN_MEM_NOT_CACHED 0x04000000 // Section is not cachable. #define IMAGE_SCN_MEM_NOT_PAGED 0x08000000 // Section is not pageable. #define IMAGE_SCN_MEM_SHARED 0x10000000 // Section is shareable. #define IMAGE_SCN_MEM_EXECUTE 0x20000000 // Section is executable. #define IMAGE_SCN_MEM_READ 0x40000000 // Section is readable. #define IMAGE_SCN_MEM_WRITE 0x80000000 // Section is writeable.

2.2.7 Imports and Exports - Linking to other modules

What is linking?

Whenever a developer writes a program, there are a number of subroutines and functions which are expected to be implemented already, saving the writer the hassle of having to write out more code or work with complex data structures. Instead, the coder need only declare one call to the subroutine, and the linker will decide what happens next.

There are two types of linking that can be used: static and dynamic. Static uses a library of precompiled functions. This precompiled code can be inserted into the final executable to implement a function, saving the programmer a lot of time. In contrast, dynamic linking allows subroutine code to reside in a different file (or *module*), which is loaded at runtime by the operating system. This is also known as a "Dynamically linked library", or DLL. A *library* is a module containing a series of functions or values that can be *exported*. This is different from the term *executable*, which *imports* things from libraries to do what it wants. From here on, "module" means any file of PE format, and a "Library" is any module which exports and imports functions and values.

Dynamically linking has the following benefits:

- It saves disk space, if more than one executable links to the library module
- Allows instant updating of routines, without providing new executables for all applications
- Can save space in memory by mapping the code of a library into more than one process
- Increases abstraction of implementation. The method by which an action is achieved can be modified without the need for reprogramming of applications. This is extremely useful for backward compatibility with operating systems.

This section discusses how this is achieved using the PE file format. An important point to note at this point is that *anything* can be imported or exported between modules, including variables as well as subroutines.

Loading

The downside of dynamically linking modules together is that, at runtime, the software which is initialising an executable must link these modules together. For various reasons, you cannot declare that "The function in this dynamic library will always exist in memory *here*". If that memory address is unavailable or the library is updated, the function will no longer exist there, and the application trying to use it will break. Instead, each module (library or executable) must declare what functions or values it *exports* to other modules, and also what it wishes to *import* from other modules.

As said above, a module cannot declare where in memory it expects a function or value to be. Instead, it declares where *in its own memory* it expects to find a **pointer** to the value it wishes to import. This permits the module to address any imported value, wherever it turns up in memory.

2.2.8 Exports

Exports are functions and values in one module that have been declared to be shared with other modules. This is done through the use of the "Export Directory", which is used to translate between the name of an export (or "Ordinal", see below), and a location in memory where the code or data can be found. The start of the export directory is identified by the IMAGE_DIRECTORY_ENTRY_EXPORT entry of the resource directory. All export data must exist in the same section. The directory is headed by the following structure:

struct IMAGE_EXPORT_DIRECTORY { long Characteristics; long TimeDateStamp; short MajorVersion; short MinorVersion; long Name; long Base; long NumberOfFunctions; long NumberOfNames; long *AddressOfFunctions; long *AddressOfNames; long *AddressOfNameOrdinals; }

The "Characteristics" value is generally unused, TimeDateStamp describes the time the export directory was generated, MajorVersion and MinorVersion should describe the version details of the directory, but their nature is undefined. These values have little or no impact on the actual exports themselves. The "Name" value is an RVA to a zero terminated ASCII string, the name of this library name, or module.

Names and Ordinals Each exported value has both a name and an "ordinal" (a kind of index). The actual exports themselves are described through AddressOfFunctions, which is an RVA to an array of RVA's, each pointing to a different function or value to be exported. The size of this array is in the value NumberOfFunctions. Each of these functions has an ordinal. The "Base" value is used as the ordinal of the first export, and the next RVA in the array is Base+1, and so forth.

Each entry in the AddressOfFunctions array is identified by a name, found through the RVA AddressOfNames. The data where AddressOfNames points to is an array of RVA's, of the size NumberOfNames. Each RVA points to a zero terminated ASCII string, each being the name of an export. There is also a second array, pointed to by the RVA in AddressOfNameOrdinals. This is also of size NumberOfNames, but each value is a 16 bit word, each value being an ordinal. These two arrays are parallel and are used to get an export value from AddressOfFunctions. To find an export by name, search the AddressOfNames array for the correct string and then take the corresponding ordinal from the AddressOfNameOrdinals array. This ordinal is then used to get an index to a value in AddressOfFunctions.

Forwarding As well as being able to export functions and values in a module, the export directory can *forward* an export to another library. This allows more flexibility when re-organising libraries: perhaps some functionality has branched into another module. If so, an export can be forwarded to that library, instead of messy reorganising inside the original module.

Forwarding is achieved by making an RVA in the AddressOfFunctions array point into the section which contains the export directory, something that normal exports should not do. At that location, there should be a zero terminated ASCII string of format "LibraryName.ExportName" for the appropriate place to forward this export to.

2.2.9 Imports

The other half of dynamic linking is importing functions and values into an executable or other module. Before runtime, compilers and linkers do not know where in memory a value that needs to be imported could exist. The import table solves this by creating an array of pointers at runtime, each one pointing to the memory location of an imported value. This array of pointers exists inside of the module at a defined RVA location. In this way, the linker can use addresses inside of the module to access values outside of it.

The Import directory

The start of the import directory is pointed to by both the IMAGE_DIRECTORY_ENTRY_IAT and IMAGE_DIRECTORY_ENTRY_IMPORT entries of the resource directory (the reason for this is uncertain). At that location, there is an array of IMAGE_IMPORT_DESCRIPTORS structures. Each of these identify a library or module that has a value we need

to import. The array continues until an entry where all the values are zero. The structure is as follows:

struct IMAGE_IMPORT_DESCRIPTOR { long *OriginalFirstThunk; long TimeDateStamp; long ForwarderChain; long Name; long *FirstThunk; }

The TimeDateStamp is relevant to the act of "Binding", see below. The Name value is an RVA to an ASCII string, naming the library to import. ForwarderChain will be explained later. The only thing of interest at this point, are the RVA's OriginalFirstThunk and FirstThunk. Both these values point to arrays of RVA's, each of which point to a IMAGE_IMPORT_BY_NAMES struct. The arrays are terminated with an entry that is equal to zero. These two arrays are parallel and point to the same structure, in the same order. The reason for this will become apparent shortly.

Each of these IMAGE_IMPORT_BY_NAMES structs has the following form:

struct IMAGE_IMPORT_BY_NAME { short Hint; char Name[1]; }

"Name" is an ASCII string of any size that names the value to be imported. This is used when looking up a value in the export directory (see above) through the AddressOfNames array. The "Hint" value is an index into the AddressOfNames array; to save searching for a string, the loader first checks the AddressOfNames entry corresponding to "Hint".

To summarise: The import table consists of a large array of IMAGE_IMPORT_DESCRIPTOR's, terminated by an all-zero entry. These descriptors identify a library to import things from. There are then two parallel RVA arrays, each pointing at IMAGE_IMPORT_BY_NAME structures, which identify a specific value to be imported.

Imports at runtime

Using the above import directory at runtime, the loader finds the appropriate modules, loads them into memory, and seeks the correct export. However, to be able to use the export, a pointer to it must be stored somewhere in the importing module's memory. This is why there are two parallel arrays, OriginalFirstThunk and FirstThunk, identifying IMAGE_IMPORT_BY_NAME structures. Once an imported value has been resolved, the pointer to it is stored in the FirstThunk array. It can then be used at runtime to address imported values.

Bound imports

The PE file format also supports a peculiar feature known as "binding". The process of loading and resolving import addresses can be time consuming, and in some situations this is to be avoided. If a developer is fairly certain that a library is not going to be updated or changed, then the addresses in memory of imported values will not change each time the application is loaded. So, the import address can be precomputed and stored in the FirstThunk array *before* runtime, allowing the loader to skip resolving the imports - the imports are "bound" to a particular memory location. However, if the versions numbers between modules do not match, or the imported library needs to be relocated, the loader will assume the bound addresses are invalid, and resolve the imports anyway.

The "TimeDateStamp" member of the import directory entry for a module controls binding; if it is set to zero, then the import directory is not bound. If it is non-zero, then it is bound to another module. However, the TimeDateStamp in the import table must match the TimeDateStamp in the bound module's FileHeader, otherwise the bound values will be discarded by the loader.

Forwarding and binding Binding can of course be a problem if the bound library / module forwards its exports to another module. In these cases, the non-forwarded imports can be bound, but the values which get forwarded must be identified so the loader can resolve them. This is done through the ForwarderChain member of the import descriptor. The value of "ForwarderChain" is an index into the FirstThunk and OriginalFirstThunk arrays. The OriginalFirstThunk for that index identifies the IMAGE_IMPORT_BY_NAME structure for a import that needs to be resolved, and the FirstThunk for that index is the index of another entry that needs to be resolved. This continues until the FirstThunk value is -1, indicating no more forwarded values to import.

2.2.10 Resources

Resource structures

Resources are data items in modules which are difficult to be stored or described using the chosen programming language. This requires a separate compiler or resource builder, allowing insertion of dialog boxes, icons, menus, images, and other types of resources, including arbitrary binary data. A number of API calls can then be used to retrieve resources from the module. The base of resource data is pointed to by the IMAGE_DIRECTORY_ENTRY_RESOURCE entry of the data directory, and at that location there is an IMAGE_RESOURCE_DIRECTORY structure:

struct IMAGE_RESOURCE_DIRECTORY { long Characteristics; long TimeDateStamp; short MajorVersion; short MinorVersion; short NumberOfNamedEntries; short NumberOfIdEntries; }

Characteristics is unused, and TimeDateStamp is normally the time of creation, although it doesn't matter if

it's set or not. MajorVersion and MinorVersion relate to the versioning info of the resources: the fields have no defined values. Immediately following the IMAGE_RESOURCE_DIRECTORY structure is a series of IMAGE_RESOURCE_DIRECTORY_ENTRY's, the number of which are defined by the total of NumberOfNamedEntries and NumberOfIdEntries. The first portion of these entries are for named resources, the latter for ID resources, depending on the values in the IMAGE_RESOURCE_DIRECTORY struct. The actual shape of the resource entry structure is as follows:

struct IMAGE_RESOURCE_DIRECTORY_ENTRY { long NameId; long *Data; }

The NameId value has dual purpose: if the most significant bit (or sign bit) is clear, then the lower 16 bits are an ID number of the resource. Alternatly, if the top bit is set, then the lower 31 bits make up an offset from the start of the resource data to the name string of this particular resource. The Data value also has a dual purpose: if the most significant bit is set, the remaining 31 bits form an offset from the start of the resource data to another IMAGE_RESOURCE_DIRECTORY (i.e. this entry is an interior node of the resource tree). Otherwise, this is a leaf node, and Data contains the offset from the start of the resource data to a structure which describes the specifics of the resource data itself (which can be considered to be an ordered stream of bytes):

struct IMAGE_RESOURCE_DATA_ENTRY { long *Data; long Size; long CodePage; long Reserved; }

The Data value contains an RVA to the actual resource data, Size is self-explanatory, and CodePage contains the Unicode codepage to be used for decoding Unicode-encoded strings in the resource (if any). Reserved should be set to 0.

Layout

The above system of resource directory and entries allows simple storage of resources, by name or ID number. However, this can get very complicated very quickly. Different types of resources, the resources themselves, and instances of resources in other languages can become muddled in just one directory of resources. For this reason, the resource directory has been given a structure to work by, allowing separation of the different resources.

For this purpose, the "Data" value of resource entries points at another IMAGE_RESOURCE_DIRECTORY structure, forming a tree-diagram like organisation of resources. The first level of resource entries identifies the *type* of the resource: cursors, bitmaps, icons and similar. They use the ID method of identifying the resource entries, of which there are twelve defined values in total. More user defined resource types can be added. Each of these resource entries points at a resource directory, naming the actual resources themselves. These can be of any name or value. These point at yet another resource directory, which uses ID numbers to distinguish languages, allowing different specific resources for systems using a different language. Finally, the entries in the language directory actually provide the offset to the resource data itself, the format of which is not defined by the PE specification and can be treated as an arbitrary stream of bytes.

2.2.11 Relocations

2.2.12 Alternate Bound Import Structure

2.2.13 Windows DLL Files

Windows DLL files are a brand of PE file with a few key differences:

- A .DLL file extension
- A DllMain() entry point, instead of a WinMain() or main().
- The DLL flag set in the PE header.

DLLs may be loaded in one of two ways, a) at load-time, or b) by calling the LoadModule() Win32 API function.

Function Exports

Functions are exported from a DLL file by using the following syntax:

__declspec(dllexport) void MyFunction() ...

The "__declspec" keyword here is not a C language standard, but is implemented by many compilers to set extendable, compiler-specific options for functions and variables. Microsoft C Compiler and GCC versions that run on windows allow for the __declspec keyword, and the dllexport property.

Functions may also be exported from regular .exe files, and .exe files with exported functions may be called dynamically in a similar manner to .dll files. This is a rare occurrence, however.

Identifying DLL Exports

There are several ways to determine which functions are exported by a DLL. The method that this book will use (often implicitly) is to use **dumpbin** in the following manner:

dumpbin /EXPORTS <dll file>

This will post a list of the function exports, along with their ordinal and RVA to the console.

Function Imports

In a similar manner to function exports, a program may import a function from an external DLL file. The dll file will load into the process memory when the program is started, and the function will be used like a local function. DLL imports need to be prototyped in the following manner, for the compiler and linker to recognize that the function is coming from an external library:

__declspec(dllimport) void MyFunction();

Identifying DLL Imports

If is often useful to determine which functions are imported from external libraries when examining a program. To list import files to the console, use **dumpbin** in the following manner:

dumpbin /IMPORTS <dll file>

You can also use depends.exe to list imported and exported functions. Depends is a a GUI tool and comes with Microsoft Platform SDK.

2.3 x86 Disassembly/Linux

The Linux page of the X86 Disassembly Wikibook is a stub. You can help by expanding this section.

2.3.1 Linux

The **GNU/Linux operating system** is open source, but at the same time there is so much that constitutes "GNU/Linux" that it can be difficult to stay on top of all aspects of the system. Here we will attempt to boil down some of the most important concepts of the GNU/Linux Operating System, especially from a reverser's standpoint

2.3.2 System Architecture

The concept of "GNU/Linux" is mostly a collection of a large number of software components that are based off the GNU tools and the Linux kernel. GNU/Linux is itself broken into a number of variants called "distros" which share some similarities, but may also have distinct peculiarities. In a general sense, all GNU/Linux distros are based on a variant of the Linux kernel. However, since each user may edit and recompile their own kernel at will, and since some distros may make certain edits to their kernels, it is hard to proclaim any one version of any one kernel as "the standard". Linux kernels are generally based off the philosophy that system configuration details should be stored in aptly-named, human-readable (and therefore human-editable) configuration files.

The Linux kernel implements much of the core API, but certainly not all of it. Much API code is stored in external modules (although users have the option of compiling all these modules together into a "Monolithic Kernel").

On top of the kernel generally runs one or more **shells**. Bash is one of the more popular shells, but many users prefer other shells, especially for different tasks.

Beyond the shell, Linux distros frequently offer a GUI (although many distros do not have a GUI at all, usually for performance reasons).

Since each GUI often supplies its own underlying framework and API, certain graphical applications may run on only one GUI. Some applications may need to be recompiled (and a few completely rewritten) to run on another GUI.

2.3.3 Configuration Files

2.3.4 Shells

Here are some popular shells:

Bash An acronym for "Bourne Again SHell."

Bourne A precursor to Bash.

Csh C Shell

Ksh Korn Shell

TCsh A Terminal oriented Csh.

Zsh Z Shell

2.3.5 GUIs

Some of the more-popular GUIs:

KDE K Desktop Environment

GNOME GNU Network Object Modeling Environment

2.3.6 Debuggers

gdb The GNU Debugger. It comes pre-installed on most Linux distributions and is primarily used to debug ELF executables. manpage

winedbg A debugger for Wine, used to debug Win32 executables under Linux. manpage

edb A fully featured plugin-based debugger inspired by the famous OllyDbg. Project page

2.3.7 File Analyzers

strings Finds printable strings in a file. When, for example, a password is stored in the binary itself (defined statically in the source), the string can then be extracted from the binary without ever needing to execute it. manpage

file Determines a file type, useful for determining whether an executable has been stripped and whether it's been dynamically (or statically) linked. manpage

objdump Disassembles object files, executables and libraries. Can list internal file structure and disassemble specific sections. Supports both Intel and AT&T syntax

nm Lists symbols from executable files. Doesn't work on stripped binaries. Used mostly on debugging version of executables.

2.4 x86 Disassembly/Linux Executable Files

The Linux Executable Files page of the X86 Disassembly Wikibook is a stub. You can help by expanding this section.

2.4.1 ELF Files

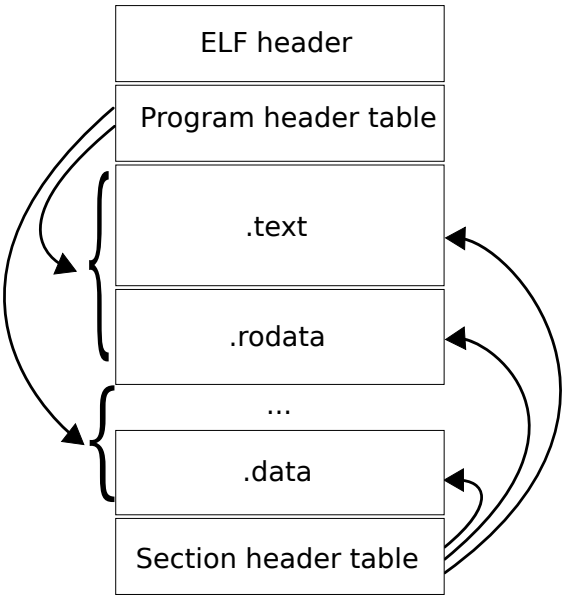
The **ELF file format** (short for Executable and Linking Format) was developed by Unix System Laboratories to be a successor to previous file formats such as COFF and a.out. In many respects, the ELF format is more powerful and versatile than previous formats, and has widely become the standard on Linux, Solaris, IRIX, and FreeBSD (although the FreeBSD-derived Mac OS X uses the Mach-O format instead). ELF has also been adopted by OpenVMS for Itanium and BeOS for x86.

Historically, Linux has not always used ELF; Red Hat Linux 4 was the first time that distribution used ELF; previous versions had used the a.out format.

ELF Objects are broken down into different segments and/or sections. These can be located by using the ELF header found at the first byte of the object. The ELF header provides the location for both the program header and the section header. Using these data structures the rest of the ELF objects contents can be found, this includes .text and .data segments which contain code and data respectively.

The GNU readelf utility, from the binutils package, is a common tool for parsing ELF objects.

File Format



An ELF file has two views: the program header shows the segments *used at run-time, while the section header lists the set of* sections *of the binary.*

Each ELF file is made up of one ELF header, followed by file data. The file data can include:

- Program header table, describing zero or more segments
- Section header table, describing zero or more sections
- Data referred to by entries in the program or section header table

The segments contain information that is necessary for runtime execution of the file, while sections contain important data for linking and relocation. Each byte in the entire file is taken by no more than one section at a time, but there can be orphan bytes, which are not covered by a section. In the normal case of a Unix executable one or more sections are enclosed in one segment.

2.4.2 Relocatable ELF Files

Relocatable ELF files are created by compilers. They need to be linked before running.

Those files are often found in .a archives, with a .o extension.

2.4.3 a.out Files

a.out is a very simple format consisting of a header (at offset 0) which contains the size of 3 executable sections

(code, data, bss), plus pointers to additional information such as relocations (for .o files), symbols and symbols' strings. The actual sections contents follows the header. Offsets of different sections are computed from the size of the previous section.

The a.out format is now rarely used.

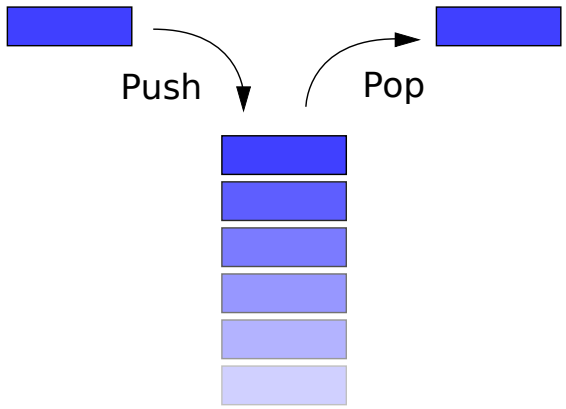
File Format

Chapter 3

Code Patterns

3.1 x86 Disassembly/The Stack

3.1.1 The Stack



Generally speaking, a **stack** is a data structure that stores data values contiguously in memory. Unlike an array, however, you access (read or write) data only at the “top” of the stack. To read from the stack is said "**to pop**" and to write to the stack is said "**to push**". A stack is also known as a LIFO queue (Last In First Out) since values are popped from the stack in *the reverse order* that they are pushed onto it (think of how you pile up plates on a table). Popped data disappears from the stack.

All x86 architectures use a stack as a temporary storage area in RAM that allows the processor to quickly store and retrieve data in memory. The current top of the stack is pointed to by the **esp** register. The stack “grows” downward, from high to low memory addresses, so values recently pushed onto the stack are located in memory addresses *above* the esp pointer. No register specifically points to the bottom of the stack, although most operating systems monitor the stack bounds to detect both “underflow” (popping an empty stack) and “overflow” (pushing too much information on the stack) conditions.

When a value is popped off the stack, the value remains sitting in memory until overwritten. However, you should never rely on the content of memory addresses below esp, because other functions may overwrite these values without your knowledge.

Users of Windows ME, 98, 95, 3.1 (and earlier) may fondly remember the infamous “Blue Screen of Death” -- that was sometimes caused by a stack overflow exception. This occurs when too much data is written to the stack, and the stack “grows” beyond its limits. Modern operating systems use better bounds-checking and error recovery to reduce the occurrence of stack overflows, and to maintain system stability after one has occurred.

3.1.2 Push and Pop

The following lines of ASM code are basically equivalent:

but the single command actually performs much faster than the alternative. It can be visualized that the stack grows from right to left, and esp decreases as the stack grows in size.

3.1.3 ESP In Action

Let’s say we want to quickly discard 3 items we pushed earlier onto the stack, without saving the values (in other words “clean” the stack). The following works:

pop eax pop eax pop eax

However there is a faster method. We can simply perform some basic arithmetic on esp to make the pointer go “above” the data items, so they cannot be read anymore, and can be overwritten with the next round of **push** commands.

add esp, 12 ; 12 is 3 DWORDs (4 bytes * 3)

Likewise, if we want to reserve room on the stack for an item bigger than a DWORD, we can use a subtraction

to artificially move esp forward. We can then access our reserved memory directly as a memory pointer, or we can access it indirectly as an offset value from esp itself.

Say we wanted to create an array of byte values on the stack, 100 items long. We want to store the pointer to the base of this array in **edi**. How do we do it? Here is an example:

sub esp, 100 ; num of bytes in our array mov edi, esp ; copy address of 100 bytes area to edi

To destroy that array, we simply write the instruction

add esp, 100

3.1.4 Reading Without Popping

To read values on the stack without popping them off the stack, **esp** can be used with an offset. For instance, to read the 3 DWORD values from the top of the stack into eax (but without using a pop instruction), we would use the instructions:

mov eax, DWORD PTR SS:[esp] mov eax, DWORD PTR SS:[esp + 4] mov eax, DWORD PTR SS:[esp + 8]

Remember, since esp moves downward as the stack grows, data on the stack can be accessed with a positive offset. A negative offset should never be used because data “above” the stack cannot be counted on to stay the way you left it. The operation of reading from the stack without popping is often referred to as “peeking”, but since this isn't the official term for it this wikibook won't use it.

3.1.5 Data Allocation

There are two areas in the computer memory where a program can store data. The first, the one that we have been talking about, is the stack. It is a linear LIFO buffer that allows fast allocations and deallocations, but has a limited size. The **heap** is typically a non-linear data storage area, typically implemented using linked lists, binary trees, or other more exotic methods. Heaps are slightly more difficult to interface with and to maintain than a stack, and allocations/deallocations are performed more slowly. However, heaps can grow as the data grows, and new heaps can be allocated when data quantities become too large.

As we shall see, explicitly declared variables are allocated on the stack. Stack variables are finite in number, and have a definite size. Heap variables can be variable in number and in size. We will discuss these topics in more detail later.

3.2 x86 Disassembly/Functions and Stack Frames

3.2.1 Functions and Stack Frames

To allow for many unknowns in the execution environment, functions are frequently set up with a "**stack frame**" to allow access to both function parameters, and automatic function variables. The idea behind a stack frame is that each subroutine can act independently of its location on the stack, and each subroutine can act as if it is the top of the stack.

When a function is called, a new stack frame is created at the current **esp** location. A stack frame acts like a partition on the stack. All items from previous functions are higher up on the stack, and should not be modified. Each current function has access to the remainder of the stack, from the stack frame until the end of the stack page. The current function always has access to the “top” of the stack, and so functions do not need to take account of the memory usage of other functions or programs.

3.2.2 Standard Entry Sequence

For many compilers, the standard function entry sequence is the following piece of code (*X* is the total size, in bytes, of all *automatic* variables used in the function):

push ebp mov ebp, esp sub esp, X

For example, here is a C function code fragment and the resulting assembly instructions:

void MyFunction() { int a, b, c; ...
_MyFunction: push ebp ; save the value of ebp mov ebp, esp ; ebp now points to the top of the stack sub esp, 12 ; space allocated on the stack for the local variables

This means local variables can be accessed by referencing ebp. Consider the following C code fragment and corresponding assembly code:

a = 10; b = 5; c = 2;
mov [ebp - 4], 10 ; location of variable a mov [ebp - 8], 5 ; location of b mov [ebp - 12], 2 ; location of c

This all seems well and good, but what is the purpose of **ebp** in this setup? Why save the old value of ebp and then point ebp to the top of the stack, only to change the value of esp with the next instruction? The answer is *function parameters*.

Consider the following C function declaration:

void MyFunction2(int x, int y, int z) { ... }

It produces the following assembly code:

_MyFunction2: push ebp mov ebp, esp sub esp, 0 ; no local variables, most compilers will omit this line

Which is exactly as one would expect. So, what exactly does **ebp** do, and where are the function parameters stored? The answer is found when we call the function.

Consider the following C function call:

MyFunction2(10, 5, 2);

This will create the following assembly code (using a Right-to-Left calling convention called CDECL, explained later):

push 2 push 5 push 10 call _MyFunction2

Note: Remember that the **call** x86 instruction is basically equivalent to

push eip + 2 ; return address is current address + size of two instructions jmp _MyFunction2

It turns out that the function arguments are all passed on the stack! Therefore, when we move the current value of the stack pointer (**esp**) into **ebp**, we are pointing ebp directly at the function arguments. As the function code pushes and pops values, ebp is not affected by esp. Remember that pushing basically does this:

sub esp, 4 ; "allocate" space for the new stack item mov [esp], X ; put new stack item value X in

This means that first the return address and then the old value of **ebp** are put on the stack. Therefore [ebp] points to the location of the old value of ebp, [ebp + 4] points to the return address, and [ebp + 8] points to the first function argument. Here is a (crude) representation of the stack at this point:

: : | 2 | [ebp + 16] (3rd function argument) | 5 | [ebp + 12] (2nd argument) | 10 | [ebp + 8] (1st argument) | RA | [ebp + 4] (return address) | FP | [ebp] (old ebp value) | | [ebp - 4] (1st local variable) : :

The stack pointer value may change during the execution of the current function. In particular this happens when:

- parameters are passed to another function;
- the pseudo-function "alloca()" is used.

[FIXME: When parameters are passed into another function the esp changing is not an issue. When that function returns the esp will be back to its old value. So why does ebp help there. This needs better explanation. (The real explanation is here, ESP is not really needed: http://blogs.msdn.com/larryosterman/archive/2007/03/12/fpo.aspx)] This means that the value of **esp** cannot be reliably used to determine (using the appropriate offset) the memory location of a specific local variable. To solve this problem, many compilers access local variables using negative offsets from the **ebp** registers. This allows us to assume that the same offset is always used to access the same variable (or parameter). For this reason, the ebp register is called the **frame pointer**, or FP.

3.2.3 Standard Exit Sequence

The Standard Exit Sequence must undo the things that the Standard Entry Sequence does. To this effect, the Standard Exit Sequence must perform the following tasks, in the following order:

1. Remove space for local variables, by reverting **esp** to its old value.
2. Restore the old value of **ebp** to its old value, which is on top of the stack.
3. Return to the calling function with a *ret* command.

As an example, the following C code:

void MyFunction3(int x, int y, int z) { int a, int b, int c; ... return; }

Will create the following assembly code:

_MyFunction3: push ebp mov ebp, esp sub esp, 12 ; sizeof(a) + sizeof(b) + sizeof(c) ;x = [ebp + 8], y = [ebp + 12], z = [ebp + 16] ;a = [ebp - 4] = [esp + 8], b = [ebp - 8] = [esp + 4], c = [ebp - 12] = [esp] mov esp, ebp pop ebp ret 12 ; sizeof(x) + sizeof(y) + sizeof(z)

3.2.4 Non-Standard Stack Frames

Frequently, reversers will come across a subroutine that doesn't set up a standard stack frame. Here are some things to consider when looking at a subroutine that does not start with a standard sequence:

Using Uninitialized Registers

When a subroutine starts using data in an *uninitialized* register, that means that the subroutine expects external functions to put data into that register before it gets called. Some calling conventions pass arguments in registers, but sometimes a compiler will not use a standard calling convention.

"static" Functions

In C, functions may optionally be declared with the **static** keyword, as such:

static void MyFunction4();

The **static** keyword causes a function to have only local scope, meaning it may not be accessed by any external functions (it is strictly internal to the given code file). When an optimizing compiler sees a static function that is only referenced by calls (no references through function pointers), it "knows" that external functions cannot possibly interface with the static function (the compiler controls all access to the function), so the compiler doesn't bother making it standard.

Hot Patch Prologue

Some Windows functions set up a regular stack frame as explained above, but start out with the apparently nonsensical line

mov edi, edi;

This instruction is assembled into 2 bytes which serve as a placeholder for future function patches. Taken as a whole such a function might look like this:

nop ; each nop is 1 byte long nop nop nop nop FUNCTION: ; <-- This is the function entry point as used by call instructions mov edi, edi ; mov edi,edi is 2 bytes long push ebp ; regular stack frame setup mov ebp, esp

If such a function needs to be replaced without reloading the application (or restarting the machine in case of kernel patches) it can be achieved by inserting a jump to the replacement function. A short jump instruction (which can jump +/- 127 bytes) requires 2 bytes of storage space - just the amount that the "mov edi,edi" placeholder provides. A jump to any memory location, in this case to our replacement function, requires 5 bytes. These are provided by the 5 no-operation bytes just preceding the function. If a function thus patched gets called it will first jump back by 5 bytes and then do a long jump to the replacement function. After the patch the memory might look like this

LABEL: jmp REPLACEMENT_FUNCTION ; <-- 5 NOPs replaced by jmp FUNCTION: jmp short LABEL ; <-- mov edi has been replaced by short jump backwards push ebp mov ebp, esp ; <-- regular stack frame setup as before

The reason for using a 2-byte mov instruction at the beginning instead of putting 5 nops there directly, is to prevent corruption during the patching process. There would be a risk with replacing 5 individual instructions if the instruction pointer is currently pointing at any one of them. Using a single mov instruction as placeholder on the other hand guarantees that the patching can be completed as an atomic transaction.

3.2.5 Local Static Variables

Local static variables cannot be created on the stack, since the value of the variable is preserved across function calls. We'll discuss local static variables and other types of variables in a later chapter.

3.3 x86 Disassembly/Functions and Stack Frame Examples

3.3.1 Example: Number of Parameters

Given the following disassembled function (in MASM syntax), how many 4-byte parameters does this function receive? How many variables are created on the stack? What does this function do?

_Question1: push ebp mov ebp, esp sub esp, 4 mov eax, [ebp + 8] mov ecx, 2 mul ecx mov [esp + 0], eax mov eax, [ebp + 12] mov edx, [esp + 0] add eax, edx mov esp, ebp pop ebp ret

The function above takes 2 4-byte parameters, accessed by offsets +8 and +12 from ebp. The function also has 1 variable created on the stack, accessed by offset +0 from esp. The function is nearly identical to this C code:

int Question1(int x, int y) { int z; z = x * 2; return y + z; }

3.3.2 Example: Standard Entry Sequences

Does the following function follow the Standard Entry and Exit Sequences? if not, where does it differ?

_Question2: call _SubQuestion2 mov ecx, 2 mul ecx ret

The function does not follow the standard entry sequence, because it doesn't set up a proper stack frame with ebp and esp. The function basically performs the following C instructions:

int Question2() { return SubQuestion2() * 2; }

Although an optimizing compiler has chosen to take a few shortcuts.

3.4 x86 Disassembly/Calling Conventions

3.4.1 Calling Conventions

Calling conventions are a standardized method for functions to be implemented and called by the machine. A calling convention specifies the method that a compiler sets up to access a subroutine. In theory, code from any compiler can be interfaced together, so long as the functions all have the same calling conventions. In practice however, this is not always the case.

Calling conventions specify how arguments are passed to a function, how return values are passed back out of a function, how the function is called, and how the function manages the stack and its stack frame. In short, the calling convention specifies how a function call in C or C++ is converted into assembly language. Needless to say, there are many ways for this translation to occur, which is why it's so important to specify certain standard methods. If these standard conventions did not exist, it would be nearly impossible for programs created using different compilers to communicate and interact with one another.

There are three major calling conventions that are used with the C language: STDCALL, CDECL, and FASTCALL. In addition, there is another calling convention typically used with C++: THISCALL. There are other calling conventions as well, including PASCAL and FORTRAN conventions, among others. We will not consider those conventions in this book.

3.4.2 Notes on Terminology

There are a few terms that we are going to be using in this chapter, which are mostly common sense, but which are worthy of stating directly:

Passing arguments "passing arguments" is a way of saying that the calling function is writing data in the place where the called function will look for them. Arguments are passed before the *call* instruction is executed.

Right-to-Left and Left-to-Right These describe the manner that arguments are passed to the subroutine, in terms of the High-level code. For instance, the following C function call:

MyFunction1(a, b);

will generate the following code if passed Left-to-Right:

push a push b call _MyFunction

and will generate the following code if passed Right-to-Left:

push b push a call _MyFunction

Return value Some functions return a value, and that value must be received reliably by the function's caller. The called function places its return value in a place where the calling function can get it when execution returns. The called function stores the return value before executing the *ret* instruction.

Cleaning the stack When arguments are pushed onto the stack, eventually they must be popped back off again. Whichever function, the caller or the callee, is responsible for cleaning the stack must reset the stack pointer to eliminate the passed arguments.

Calling function (the caller) The "parent" function that calls the subroutine. Execution resumes in the calling function directly after the subroutine call, unless the program terminates inside the subroutine.

Called function (the callee) The "child" function that gets called by the "parent."

Name Decoration When C code is translated to assembly code, the compiler will often "decorate" the function name by adding extra information that the linker will use to find and link to the correct functions. For most calling conventions, the decoration is very simple (often only an extra symbol or two to denote the calling convention), but in some extreme cases (notably C++ "thiscall" convention), the names are "mangled" severely.

Entry sequence (the function prologue) a few instructions at the beginning of a function, which prepare the stack and registers for use within the function.

Exit sequence (the function epilogue) a few instructions at the end of a function, which restore the stack and registers to the state expected by the caller, and return to the caller. Some calling conventions clean the stack in the exit sequence.

Call sequence a few instructions in the middle of a function (the caller) which pass the arguments and call the called function. After the called function has returned, some calling conventions have one more instruction in the call sequence to clean the stack.

3.4.3 Standard C Calling Conventions

The C language, by default, uses the CDECL calling convention, but most compilers allow the programmer to specify another convention via a specifier keyword. These keywords **are not** part of the ISO-ANSI C standard, so you should always check with your compiler documentation about implementation specifics.

If a calling convention other than CDECL is to be used, or if CDECL is not the default for your compiler, and you want to manually use it, you must specify the calling convention keyword in the function declaration itself, and in any prototypes for the function. This is important because both the calling function and the called function need to know the calling convention.

CDECL

In the CDECL calling convention the following holds:

- Arguments are passed on the stack in Right-to-Left order, and return values are passed in eax.
- The *calling* function cleans the stack. This allows CDECL functions to have *variable-length argument lists* (aka variadic functions). For this reason the number of arguments is not appended to the name of the function by the compiler, and the assembler and the linker are therefore unable to determine if an incorrect number of arguments is used.

Variadic functions usually have special entry code, generated by the va_start(), va_arg() C pseudo-functions.

Consider the following C instructions:

_cdecl int MyFunction1(int a, int b) { return a + b; }

and the following function call:

x = MyFunction1(2, 3);

These would produce the following assembly listings, respectively:

_MyFunction1: push ebp mov ebp, esp mov eax, [ebp + 8] mov edx, [ebp + 12] add eax, edx pop ebp ret

and

push 3 push 2 call _MyFunction1 add esp, 8

When translated to assembly code, CDECL functions are almost always prepended with an underscore (that's why all previous examples have used "_" in the assembly code).

STDCALL

STDCALL, also known as "WINAPI" (and a few other names, depending on where you are reading it) is used almost exclusively by Microsoft as the standard calling convention for the Win32 API. Since STDCALL is strictly defined by Microsoft, all compilers that implement it do it the same way.

- STDCALL passes arguments right-to-left, and returns the value in eax. (The Microsoft documentation erroneously claimed that arguments are passed left-to-right, but this is not the case.)
- The called function cleans the stack, unlike CDECL. This means that STDCALL doesn't allow variable-length argument lists.

Consider the following C function:

_stdcall int MyFunction2(int a, int b) { return a + b; }

and the calling instruction:

x = MyFunction2(2, 3);

These will produce the following respective assembly code fragments:

:_MyFunction@8 push ebp mov ebp, esp mov eax, [ebp + 8] mov edx, [ebp + 12] add eax, edx pop ebp ret 8

and

push 3 push 2 call _MyFunction@8

There are a few important points to note here:

1. In the function body, the *ret* instruction has an (optional) argument that indicates how many bytes to pop off the stack when the function returns.
2. STDCALL functions are name-decorated with a leading underscore, followed by an @, and then the number (in bytes) of arguments passed on the stack. This number will always be a multiple of 4, on a 32-bit aligned machine.

FASTCALL

The FASTCALL calling convention is not completely standard across all compilers, so it should be used with caution. In FASTCALL, the first 2 or 3 32-bit (or smaller) arguments are passed in registers, with the most commonly used registers being edx, eax, and ecx. Additional arguments, or arguments larger than 4-bytes are passed on the stack, often in Right-to-Left order (similar to CDECL). The calling function most frequently is responsible for cleaning the stack, if needed.

Because of the ambiguities, it is recommended that FASTCALL be used only in situations with 1, 2, or 3 32-bit arguments, where speed is essential.

The following C function:

_fastcall int MyFunction3(int a, int b) { return a + b; }

and the following C function call:

x = MyFunction3(2, 3);

Will produce the following assembly code fragments for the called, and the calling functions, respectively:

:@MyFunction3@8 push ebp mov ebp, esp ;many compilers create a stack frame even if it isn't used add eax, edx ;a is in eax, b is in edx pop ebp ret

and

;the calling function mov eax, 2 mov edx, 3 call @MyFunction3@8

The name decoration for FASTCALL prepends an @ to the function name, and follows the function name with @x, where x is the number (in bytes) of arguments passed to the function.

Many compilers still produce a stack frame for FASTCALL functions, especially in situations where the FASTCALL function itself calls another subroutine. However, if a FASTCALL function doesn't need a stack frame, optimizing compilers are free to omit it.

3.4.4 C++ Calling Convention

C++ requires that non-static methods of a class be called by an instance of the class. Therefore it uses its own standard calling convention to ensure that pointers to the object are passed to the function: **THISCALL**.

THISCALL

In THISCALL, the pointer to the class object is passed in ecx, the arguments are passed Right-to-Left on the stack, and the return value is passed in eax.

For instance, the following C++ instruction:

MyObj.MyMethod(a, b, c);

Would form the following asm code:

mov ecx, MyObj push c push b push a call _MyMethod

At least, it *would* look like the assembly code above if it weren't for **name mangling**.

Name Mangling

Because of the complexities inherent in function overloading, C++ functions are heavily name-decorated to the point that people often refer to the process as "Name Mangling." Unfortunately C++ compilers are free to do the name-mangling differently since the standard does not enforce a convention. Additionally, other issues such as exception handling are also not standardized.

Since every compiler does the name-mangling differently, this book will not spend too much time discussing the specifics of the algorithm. Notice that in many cases, it's possible to determine which compiler created the executable by examining the specifics of the name-mangling format. We will not cover this topic in this much depth in this book, however.

Here are a few general remarks about THISCALL name-mangled functions:

- They are recognizable on sight because of their complexity when compared to CDECL, FASTCALL, and STDCALL function name decorations
- They sometimes include the name of that function's class.
- They almost always include the number and type of the arguments, so that overloaded functions can be differentiated by the arguments passed to it.

Here is an example of a C++ class and function declaration:

class MyClass { MyFunction(int a); } MyClass::MyFunction(2) { }

And here is the resultant mangled name:

?MyFunction@MyClass@@QAEHH@Z

Extern "C"

In a C++ source file, functions placed in an extern "C" block are guaranteed not to be mangled. This is done frequently when libraries are written in C++, and the functions need to be exported without being mangled. Even though the program is written in C++ and compiled with a C++ compiler, some of the functions might therefore not be mangled and will use one of the ordinary C calling conventions (typically CDECL).

3.4.5 Note on Name Decorations

We've been discussing name decorations in this chapter, but the fact is that in pure disassembled code there typically are no names whatsoever, especially not names with fancy decorations. The assembly stage removes all these

readable identifiers, and replaces them with the binary locations instead. Function names really only appear in two places:

1. Listing files produced during compilation
2. In export tables, if functions are exported

When disassembling raw machine code, there will be no function names and no name decorations to examine. For this reason, you will need to pay more attention to the way parameters are passed, the way the stack is cleaned, and other similar details.

While we haven't covered optimizations yet, suffice it to say that optimizing compilers can even make a mess out of these details. Functions which are not exported do not necessarily need to maintain standard interfaces, and if it is determined that a particular function does not need to follow a standard convention, some of the details will be optimized away. In these cases, it can be difficult to determine what calling conventions were used (if any), and it is even difficult to determine where a function begins and ends. This book cannot account for all possibilities, so we try to show as much information as possible, with the knowledge that much of the information provided here will not be available in a true disassembly situation.

3.4.6 further reading

- x86 Disassembly/Calling Convention Examples
- Embedded Systems/Mixed C and Assembly Programming describes calling conventions on other CPUs.

3.5 x86 Disassembly/Calling Convention Examples

3.5.1 Microsoft C Compiler

Here is a simple function in C:

int MyFunction(int x, int y) { return (x * 2) + (y * 3); }

Using cl.exe, we are going to generate 3 separate listings for MyFunction, one with CDECL, one with FASTCALL, and one with STDCALL calling conventions. On the commandline, there are several switches that you can use to force the compiler to change the default:

- /Gd : The default calling convention is CDECL
- /Gr : The default calling convention is FASTCALL
- /Gz : The default calling convention is STDCALL

Using these commandline options, here are the listings:

CDECL

int MyFunction(int x, int y) { return (x * 2) + (y * 3); }

becomes:

PUBLIC _MyFunction _TEXT SEGMENT _x$ = 8 ; size = 4 _y$ = 12 ; size = 4 _MyFunction PROC NEAR ; Line 4 push ebp mov ebp, esp ; Line 5 mov eax, _y$[ebp] imul eax, 3 mov ecx, _x$[ebp] lea eax, [eax+ecx*2] ; Line 6 pop ebp ret 0 _MyFunction ENDP _TEXT ENDS END

On entry of a function, ESP points to the return address pushed on the stack by the call instruction (that is, previous contents of EIP). Any argument in stack of higher address than entry ESP is pushed by caller before the call is made; in this example, the first argument is at offset +4 from ESP (EIP is 4 bytes wide), plus 4 more bytes once the EBP is pushed on the stack. Thus, at line 5, ESP points to the saved frame pointer EBP, and arguments are located at addresses ESP+8 (x) and ESP+12 (y).

For CDECL, caller pushes arguments into stack in a right to left order. Because ret 0 is used, it must be the caller who cleans up the stack.

As a point of interest, notice how **lea** is used in this function to simultaneously perform the multiplication (ecx * 2), and the addition of that quantity to eax. Unintuitive instructions like this will be explored further in the chapter on unintuitive instructions.

FASTCALL

int MyFunction(int x, int y) { return (x * 2) + (y * 3); }

becomes:

PUBLIC @MyFunction@8 _TEXT SEGMENT _y$ = −8 ; size = 4 _x$ = −4 ; size = 4 @MyFunction@8 PROC NEAR ; _x$ = ecx ; _y$ = edx ; Line 4 push ebp mov ebp, esp sub esp, 8 mov _y$[ebp], edx mov _x$[ebp], ecx ; Line 5 mov eax, _y$[ebp] imul eax, 3 mov ecx, _x$[ebp] lea eax, [eax+ecx*2] ; Line 6 mov esp, ebp pop ebp ret 0 @MyFunction@8 ENDP _TEXT ENDS END

This function was compiled with optimizations turned off. Here we see arguments are first saved in stack then fetched from stack, rather than be used directly. This is because the compiler wants a consistent way to use all arguments via stack access, not only one compiler does like that.

There is no argument is accessed with positive offset to entry SP, it seems caller doesn't pushed in them, thus it can use ret 0. Let's do further investigation:

int FastTest(int x, int y, int z, int a, int b, int c) { return x * y * z * a * b * c; }

and the corresponding listing:

PUBLIC @FastTest@24 _TEXT SEGMENT _y$ = −8 ; size = 4 _x$ = −4 ; size = 4 _z$ = 8 ; size = 4 _a$ = 12 ; size = 4 _b$ = 16 ; size = 4 _c$ = 20 ; size = 4 @FastTest@24 PROC NEAR ; _x$ = ecx ; _y$ = edx ; Line 2 push ebp mov ebp, esp sub esp, 8 mov _y$[ebp], edx mov _x$[ebp], ecx ; Line 3 mov eax, _x$[ebp] imul eax, DWORD PTR _y$[ebp] imul eax, DWORD PTR _z$[ebp] imul eax, DWORD PTR _a$[ebp] imul eax, DWORD PTR _b$[ebp] imul eax, DWORD PTR _c$[ebp] ; Line 4 mov esp, ebp pop ebp ret 16 ; 00000010H

Now we have 6 arguments, four are pushed in by caller from right to left, and last two are passed again in cx/dx, and processed the same way as previous example. Stack cleanup is done by ret 16, which corresponding to 4 arguments pushed before call executed.

For FASTCALL, compiler will try to pass arguments in registers, if not enough caller will pushed them into stack still in an order from right to left. Stack cleanup is done by callee. It is called FASTCALL because if arguments can be passed in registers (for 64bit CPU the maximum number is 6), no stack push/clean up is needed.

The name-decoration scheme of the function: @MyFunction@n, here n is stack size needed for all arguments.

STDCALL

int MyFunction(int x, int y) { return (x * 2) + (y * 3); }

becomes:

PUBLIC _MyFunction@8 _TEXT SEGMENT _x$ = 8 ; size = 4 _y$ = 12 ; size = 4 _MyFunction@8 PROC NEAR ; Line 4 push ebp mov ebp, esp ; Line 5 mov eax, _y$[ebp] imul eax, 3 mov ecx, _x$[ebp] lea eax, [eax+ecx*2] ; Line 6 pop ebp ret 8 _MyFunction@8 ENDP _TEXT ENDS END

The STDCALL listing has only one difference than the CDECL listing that it uses "ret 8" for self clean up of stack. Lets do an example with more parameters:

int STDCALLTest(int x, int y, int z, int a, int b, int c) { return x * y * z * a * b * c; }

Let's take a look at how this function gets translated into assembly by cl.exe:

PUBLIC _STDCALLTest@24 _TEXT SEGMENT _x$ = 8 ; size = 4 _y$ = 12 ; size = 4 _z$ = 16 ; size = 4 _a$ = 20 ; size = 4 _b$ = 24 ; size = 4 _c$ = 28 ; size = 4 _STDCALLTest@24 PROC NEAR ; Line 2 push ebp mov ebp, esp ; Line 3 mov eax, _x$[ebp] imul eax, DWORD PTR _y$[ebp] imul eax, DWORD PTR _z$[ebp] imul eax, DWORD PTR _a$[ebp] imul eax, DWORD PTR _b$[ebp] imul eax, DWORD PTR _c$[ebp] ; Line 4 pop ebp ret 24 ; 00000018H _STDCALLTest@24 ENDP _TEXT ENDS END

Yes the only difference between STDCALL and CDECL is that the former does stack clean up in callee, the later in caller. This saves a little bit in X86 due to its "ret n".

3.5.2 GNU C Compiler

We will be using 2 example C functions to demonstrate how GCC implements calling conventions:

int MyFunction1(int x, int y) { return (x * 2) + (y * 3); }

and

int MyFunction2(int x, int y, int z, int a, int b, int c) { return x * y * (z + 1) * (a + 2) * (b + 3) * (c + 4); }

GCC does not have commandline arguments to force the default calling convention to change from CDECL (for C), so they will be manually defined in the text with the directives: __cdecl, __fastcall, and __stdcall.

CDECL

The first function (MyFunction1) provides the following assembly listing:

_MyFunction1: pushl %ebp movl %esp, %ebp movl 8(%ebp), %eax leal (%eax,%eax), %ecx movl 12(%ebp), %edx movl %edx, %eax addl %eax, %eax addl %edx, %eax leal (%eax,%ecx), %eax popl %ebp ret

First of all, we can see the name-decoration is the same as in cl.exe. We can also see that the ret instruction doesn't have an argument, so the calling function is cleaning the stack. However, since GCC doesn't provide us with the variable names in the listing, we have to deduce which parameters are which. After the stack frame is set up, the first instruction of the function is "movl 8(%ebp), %eax". One we remember (or learn for the first time) that GAS instructions have the general form:

instruction src, dest

We realize that the value at offset +8 from ebp (the last parameter pushed on the stack) is moved into eax. The leal instruction is a little more difficult to decipher, especially if we don't have any experience with GAS instruc-

tions. The form "leal(reg1,reg2), dest" adds the values in the parenthesis together, and stores the value in *dest*. Translated into Intel syntax, we get the instruction:

lea ecx, [eax + eax]

Which is clearly the same as a multiplication by 2. The first value accessed must then have been the last value passed, which would seem to indicate that values are passed right-to-left here. To prove this, we will look at the next section of the listing:

movl 12(%ebp), %edx movl %edx, %eax addl %eax, %eax addl %edx, %eax leal (%eax,%ecx), %eax

the value at offset +12 from ebp is moved into edx. edx is then moved into eax. eax is then added to itselt (eax * 2), and then is added back to edx (edx + eax). remember though that eax = 2 * edx, so the result is edx * 3. This then is clearly the y parameter, which is furthest on the stack, and was therefore the first pushed. CDECL then on GCC is implemented by passing arguments on the stack in right-to-left order, same as cl.exe.

FASTCALL

.globl @MyFunction1@8 .def @MyFunction1@8; .scl 2; .type 32; .endef @MyFunction1@8: pushl %ebp movl %esp, %ebp subl $8, %esp movl %ecx, −4(%ebp) movl %edx, −8(%ebp) movl −4(%ebp), %eax leal (%eax,%eax), %ecx movl −8(%ebp), %edx movl %edx, %eax addl %eax, %eax addl %edx, %eax leal (%eax,%ecx), %eax leave ret

Notice first that the same name decoration is used as in cl.exe. The astute observer will already have realized that GCC uses the same trick as cl.exe, of moving the fastcall arguments from their registers (ecx and edx again) onto a negative offset on the stack. Again, optimizations are turned off. ecx is moved into the first position (−4) and edx is moved into the second position (−8). Like the CDECL example above, the value at −4 is doubled, and the value at −8 is tripled. Therefore, −4 (ecx) is x, and −8 (edx) is y. It would seem from this listing then that values are passed left-to-right, although we will need to take a look at the larger, MyFunction2 example:

.globl @MyFunction2@24 .def @MyFunction2@24; .scl 2; .type 32; .endef @MyFunction2@24: pushl %ebp movl %esp, %ebp subl $8, %esp movl %ecx, −4(%ebp) movl %edx, −8(%ebp) movl −4(%ebp), %eax imull −8(%ebp), %eax movl 8(%ebp), %edx incl %edx imull %edx, %eax movl 12(%ebp), %edx addl $2, %edx imull %edx, %eax movl 16(%ebp), %edx addl $3, %edx imull %edx, %eax movl 20(%ebp), %edx addl $4, %edx imull %edx, %eax leave ret $16

By following the fact that in MyFunction2, successive parameters are added to increasing constants, we can deduce the positions of each parameter. −4 is still x, and −8 is still y. +8 gets incremented by 1 (z), +12 gets increased by 2 (a). +16 gets increased by 3 (b), and +20 gets increased by 4 (c). Let's list these values then:

z = [ebp + 8] a = [ebp + 12] b = [ebp + 16] c = [ebp + 20]

c is the furthest down, and therefore was the first pushed. z is the highest to the top, and was therefore the last pushed. Arguments are therefore pushed in right-to-left order, just like cl.exe.

STDCALL

Let's compare then the implementation of MyFunction1 in GCC:

.globl _MyFunction1@8 .def _MyFunction1@8; .scl 2; .type 32; .endef _MyFunction1@8: pushl %ebp movl %esp, %ebp movl 8(%ebp), %eax leal (%eax,%eax), %ecx movl 12(%ebp), %edx movl %edx, %eax addl %eax, %eax addl %edx, %eax leal (%eax,%ecx), %eax popl %ebp ret $8

The name decoration is the same as in cl.exe, so STDCALL functions (and CDECL and FASTCALL for that matter) can be assembled with either compiler, and linked with either linker, it seems. The stack frame is set up, then the value at [ebp + 8] is doubled. After that, the value at [ebp + 12] is tripled. Therefore, +8 is x, and +12 is y. Again, these values are pushed in right-to-left order. This function also cleans its own stack with the "ret 8" instruction.

Looking at a bigger example:

.globl _MyFunction2@24 .def _MyFunction2@24; .scl 2; .type 32; .endef _MyFunction2@24: pushl %ebp movl %esp, %ebp movl 8(%ebp), %eax imull 12(%ebp), %eax movl 16(%ebp), %edx incl %edx imull %edx, %eax movl 20(%ebp), %edx addl $2, %edx imull %edx, %eax movl 24(%ebp), %edx addl $3, %edx imull %edx, %eax movl 28(%ebp), %edx addl $4, %edx imull %edx, %eax popl %ebp ret $24

We can see here that values at +8 and +12 from ebp are still x and y, respectively. The value at +16 is incremented by 1, the value at +20 is incremented by 2, etc all the way to the value at +28. We can therefore create the following table:

x = [ebp + 8] y = [ebp + 12] z = [ebp + 16] a = [ebp + 20] b = [ebp + 24] c = [ebp + 28]

With c being pushed first, and x being pushed last. Therefore, these parameters are also pushed in right-to-left order. This function then also cleans 24 bytes off the stack with the "ret 24" instruction.

3.5.3 Example: C Calling Conventions

Identify the calling convention of the following C function:

int MyFunction(int a, int b) { return a + b; }

The function is written in C, and has no other specifiers, so it is CDECL by default.

3.5.4 Example: Named Assembly Function

Identify the calling convention of the function **MyFunction**:

:_MyFunction@12 push ebp mov ebp, esp ... pop ebp ret 12

The function includes the decorated name of an STDCALL function, and cleans up its own stack. It is therefore an STDCALL function.

3.5.5 Example: Unnamed Assembly Function

This code snippet is the entire body of an unnamed assembly function. Identify the calling convention of this function.

push ebp mov ebp, esp add eax, edx pop ebp ret

The function sets up a stack frame, so we know the compiler hasnt done anything "funny" to it. It accesses registers which arent initialized yet, in the edx and eax registers. It is therefore a FASTCALL function.

3.5.6 Example: Another Unnamed Assembly Function

push ebp mov ebp, esp mov eax, [ebp + 8] pop ebp ret 16

The function has a standard stack frame, and the *ret* instruction has a parameter to clean its own stack. Also, it accesses a parameter from the stack. It is therefore an STDCALL function.

3.5.7 Example: Name Mangling

What can we tell about the following function call?

mov ecx, x push eax mov eax, ss:[ebp - 4] push eax mov al, ss:[ebp - 3] call @__Load?$Container__XXXY_?Fcii

Two things should get our attention immediately. The first is that before the function call, a value is stored into ecx. Also, the function name itself is heavily mangled. This example must use the C++ THISCALL convention. Inside the mangled name of the function, we can pick out two english words, "Load" and "Container". Without knowing the specifics of this name mangling scheme, it is not possible to determine which word is the function name, and which word is the class name.

We can pick out two 32-bit variables being passed to the function, and a single 8-bit variable. The first is located in **eax**, the second is originally located on the stack from offset -4 from **ebp**, and the third is located at **ebp** offset -3. In C++, these would likely correspond to two **int** variables, and a single **char** variable. Notice at the end of the mangled function name are three lower-case characters "cii". We can't know for certain, but it appears these three letters correspond to the three parameters (char, int, int). We do not know from this whether the function returns a value or not, so we will assume the function returns **void**.

Assuming that "Load" is the function name and "Container" is the class name (it could just as easily be the other way around), here is our function definition:

class Container { void Load(char, int, int); }

3.6 x86 Disassembly/Branches

3.6.1 Branching

Computer science professors tell their students to avoid jumps and **goto** instructions, to avoid the proverbial "spaghetti code." Unfortunately, assembly only has jump instructions to control program flow. This chapter will explore the subject that many people avoid like the plague, and will attempt to show how the spaghetti of assembly can be translated into the more familiar control structures of high-level language. Specifically, this chapter will focus on **If-Then-Else** and **Switch** branching instructions.

3.6.2 If-Then

Let's consider a generic **if** statement in pseudo-code followed by its equivalent form using jumps:

What does this code do? In English, the code checks the condition and performs a jump only if it is *false*. With that in mind, let's compare some actual C code and its Assembly translation:

Note that when we translate to assembly, we need to *negate* the condition of the jump because--like we said

above--we only jump if the condition is false. To recreate the high-level code, simply negate the condition once again.

Negating a comparison may be tricky if you're not paying attention. Here are the correct dual forms:

And here are some examples.

mov eax, $x //move x into eax cmp eax, $y //compare eax with y jg end //jump if greater than inc eax move $x, eax //increment x end: ...

Is produced by these C statements:

if(x <= y) { x++; }

As you can see, x is incremented only if it is **less than or equal to** y. Thus, if it is greater than y, it will not be incremented as in the assembler code. Similarly, the C code

if(x < y) { x++; }

produces this assembler code:

mov eax, $x //move x into eax cmp eax, $y //compare eax with y jge end //jump if greater than or equal to inc eax move $x, eax //increment x end: ...

X is incremented in the C code only if it is **less than** y, so the assembler code now jumps if it's greater than or equal to y. This kind of thing takes practice, so we will try to include lots of examples in this section.

3.6.3 If-Then-Else

Let us now look at a more complicated case: the **If-Then-Else** instruction.

Now, what happens here? Like before, the if statement only jumps to the else clause when the condition is false. However, we must also install an *unconditional* jump at the end of the "then" clause, so we don't perform the else clause directly afterwards.

Now, here is an example of a real C If-Then-Else:

if(x == 10) { x = 0; } else { x++; }

Which gets translated into the following assembly code:

mov eax, $x cmp eax, 0x0A ;0x0A = 10 jne else mov eax, 0 jmp end else: inc eax end: mov $x, eax

As you can see, the addition of a single unconditional jump can add an entire extra option to our conditional.

3.6.4 Switch-Case

Switch-Case structures can be very complicated when viewed in assembly language, so we will examine a few examples. First, keep in mind that in C, there are several keywords that are commonly used in a switch statement. Here is a recap:

Switch This keyword tests the argument, and starts the switch structure

Case This creates a label that execution will switch to, depending on the value of the argument.

Break This statement jumps to the end of the switch block

Default This is the label that execution jumps to if and only if it doesn't match up to any other conditions

Lets say we have a general switch statement, but with an extra label at the end, as such:

switch (x) { //body of switch statement } end_of_switch:

Now, every **break** statement will be immediately replaced with the statement

jmp end_of_switch

But what do the rest of the statements get changed to? The case statements can each resolve to any number of arbitrary integer values. How do we test for that? The answer is that we use a "Switch Table". Here is a simple, C example:

int main(int argc, char **argv) { //line 10 switch(argc) { case 1: MyFunction(1); break; case 2: MyFunction(2); break; case 3: MyFunction(3); break; case 4: MyFunction(4); break; default: MyFunction(5); } return 0; }

And when we compile this with **cl.exe**, we can generate the following listing file:

tv64 = −4 ; size = 4 _argc$ = 8 ; size = 4 _argv$ = 12 ; size = 4 _main PROC NEAR ; Line 10 push ebp mov ebp, esp push ecx ; Line 11 mov eax, DWORD PTR _argc$[ebp] mov DWORD PTR tv64[ebp], eax mov ecx, DWORD PTR tv64[ebp] sub ecx, 1 mov DWORD PTR tv64[ebp], ecx cmp DWORD PTR tv64[ebp], 3 ja SHORT $L810 mov edx, DWORD PTR tv64[ebp] jmp DWORD PTR $L818[edx*4] $L806: ; Line 14 push 1 call _MyFunction add esp, 4 ; Line 15 jmp SHORT $L803 $L807: ; Line 17 push 2 call _MyFunction add esp, 4 ; Line 18 jmp SHORT $L803 $L808: ; Line 19 push 3 call _MyFunction add esp, 4 ; Line 20 jmp SHORT $L803 $L809: ; Line 22 push 4 call _MyFunction add esp, 4 ; Line 23 jmp SHORT $L803 $L810: ; Line 25 push 5 call _MyFunction add esp, 4 $L803: ; Line 27 xor eax, eax ; Line 28 mov esp, ebp

pop ebp ret 0 $L818: DD $L806 DD $L807 DD $L808 DD $L809 _main ENDP

Lets work our way through this. First, we see that line 10 sets up our standard stack frame, and it also saves ecx. Why does it save ecx? Scanning through the function, we never see a corresponding "pop ecx" instruction, so it seems that the value is never restored at all. In fact, the compiler isn't saving ecx at all, but is instead simply reserving space on the stack: it's creating a local variable. The original C code didn't have any local variables, however, so perhaps the compiler just needed some extra scratch space to store intermediate values. Why doesn't the compiler execute the more familiar "sub esp, 4" command to create the local variable? **push ecx** is just a faster instruction that does the same thing. This "scratch space" is being referenced by a *negative offset* from ebp. **tv64** was defined in the beginning of the listing as having the value -4, so every call to "tv64[ebp]" is a call to this scratch space.

There are a few things that we need to notice about the function in general:

- Label $L803 is the end_of_switch label. Therefore, every "jmp SHORT $L803" statement is a **break**. This is verifiable by comparing with the C code line-by-line.
- Label $L818 contains a list of hard-coded memory addresses, which here are labels in the code section! Remember, labels resolve to the memory address of the instruction. This must be an important part of our puzzle.

To solve this puzzle, we will take an in-depth look at line 11:

mov eax, DWORD PTR _argc$[ebp] mov DWORD PTR tv64[ebp], eax mov ecx, DWORD PTR tv64[ebp] sub ecx, 1 mov DWORD PTR tv64[ebp], ecx cmp DWORD PTR tv64[ebp], 3 ja SHORT $L810 mov edx, DWORD PTR tv64[ebp] jmp DWORD PTR $L818[edx*4]

This sequence performs the following pseudo-C operation:

if(argc - 1 >= 4) { goto $L810; /* the default */ } label *L818[] = { $L806, $L807, $L808, $L809 }; /* define a table of jumps, one per each case */ // goto L818[argc - 1]; /* use the address from the table to jump to the correct case */

Here's why...

The Setup

mov eax, DWORD PTR _argc$[ebp] mov DWORD PTR tv64[ebp], eax mov ecx, DWORD PTR tv64[ebp] sub ecx, 1 mov DWORD PTR tv64[ebp], ecx

The value of argc is moved into eax. The value of eax is then immediately moved to the scratch space. The value of the scratch space is then moved into ecx. Sounds like an awfully convoluted way to get the same value into so many different locations, but remember: I turned off the optimizations. The value of ecx is then decremented by 1. Why didn't the compiler use a **dec** instruction instead? Perhaps the statement is a general statement, that in this case just happens to have an argument of 1. We don't know why exactly, all we know is this:

- **eax = "scratch pad"**
- **ecx = eax - 1**

Finally, the last line moves the new, decremented value of ecx *back into the scratch pad*. Very inefficient.

The Compare and Jumps

cmp DWORD PTR tv64[ebp], 3 ja SHORT $L810

The value of the scratch pad is compared with the value 3, and if the *unsigned* value is above 3 (4 or more), execution jumps to label $L810. How do I know the value is unsigned? I know because **ja** is an unsigned conditional jump. Let's look back at the original C code switch:

switch(argc) { case 1: MyFunction(1); break; case 2: MyFunction(2); break; case 3: MyFunction(3); break; case 4: MyFunction(4); break; default: MyFunction(5); }

Remember, the scratch pad contains the value (argc - 1), which means that this condition is only triggered when argc > 4. What happens when argc is greater than 4? The function goes to the default condition. Now, let's look at the next two lines:

mov edx, DWORD PTR tv64[ebp] jmp DWORD PTR $L818[edx*4]

edx gets the value of the scratch pad (argc - 1), and then there is a very weird jump that takes place: execution jumps to a location pointed to by the value (edx * 4 + $L818). What is $L818? We will examine that right now.

The Switch Table

$L818: DD $L806 DD $L807 DD $L808 DD $L809

$L818 is a pointer, in the code section, to a list of other code section pointers. These pointers are all 32bit values (DD is a DWORD). Let's look back at our jump statement:

jmp DWORD PTR $L818[edx*4]

In this jump, $L818 *isn't the offset, it's the base*, edx*4 is the offset. As we said earlier, edx contains the value of (argc - 1). If argc == 1, we jump to [$L818 + 0] which is $L806. If argc == 2, we jump to [$L818 + 4], which is $L807. Get the picture? A quick look at labels $L806, $L807, $L808, and $L809 shows us exactly what we expect to see: the bodies of the **case** statements from the original C code, above. Each one of the case statements calls the function "MyFunction", then breaks, and then jumps to the end of the switch block.

3.6.5 Ternary Operator ?:

Again, the best way to learn is by doing. Therefore we will go through a mini example to explain the ternary operator. Consider the following C code program:

int main(int argc, char **argv) { return (argc > 1)?(5):(0); }

cl.exe produces the following assembly listing file:

_argc$ = 8 ; size = 4 _argv$ = 12 ; size = 4 _main PROC NEAR ; File c:\documents and settings\andrew\desktop\test2.c ; Line 2 push ebp mov ebp, esp ; Line 3 xor eax, eax cmp DWORD PTR _argc$[ebp], 1 setle al dec eax and eax, 5 ; Line 4 pop ebp ret 0 _main ENDP

Line 2 sets up a stack frame, and line 4 is a standard exit sequence. There are no local variables. It is clear that Line 3 is where we want to look.

The instruction "xor eax, eax" simply sets eax to 0. For more information on that line, see the chapter on unintuitive instructions. The **cmp** instruction tests the condition of the ternary operator. The **setle** function is one of a set of x86 functions that works like a conditional move: al gets the value 1 if argc <= 1. Isn't that the exact opposite of what we wanted? In this case, it is. Let's look at what happens when argc = 0: **al** gets the value 1. **al** is decremented (al = 0), and then eax is logically anded with 5. 5 & 0 = 0. When argc == 2 (greater than 1), the **setle** instruction doesn't do anything, and eax still is zero. eax is then decremented, which means that eax == −1. What is −1?

In x86 processors, negative numbers are stored in **two's-complement** format. For instance, let's look at the following C code:

BYTE x; x = −1;

At the end of this C code, **x** will have the value 11111111: all ones!

When argc is greater than 1, setle sets al to zero. Decrementing this value sets every bit in eax to a logical 1. Now, when we perform the logical **and** function we get:

...11111111 &...00000101 ;101 is 5 in binary ---------- ...00000101

eax gets the value 5. In this case, it's a roundabout method of doing it, but as a reverser, this is the stuff you need to worry about.

For reference, here is the GCC assembly output of the same ternary operator from above:

_main: pushl %ebp movl %esp, %ebp subl $8, %esp xorl %eax, %eax andl $−16, %esp call __alloca call ___main xorl %edx, %edx cmpl $2, 8(%ebp) setge %dl leal (%edx,%edx,4), %eax leave ret

Notice that GCC produces slightly different code than cl.exe produces. However, the stack frame is set up the same way. Notice also that GCC doesn't give us line numbers, or other hints in the code. The ternary operator line occurs after the instruction "call __main". Let's highlight that section here:

xorl %edx, %edx cmpl $2, 8(%ebp) setge %dl leal (%edx,%edx,4), %eax

Again, **xor** is used to set edx to 0 quickly. Argc is tested against 2 (instead of 1), and dl is set if argc is *greater then or equal*. If dl gets set to 1, the **leal** instruction directly thereafter will move the value of 5 into eax (because lea (edx,edx,4) means edx + edx * 4, i.e. edx * 5).

3.7 x86 Disassembly/Branch Examples

3.7.1 Example: Number of Parameters

What parameters does this function take? What calling convention does it use? What kind of value does it return? Write the entire C prototype of this function. Assume all values are unsigned values.

push ebp mov ebp, esp mov eax, 0 mov ecx, [ebp + 8] cmp ecx, 0 jne _Label_1 inc eax jmp _Label_2 :_Label_1 dec eax : _Label_2 mov ecx, [ebp + 12] cmp ecx, 0 jne _Label_3 inc eax : _Label_3 mov esp, ebp pop ebp ret

This function accesses parameters on the stack at [ebp + 8] and [ebp + 12]. Both of these values are loaded into ecx, and we can therefore assume they are 4-byte values. This function doesn't clean its own stack, and the values aren't passed in registers, so we know the function is CDECL. The return value in eax is a 4-byte value, and we are told to assume that all the values are unsigned.

Putting all this together, we can construct the function prototype:

unsigned int CDECL MyFunction(unsigned int param1, unsigned int param2);

3.7.2 Example: Identify Branch Structures

How many separate branch structures are in this function? What types are they? Can you give more descriptive names to _Label_1, _Label_2, and _Label_3, based on the structures of these branches?

push ebp mov ebp, esp mov eax, 0 mov ecx, [ebp + 8] cmp ecx, 0 jne _Label_1 inc eax jmp _Label_2 :_Label_1 dec eax : _Label_2 mov ecx, [ebp + 12] cmp ecx, 0 jne _Label_3 inc eax : _Label_3 mov esp, ebp pop ebp ret

How many separate branch structures are there in this function? Stripping away the entry and exit sequences, here is the code we have left:

mov ecx, [ebp + 8] cmp ecx, 0 jne _Label_1 inc eax jmp _Label_2 :_Label_1 dec eax : _Label_2 mov ecx, [ebp + 12] cmp ecx, 0 jne _Label_3 inc eax : _Label_3

Looking through, we see 2 **cmp** statements. The first cmp statement compares ecx to zero. If ecx is not zero, we go to _Label_1, decrement eax, and then fall through to _Label_2. If ecx is zero, we increment eax, and go to directly to _Label_2. Writing out some pseudocode, we have the following result for the first section:

if(ecx doesnt equal 0) goto _Label_1 eax++; goto _Label_2 :_Label_1 eax--; :_Label_2

Since _Label_2 occurs at the end of this structure, we can rename it to something more descriptive, like "End_of_Branch_1", or "Branch_1_End". The first comparison tests ecx against 0, and then jumps on not-equal. We can reverse the conditional, and say that _Label_1 is an **else** block:

if(ecx == 0) ;ecx is param1 here { eax++; } else { eax--; }

So we can rename _Label_1 to something else descriptive, such as "Else_1". The rest of the code block, after Branch_1_End (_Label_2) is as follows:

mov ecx, [ebp + 12] cmp ecx, 0 jne _Label_3 inc eax : _Label_3

We can see immediately that _Label_3 is the end of this branch structure, so we can immediately call it "Branch_2_End", or something else. Here, we are again comparing ecx to 0, and if it is not equal, we jump to the end of the block. If it is equal to zero, however, we increment eax, and then fall out the bottom of the branch. We can see that there is no **else** block in this branch structure, so we don't need to invert the condition. We can write an **if** statement directly:

if(ecx == 0) ;ecx is param2 here { eax++; }

3.7.3 Example: Convert To C

Write the equivalent C code for this function. Assume all parameters and return values are unsigned values.

push ebp mov ebp, esp mov eax, 0 mov ecx, [ebp + 8] cmp ecx, 0 jne _Label_1 inc eax jne _Label_2 :_Label_1 dec eax : _Label_2 mov ecx, [ebp + 12] cmp ecx, 0 jne _Label_3 inc eax : _Label_3 mov esp, ebp pop ebp ret

Starting with the C function prototype from answer 1, and the conditional blocks in answer 2, we can put together a pseudo-code function, without variable declarations, or a return value:

unsigned int CDECL MyFunction(unsigned int param1, unsigned int param2) { if(param1 == 0) { eax++; } else { eax--; } if(param2 == 0) { eax++; } }

Now, we just need to create a variable to store the value from eax, which we will call "a", and we will declare as a **register** type:

unsigned int CDECL MyFunction(unsigned int param1, unsigned int param2) { register unsigned int a = 0; if(param1 == 0) { a++; } else { a--; } if(param2 == 0) { a++; } return a; }

Granted, this function isn't a particularly useful function, but at least we know what it does.

3.8 x86 Disassembly/Loops

3.8.1 Loops

To complete repetitive tasks, programmers often implement **loops**. There are many sorts of loops, but they can all be boiled down to a few similar formats in assembly code. This chapter will discuss loops, how to identify them, and how to "decompile" them back into high-level representations.

3.8.2 Do-While Loops

It seems counterintuitive that this section will consider **Do-While** loops first, considering that they might be the

least used of all the variations in practice. However, there is method to our madness, so read on.

Consider the following generic Do-While loop:

What does this loop do? The loop body simply executes, the condition is tested at the end of the loop, and the loop jumps back to the beginning of the loop if the condition is satisfied. Unlike **if** statements, Do-While conditions are not reversed.

Let us now take a look at the following C code:

do { x++; } while(x != 10);

Which can be translated into assembly language as such:

mov eax, $x beginning: inc eax cmp eax, 0x0A ;0x0A = 10 jne beginning mov $x, eax

3.8.3 While Loops

While loops look almost as simple as a **Do-While** loop, but in reality they aren't as simple at all. Let's examine a generic while-loop:

while(x) { //loop body }

What does this loop do? First, the loop checks to make sure that x is true. If x is not true, the loop is skipped. The loop body is then executed, followed by another check: is x still true? If x is still true, execution jumps back to the top of the loop, and execution continues. Keep in mind that there needs to be a jump at the bottom of the loop (to get back up to the top), but it makes no sense to jump back to the top, retest the conditional, and then jump *back to the bottom of the loop* if the conditional is found to be false. The while-loop then, performs the following steps:

1. check the condition. if it is false, go to the end
2. perform the loop body
3. check the condition, if it is true, jump to 2.
4. if the condition is not true, fall-through the end of the loop.

Here is a while-loop in C code:

while(x <= 10) { x++; }

And here then is that same loop translated into assembly:

mov eax, $x cmp eax, 0x0A jg end beginning: inc eax cmp eax, 0x0A jle beginning end:

If we were to translate that assembly code **back into C**, we would get the following code:

if(x <= 10) //remember: in If statements, we reverse the condition from the asm { do { x++; } while(x <= 10) }

See why we covered the Do-While loop first? Because the While-loop becomes a Do-While when it gets assembled.

So why can't the jump label occur before the test?

mov eax, $x beginning: cmp eax, 0x0A jg end inc eax jmp begnning end: mov $x, eax

3.8.4 For Loops

What is a For-Loop? In essence, it's a While-Loop with an initial state, a condition, and an iterative instruction. For instance, the following generic For-Loop:

gets translated into the following pseudocode while-loop:

initialization; while(condition) { action; increment; }

Which in turn gets translated into the following Do-While Loop:

initialization; if(condition) { do { action; increment; } while(condition); }

Note that often in for() loops you assign an initial constant value in A (for example x = 0), and then compare that value with another constant in B (for example x < 10). Most optimizing compilers will be able to notice that the first time x IS less than 10, and therefore there is no need for the initial if(B) statement. In such cases, the compiler will simply generate the following sequence:

initialization; do { action increment; } while(condition);

rendering the code indistinguishable from a while() loop.

3.8.5 Other Loop Types

C only has Do-While, While, and For Loops, but some other languages may very well implement their own types. Also, a good C-Programmer could easily "home brew" a new type of loop using a series of good macros, so they bear some consideration:

Do-Until Loop

A common Do-Until Loop will take the following form:

do { //loop body } until(x);

which essentially becomes the following Do-While loop:

do { //loop body } while(!x);

Until Loop

Like the Do-Until loop, the standard Until-Loop looks like the following:

until(x) { //loop body }

which (likewise) gets translated to the following While-Loop:

while(!x) { //loop body }

Do-Forever Loop

A Do-Forever loop is simply an unqualified loop with a condition that is always true. For instance, the following pseudo-code:

doforever { //loop body }

will become the following while-loop:

while(1) { //loop body }

Which can actually be reduced to a simple unconditional jump statement:

beginning: ;loop body jmp beginning

Notice that some non-optimizing compilers will produce nonsensical code for this:

mov ax, 1 cmp ax, 1 jne loopend beginning: ;loop body cmp ax, 1 je beginning loopend:

Notice that a lot of the comparisons here are not needed since the condition is a constant. Most compilers will optimize cases like this.

3.9 x86 Disassembly/Loop Examples

3.9.1 Example: Identify Purpose

What does this function do? What kinds of parameters does it take, and what kind of results (if any) does it return?

push ebp mov ebp, esp mov esi, [ebp + 8] mov ebx, 0 mov eax, 0 mov ecx, 0 _Label_1: mov ecx, [esi + ebx * 4] add eax, ecx inc ebx cmp ebx, 100 jne _Label_1 mov esp, ebp pop ebp ret 4

This function loops through an array of 4 byte integer values, pointed to by esi, and adds each entry. It returns the sum in eax. The only parameter (located in [ebp + 8]) is a pointer to an array of integer values. The comparison between ebx and 100 indicates that the input array has 100 entries in it. The pointer offset [esi + ebx * 4] shows that each entry in the array is 4 bytes wide.

3.9.2 Example: Complete C Prototype

What is this function's C prototype? Make sure to include parameters, return values, and calling convention.

push ebp mov ebp, esp mov esi, [ebp + 8] mov ebx, 0 mov eax, 0 mov ecx, 0 _Label_1: mov ecx, [esi + ebx * 4] add eax, ecx inc ebx cmp ebx, 100 jne _Label_1 mov esp, ebp pop ebp ret 4

Notice how the **ret** function cleans its parameter off the stack? That means that this function is an STDCALL function. We know that the function takes, as its only parameter, a pointer to an array of integers. We do not know, however, whether the integers are signed or unsigned, because the **je** command is used for both types of values. We can assume one or the other, and for simplicity, we can assume unsigned values (unsigned and signed values, in this function, will actually work the same way). We also know that the return value is a 4-byte integer value, of the same type as is found in the parameter array. Since the function doesnt have a name, we can just call it "MyFunction", and we can call the parameter "array" because it is an array. From this information, we can determine the following prototype in C:

unsigned int STDCALL MyFunction(unsigned int *array);

3.9.3 Example: Decompile To C Code

Decompile this code into equivalent C source code.

push ebp mov ebp, esp mov esi, [ebp + 8] mov ebx, 0 mov eax, 0 mov ecx, 0 _Label_1: mov ecx, [esi + ebx * 4] add eax, ecx inc ebx cmp ebx, 100 jne _Label_1 mov esp, ebp pop ebp ret 4

Starting with the function prototype above, and the description of what this function does, we can start to write the C code for this function. We know that this function initializes eax, ebx, and ecx before the loop. However, we can see that ecx is being used as simply an intermediate storage location, receiving successive values from the array, and then being added to eax.

We will create two unsigned integer values, a (for eax) and b (for ebx). We will define both a and b with the **register** qualifier, so that we can instruct the compiler

not to create space for them on the stack. For each loop iteration, we are adding the value of the array, at location ebx*4 to the running sum, eax. Converting this to our a and b variables, and using C syntax, we see:

```
a = a + array[b];
```

The loop could be either a **for** loop, or a **while** loop. We see that the loop control variable, b, is initialized to 0 before the loop, and is incremented by 1 each loop iteration. The loop tests b against 100, *after it gets incremented*, so we know that b never equals 100 inside the loop body. Using these simple facts, we will write the loop in 3 different ways:

First, with a **while** loop.

```
unsigned int STDCALL MyFunction(unsigned int *array) { register unsigned int b = 0; register unsigned int a = 0; while(b != 100) { a = a + array[b]; b++; } return a; }
```

Or, with a **for** loop:

```
unsigned int STDCALL MyFunction(unsigned int *array) { register unsigned int b; register unsigned int a = 0; for(b = 0; b != 100; b++) { a = a + array[b]; } return a; }
```

And finally, with a **do-while** loop:

```
unsigned int STDCALL MyFunction(unsigned int *array) { register unsigned int b = 0; register unsigned int a = 0; do { a = a + array[b]; b++; }while(b != 100); return a; }
```

Chapter 4

Data Patterns

4.1 x86 Disassembly/Variables

4.1.1 Variables

We've already seen some mechanisms to create local storage on the stack. This chapter will talk about some other variables, including **global variables**, **static variables**, variables labled "**const**," "**register**," and "**volatile**." It will also consider some general techniques concerning variables, including accessor and setter methods (to borrow from OO terminology). This section may also talk about setting memory breakpoints in a debugger to track memory I/O on a variable.

4.1.2 How to Spot a Variable

Variables come in 2 distinct flavors: those that are created on the stack (local variables), and those that are accessed via a hardcoded memory address (global variables). Any memory that is accessed via a hard-coded address is usually a global variable. Variables that are accessed as an offset from esp, or ebp are frequently local variables.

Hardcoded address Anything hardcoded is a value that is stored as-is in the binary, and is not changed at runtime. For instance, the value 0x2054 is hard-coded, whereas the current value of variable X is not hard-coded and may change at runtime.

Example of a hardcoded address:

mov eax, [0x77651010]

OR:

mov ecx, 0x77651010 mov eax, [ecx]

Example of a non-hardcoded (softcoded?) address:

mov ecx, [esp + 4] add ecx, ebx mov eax, [ecx]

In the last example, the value of ecx is calculated at runtime, whereas in the first 2 examples, the value is the same every time. RVAs are considered hard-coded addresses, even though the loader needs to "fix them up" to point to the correct locations.

4.1.3 .BSS and .DATA sections

Both .bss and .data sections contain values which can change at run-time (e.g. *variables*). Typically, variables that are initialized to a non-zero value in the source are allocated in the .data section (e.g. "int a = 10;"). Variables that are not initialized, or initialized with a zero value, can be allocated to the .bss section (e.g. "int arr[100];"). Because all values of .bss variables are guaranteed to be zero at the start of the program, there is no need for the linker to allocate space in the binary file. Therefore, .bss sections do not take space in the binary file, regardless of their size.

4.1.4 "Static" Local Variables

Local variables labeled **static** maintain their value across function calls, and therefore cannot be created on the stack like other local variables are. How are static variables created? Let's take a simple example C function:

void MyFunction(int a) { static int x = 0; printf("my number: "); printf("%d, %d\n", a, x); }

Compiling to a listing file with **cl.exe** gives us the following code:

_BSS SEGMENT ?x@?1??MyFunction@@9@9 DD 01H DUP (?) ; `MyFunction'::`2'::x _BSS ENDS _DATA SEGMENT $SG796 DB 'my number: ', 00H $SG797 DB '%d, %d', 0aH, 00H _DATA ENDS PUBLIC _MyFunction EXTRN _printf:NEAR ; Function compile flags: /Odt _TEXT SEGMENT _a$ = 8 ; size = 4 _MyFunction PROC NEAR ; Line 4 push ebp mov ebp, esp ; Line 6 push OFFSET FLAT:$SG796 call _printf add esp, 4 ; Line 7 mov eax, DWORD PTR ?x@?1??MyFunction@@9@9 push eax mov ecx, DWORD PTR _a$[ebp] push ecx push OFFSET

FLAT:$SG797 call _printf add esp, 12 ; 0000000cH ; Line 8 pop ebp ret 0 _MyFunction ENDP _TEXT ENDS

Normally when assembly listings are posted in this wikibook, most of the code gibberish is discarded to aid readability, but in this instance, the "gibberish" contains the answer we are looking for. As can be clearly seen, this function creates a standard stack frame, and it doesn't create any local variables on the stack. In the interests of being complete, we will take baby-steps here, and work to the conclusion logically.

In the code for Line 7, there is a call to _printf with 3 arguments. Printf is a standard **libc** function, and it therefore can be assumed to be cdecl calling convention. Arguments are pushed, therefore, from right to left. Three arguments are pushed onto the stack before _printf is called:

- DWORD PTR ?x@?1??MyFunction@@9@9
- DWORD PTR _a$[ebp]
- OFFSET FLAT:$SG797

The second one, _a$[ebp] is partially defined in this assembly instruction:

_a$ = 8

And therefore _a$[ebp] is the variable located at offset +8 from ebp, or the first argument to the function. OFFSET FLAT:$SG797 likewise is declared in the assembly listing as such:

SG797 DB '%d, %d', 0aH, 00H

If you have your ASCII table handy, you will notice that 0aH = 0x0A = '\n'. OFFSET FLAT:$SG797 then is the format string to our printf statement. Our last option then is the mysterious-looking "?x@?1??MyFunction@@9@9", which is defined in the following assembly code section:

_BSS SEGMENT ?x@?1??MyFunction@@9@9 DD 01H DUP (?) _BSS ENDS

This shows that the Microsoft C compiler creates static variables in the .bss section. This might not be the same for all compilers, but the lesson is the same: local static variables are created and used in a very similar, if not the exact same, manner as global values. In fact, as far as the reverser is concerned, the two are usually interchangeable. Remember, the only real difference between static variables and global variables is the idea of "scope", which is only used by the compiler.

4.1.5 Signed and Unsigned Variables

Integer formatted variables, such as **int**, **char**, **short** and **long** may be declared signed or unsigned variables in the C source code. There are two differences in how these variables are treated:

1. Signed variables use signed instructions such as **add**, and **sub**. Unsigned variables use unsigned arithmetic instructions such as **addi**, and **subi**.
2. Signed variables use signed branch instructions such as **jge** and **jl**. Unsigned variables use unsigned branch instructions such as **jae**, and **jb**.

The difference between signed and unsigned instructions is the conditions under which the various flags for greater-then or less-then (overflow flags) are set. The integer result values are exactly the same for both signed and unsigned data.

4.1.6 Floating-Point Values

Floating point values tend to be 32-bit data values (for **float**) or 64-bit data values (for **double**). These values are distinguished from ordinary integer-valued variables because they are used with floating-point instructions. Floating point instructions typically start with the letter *f*. For instance, **fadd**, **fcmp**, and similar instructions are used with floating point values. Of particular note are the **fload** instruction and variants. These instructions take an integer-valued variable and converts it into a floating point variable.

We will discuss floating point variables in more detail in a later chapter.

4.1.7 Global Variables

Global variables do not have a limited scope like lexical variables do inside a function body. Since the notion of lexical scope implies the use of the system stack, and since global variables are not lexical in nature, they are typically not found on the stack. Global variables tend to exist in the program as a hard-coded memory address, a location which never changes throughout program execution. These could exist in the DATA segment of the executable, or anywhere else that a hard-coded memory address can be used to store data.

In C, global variables are defined outside the body of any function. There is no "global" keyword. Any variable which is not defined inside a function is global. In C however, a variable which is not defined inside a function is only global to the particular source code file in which it is defined. For example, we have two files Foo.c and Bar.c, and a global variable MyGlobalVar:

In the example above, the variable MyGlobalVar is visible inside the file Foo.c, but is not visible inside the file Bar.c. To make MyGlobalVar visible inside all project files, we need to use the extern keyword, which we will discuss below.

"static" Variables

The C programming language specifies a special keyword "static" to define variables which are lexical to the function (they cannot be referenced from outside the function) but they maintain their values across function calls. Unlike ordinary lexical variables which are created on the stack when the function is entered and are destroyed from the stack when the function returns, static variables are created once and are never destroyed.

int MyFunction(void) { static int x; ... }

Static variables in C are global variables, except the compiler takes precautions to prevent the variable from being accessed outside of the parent function's scope. Like global variables, static variables are referenced using a hardcoded memory address, not a location on the stack like ordinary variables. However unlike globals, static variables are only used inside a single function. There is no difference between a global variable which is only used in a single function, and a static variable inside that same function. However, it's good programming practice to limit the number of global variables, so when disassembling, you should prefer interpreting these variables as static instead of global.

"extern" Variables

The extern keyword is used by a C compiler to indicate that a particular variable is global to the entire project, not just to a single source code file. Besides this distinction, and the slightly larger lexical scope of extern variables, they should be treated like ordinary global variables.

In static libraries, variables marked as being extern might be available for use with programs which are linked to the library.

Global Variables Summary

Here is a table to summarize some points about global variables:

When disassembling, a hard-coded memory address should be considered to be an ordinary global variable unless you can determine from the scope of the variable that it is static or extern.

4.1.8 Constants

Variables qualified with the **const** keyword (in C) are frequently stored in the .data section of the executable. Constant values can be distinguished because they are initialized at the beginning of the program, and are never modified by the program itself. For this reasons, some compilers may chose to store constant variables (especially strings) in the .text section of the executable, thus allowing the sharing of these variables across multiple instances of the same process. This creates a big problem for the reverser, who now has to decide whether the code he's looking at is part of a constant variable or part of a subroutine.

4.1.9 "Volatile" memory

In C and C++, variables can be declared "volatile," which tells the compiler that the memory location can be accessed from *external* or *concurrent* processes, and that the compiler should not perform any optimizations on the variable. For instance, if multiple threads were all accessing and modifying a single global value, it would be bad for the compiler to store that variable in a register sometimes, and flush it to memory infrequently. In general, Volatile memory must be flushed to memory after every calculation, to ensure that the most current version of the data is in memory when other processes come to look for it.

It is not always possible to determine from a disassembly listing whether a given variable is a volatile variable. However, if the variable is accessed frequently from memory, and its value is constantly updated in memory (especially if there are free registers available), that's a good hint that the variable might be volatile.

4.1.10 Simple Accessor Methods

An Accessor Method is a tool derived from OO theory and practice. In it's most simple form, an accessor method is a function that receives no parameters (or perhaps simply an offset), and returns the value of a variable. Accessor and Setter methods are ways to restrict access to certain variables. The only standard way to get the value of the variable is to use the Accessor.

Accessors can prevent some simple problems, such as out-of-bounds array indexing, and using uninitialized data. Frequently, Accessors contain little or no error-checking.

Here is an example:

push ebp mov ebp, esp mov eax, [ecx + 8] ;THISCALL function, passes "this" pointer in ecx mov esp, ebp pop ebp ret

Because they are so simple, accessor methods are frequently heavily optimized (they generally don't need a stack frame), and are even occasionally *inlined* by the compiler.

4.1.11 Simple Setter (Manipulator) Methods

Setter methods are the antithesis of an accessor method, and provide a unified way of altering the value of a given variable. Setter methods will often take as a parameter the value to be set to the variable, although some methods (Initializers) simply set the variable to a pre-defined value. Setter methods often do bounds checking, and error checking on the variable before it is set, and frequently either a) return no value, or b) return a simple boolean value to determine success.

Here is an example:

push ebp mov ebp, esp cmp [ebp + 8], 0 je error mov eax, [ebp + 8] mov [ecx + 0], eax mov eax, 1 jmp end :error mov eax, 0 :end mov esp, ebp pop ebp ret

4.2 x86 Disassembly/Variable Examples

4.2.1 Example: Identify C++ Code

Can you tell what the original C++ source code looks like, in general, for the following accessor method?

push ebp mov ebp, esp mov eax, [ecx + 8] ;THISCALL function, passes “this” pointer in ecx mov esp, ebp pop ebp ret

We don't know the name of the class, so we will use a generic name MyClass (or whatever you would like to call it). We will lay out a simple class definition, that contains a data value at offset +8. Offset +8 is the only data value accessed, so we don't know what the first 8 bytes of data looks like, but we will just assume (for our purposes) that our class looks like this:

class MyClass { int value1; int value2; int value3; //offset +8 ... }

We will then create our function, which I will call “GetValue3()". We know that the data value being accessed is located at [ecx+8], (which we have defined above to be “value3”). Also, we know that the data is being read into a 4-byte register (eax), and is not truncated. We can assume, therefore, that value3 is a 4-byte data value. We can use the **this** pointer as the pointer value stored in ecx, and we can take the element that is at offset +8 from that pointer (value3):

MyClass::GetValue3() { return this->value3; }

The **this** pointer is not necessary here, but i use it anyway to illustrate the fact that the variable was accessed as an offset from the **this** pointer.

Note: Remember, we don't know what the first 8 bytes actually look like in our class, we only have a single accessor method, that only accesses a single data value at offset +8. The class could also have looked like this:

class MyClass /*Alternate Definition*/ { byte byte1; byte byte2; short short1; long value2; long value3; ... }

Or, any other combinations of 8 bytes.

4.2.2 Example: Identify C++ Code

Can you tell what the original C++ source code looks like, in general, for the following setter method?

push ebp mov ebp, esp cmp [ebp + 8], 0 je error mov eax, [ebp + 8] mov [ecx + 0], eax mov eax, 1 jmp end :error mov eax, 0 :end mov esp, ebp pop ebp ret

This code looks a little complicated, but don't panic! We will walk through it slowly. The first two lines of code set up the stack frame:

push ebp mov ebp, esp

The next two lines of code compare the value of [ebp + 8] (which we know to be the first parameter) to zero. If [ebp+8] is zero, the function jumps to the label “error”. We see that the label “error” sets eax to 0, and returns. We haven't seen it before, but this looks conspicuously like an **if** statement. “If the parameter is zero, return zero”.

If, on the other hand, the parameter is not zero, we move the value into eax, and then move the value into [ecx + 0], which we know as the first data field in MyClass. We also see, from this code, that this first data field must be 4 bytes long (because we are using eax). After we move eax into [ecx + 0], we set eax to 1 and jump to the end of the function.

If we use the same MyClass defintion as in question 1, above, we can get the following code for our function, “SetValue1(int val)":

int MyClass::SetValue1(int val) { if(val == 0) return 0; this->value1 = val; return 1; }

Notice that since we are returning a 0 on failure, and a 1 on success, the function looks like it has a **bool** return value. However, the return value is 4 bytes wide (eax is

used), but the size of a **bool** is implementation-specific, so we can't be sure. The **bool** is usually defined to have a size of 1 byte, but it is often stored the same way as an **int**.

4.3 x86 Disassembly/Data Structures

4.3.1 Data Structures

Few programs can work by using simple memory storage; most need to utilize complex data objects, including **pointers**, **arrays**, **structures**, and other complicated types. This chapter will talk about how compilers implement complex data objects, and how the reverser can identify these objects.

4.3.2 Arrays

Arrays are simply a storage scheme for multiple data objects of the same type. Data objects are stored sequentially, often as an offset from a pointer to the beginning of the array. Consider the following C code:

x = array[25];

Which is identical to the following asm code:

mov ebx, $array mov eax, [ebx + 25] mov $x, eax

Now, consider the following example:

int MyFunction1() { int array[20]; ...

This (roughly) translates into the following asm pseudo-code:

:_MyFunction1 push ebp mov ebp, esp sub esp, 80 ;the whole array is created on the stack!!! lea $array, [esp + 0] ;a pointer to the array is saved in the array variable ...

The entire array is created on the stack, and the pointer to the bottom of the array is stored in the variable "array". An optimizing compiler could ignore the last instruction, and simply refer to the array via a +0 offset from esp (in this example), but we will do things verbosely.

Likewise, consider the following example:

void MyFunction2() { char buffer[4]; ...

This will translate into the following asm pseudo-code:

:_MyFunction2 push ebp mov ebp, esp sub esp, 4 lea $buffer, [esp + 0] ...

Which looks harmless enough. But, what if a program inadvertantly accesses buffer[4]? what about buffer[5]? what about buffer[8]? This is the makings of a buffer overflow vulnerability, and (might) will be discussed in a later section. However, this section won't talk about security issues, and instead will focus only on data structures.

Spotting an Array on the Stack

To spot an array on the stack, look for large amounts of local storage allocated on the stack ("sub esp, 1000", for example), and look for large portions of that data being accessed by an offset from a different register from esp. For instance:

:_MyFunction3 push ebp mov ebp, esp sub esp, 256 lea ebx, [esp + 0x00] mov [ebx + 0], 0x00

is a good sign of an array being created on the stack. Granted, an optimizing compiler might just want to offset from esp instead, so you will need to be careful.

Spotting an Array in Memory

Arrays in memory, such as global arrays, or arrays which have initial data (remember, initialized data is created in the .data section in memory) and will be accessed as offsets from a hardcoded address in memory:

:_MyFunction4 push ebp mov ebp, esp mov esi, 0x77651004 mov ebx, 0x00000000 mov [esi + ebx], 0x00

It needs to be kept in mind that structures and classes might be accessed in a similar manner, so the reverser needs to remember that all the data objects in an array are of the same type, that they are sequential, and they will often be handled in a loop of some sort. Also, (and this might be the most important part), each elements in an array may be accessed by a *variable offset from the base*.

Since most times an array is accessed through a computed index, not through a constant, the compiler will likely use the following to access an element of the array:

mov [ebx + eax], 0x00

If the array holds elements larger than 1 byte (for char), the index will need to be multiplied by the size of the element, yielding code similar to the following:

mov [ebx + eax * 4], 0x11223344 # access to an array of DWORDs, e.g. arr[i] = 0x11223344 ... mul eax, $20 # access to an array of structs, each 20 bytes long lea edi, [ebx + eax] # e.g. ptr = &arr[i]

This pattern can be used to distinguish between accesses to arrays and accesses to structure data members.

4.3.3 Structures

All C programmers are going to be familiar with the following syntax:

struct MyStruct { int FirstVar; double SecondVar; unsigned short int ThirdVar; }

It's called a **structure** (Pascal programmers may know a similar concept as a "record").

Structures may be very big or very small, and they may contain all sorts of different data. Structures may look very similar to arrays in memory, but a few key points need to be remembered: structures do not need to contain data fields of all the same type, structure fields are often 4-byte aligned (not sequential), and each element in a structure has its own offset. It therefore makes no sense to reference a structure element by a variable offset from the base.

Take a look at the following structure definition:

struct MyStruct2 { long value1; short value2; long value3; }

Assuming the pointer to the base of this structure is loaded into ebx, we can access these members in one of two schemes:

The first arrangement is the most common, but it clearly leaves open an entire memory word (2 bytes) at offset +6, which is not used at all. Compilers occasionally allow the programmer to manually specify the offset of each data member, but this isn't always the case. The second example also has the benefit that the reverser can easily identify that each data member in the structure is a different size.

Consider now the following function:

:_MyFunction push ebp mov ebp, esp lea ecx, SS:[ebp + 8] mov [ecx + 0], 0x0000000A mov [ecx + 4], ecx mov [ecx + 8], 0x0000000A mov esp, ebp pop ebp

The function clearly takes a pointer to a data structure as its first argument. Also, each data member is the same size (4 bytes), so how can we tell if this is an array or a structure? To answer that question, we need to remember one important distinction between structures and arrays: the elements in an array are all of the same type, the elements in a structure do not need to be the same type. Given that rule, it is clear that one of the elements in this structure is a pointer (it points to the base of the structure itself!) and the other two fields are loaded with the hex value 0x0A (10 in decimal), which is certainly not a valid pointer on any system I have ever used. We can then partially recreate the structure and the function code below:

struct MyStruct3 { long value1; void *value2; long value3; } void MyFunction2(struct MyStruct3 *ptr) { ptr->value1 = 10; ptr->value2 = ptr; ptr->value3 = 10; }

As a quick aside note, notice that this function doesn't load anything into eax, and therefore it doesn't return a value.

4.3.4 Advanced Structures

Lets say we have the following situation in a function:

:MyFunction1 push ebp mov ebp, esp mov esi, [ebp + 8] lea ecx, SS:[esi + 8] ...

what is happening here? First, esi is loaded with the value of the function's first parameter (ebp + 8). Then, ecx is loaded with a pointer to the offset +8 from esi. It looks like we have 2 pointers accessing the same data structure!

The function in question could easily be one of the following 2 prototypes:

struct MyStruct1 { DWORD value1; DWORD value2;
struct MySubStruct1 { ...
struct MyStruct2 { DWORD value1; DWORD value2; DWORD array[LENGTH]; ...

one pointer offset from another pointer in a structure often means a complex data structure. There are far too many combinations of structures and arrays, however, so this wikibook will not spend too much time on this subject.

4.3.5 Identifying Structs and Arrays

Array elements and structure fields are both accessed as offsets from the array/structure pointer. When disassembling, how do we tell these data structures apart? Here are some pointers:

1. array elements are not meant to be accessed individually. Array elements are typically accessed using a variable offset
2. Arrays are frequently accessed in a loop. Because arrays typically hold a series of similar data items, the best way to access them all is usually a loop. Specifically, for(x = 0; x < length_of_array; x++) style loops are often used to access arrays, although there can be others.
3. All the elements in an array have the same data type.
4. Struct fields are typically accessed using constant offsets.

5. Struct fields are typically not accessed in order, and are also not accessed using loops.

6. Struct fields are not typically all the same data type, or the same data width

4.3.6 Linked Lists and Binary Trees

Two common structures used when programming are linked lists and binary trees. These two structures in turn can be made more complicated in a number of ways. Shown in the images below are examples of a linked list structure and a binary tree structure.

Each node in a linked list or a binary tree contains some amount of data, and a pointer (or pointers) to other nodes. Consider the following asm code example:

loop_top: cmp [ebp + 0], 10 je loop_end mov ebp, [ebp + 4] jmp loop_top loop_end:

At each loop iteration, a data value at [ebp + 0] is compared with the value 10. If the two are equal, the loop is ended. If the two are not equal, however, the pointer in ebp is updated with a pointer at an offset from ebp, and the loop is continued. This is a classic linked-loop search technique. This is analagous to the following C code:

struct node { int data; struct node *next; }; struct node *x; ... while(x->data != 10) { x = x->next; }

Binary trees are the same, except two different pointers will be used (the right and left branch pointers).

4.4 x86 Disassembly/Objects and Classes

The Objects and Classes page of the X86 Disassembly Wikibook is a stub. You can help by expanding this section.

4.4.1 Object-Oriented Programming

Object-Oriented (OO) programming provides for us a new unit of program structure to contend with: the **Object**. This chapter will look at disassembled classes from C++. This chapter will not deal directly with COM, but it will work to set a lot of the groundwork for future discussions in reversing COM components (Windows users only).

4.4.2 Classes

A basic class that has not inherited anything can be broken into two parts, the variables and the methods. The non-static variables are shoved into a simple data structure while the methods are compiled and called like every other function.

When you start adding in inheritance and polymorphism, things get a little more complicated. For the purposes of simplicity, the structure of an object will be described in terms of having no inheritance. At the end, however, inheritance and polymorphism will be covered.

Variables

All static variables defined in a class resides in the static region of memory for the entire duration of the application. Every other variable defined in the class is placed into a data structure known as an object. Typically when the constructor is called, the variables are placed into the object in sequential order, see **Figure 1**.

A:

class ABC123 { public: int a, b, c; ABC123():a(1), b(2), c(3) {}; };

B:

0x00200000 dd 1 ;int a 0x00200004 dd 2 ;int b 0x00200008 dd 3 ;int c

However, the compiler typically needs the variables to be separated into sizes that are multiples of a word (2 bytes) in order to locate them. Not all variables fit this requirement, namely char arrays; some unused bits might be used pad the variables so they meet this size requirement. This is illustrated in **Figure 2**.

A:

class ABC123{ public: int a; char b[3]; double c; ABC123():a(1),c(3) { strcpy(b,"02"); }; };

B:

0x00200000 dd 1 ;int a ; offset = abc123 + 0*word_size 0x00200004 db '0' ;b[0] = '0' ; offset = abc123 + 2*word_size 0x00200005 db '2' ;b[1] = '2' 0x00200006 db 0 ;b[2] = null 0x00200007 db 0 ;<= UNUSED BYTE 0x00200008 dd 0x00000000 ;double c, lower 32 bits ; offset = abc123 + 4*word_size 0x0020000C dd 0x40080000 ;double c, upper 32 bits

In order for the application to access one of these object variables, an object pointer needs to be offset to find the desired variable. The offset of every variable is known by the compiler and written into the object code wherever it's needed. **Figure 3** shows how to offset a pointer to

retrieve variables.

;abc123 = pointer to object mov eax, [abc123] ;eax = &a ;offset = abc123+0*word_size = abc123 mov ebx, [abc123+4] ;ebx = &b ;offset = abc123+2*word_size = abc123+4 mov ecx, [abc123+8] ;ecx = &c ;offset = abc123+4*word_size = abc123+8

Figure 3: This shows how to offset a pointer to retrieve variables. The first line places the address of variable 'a' into eax. The second line places the address of variable 'b' into ebx. And the last line places the variable 'c' into ecx.

Methods

At a low level, there is almost no difference between a function and a method. When decompiling, it can sometimes be hard to tell a difference between the two. They both reside in the text memory space, and both are called the same way. An example of how a method is called can be seen in **Figure 4**.

A:

//method call abc123->foo(1, 2, 3);

B:

push 3 ; int c push 2 ; int b push 1 ; int a push [ebp-4] ; the address of the object call 0x00434125 ; call to method

A notable characteristic in a method call is the address of the object being passed in as an argument. This, however, is not a always a good indicator. **Figure 5** shows function with the first argument being an object passed in by reference. The result is function that looks identical to a method call.

A:

//function call foo(abc123, 1, 2, 3);

B:

push 3 ; int c push 2 ; int b push 1 ; int a push [ebp+4] ; the address of the object call 0x00498372 ; call to function

Inheritance & Polymorphism

Inheritance and polymorphism completely changes the structure of a class, the object no longer contains just variables, they also contain pointers to the inherited methods. This is due to the fact that polymorphism requires the address of a method or inner object to be figured out at runtime.

Take **Figure 6** into consideration. How does the application know to call D::one or C::one? The answer is that the compiler figures out a convention in which to order variables and method pointers inside the object such that when they're referenced, the offsets are the same for any object that has inherited its methods and variables.

The abstract class A acts as a blueprint for the compiler, defining an expected structure for any class that inherits it. Every variable defined in class A and every virtual method defined in A will have the exact same offset for any of its children. **Figure 7** declares a possible inheritance scheme as well as it structure in memory. Notice how the offset to C::one is the same as D::one, and the offset to C's copy of A::a is the same as D's copy. In this, our polymorphic loop can just iterate through the array of pointers and know exactly where to find each method.

A:

class A{ public: int a; virtual void one() = 0; }; class B{ public: int b; int c; virtual void two() = 0; }; class C: public A{ public: int d; void one(); }; class D: public A, public B{ public: int e; void one(); void two(); };

B:

;Object C 0x00200000 dd 0x00423848 ; address of C::one ;offset = 0*word_size 0x00200004 dd 1 ; C's copy of A::a ;offset = 2*word_size 0x00200008 dd 4 ; C::d ;offset = 4*word_size ;Object D 0x00200100 dd 0x00412348 ; address of D::one ;offset = 0*word_size 0x00200104 dd 1 ; D's copy of A::a ;offset = 2*word_size 0x00200108 dd 0x00431255 ; address of D::two ;offset = 4*word_size 0x0020010C dd 2 ; D's copy of B::b ;offset = 6*word_size 0x00200110 dd 3 ; D's copy of B::c ;offset = 8*word_size 0x00200114 dd 5 ; D::e ;offset = 10*word_size

4.4.3 Classes Vs. Structs

4.5 x86 Disassembly/Floating Point Numbers

4.5.1 Floating Point Numbers

This page will talk about how **floating point** numbers are used in assembly language constructs. This page will not talk about new constructs, it will not explain what the FPU instructions do, how floating point numbers are stored or manipulated, or the differences in floating-point data representations. However, this page will demon-

strate briefly how floating-point numbers are used in code and data structures that we have already considered.

The x86 architecture does not have any registers specifically for floating point numbers, but it does have a special stack for them. The floating point stack is built directly into the processor, and has access speeds similar to those of ordinary registers. Notice that the FPU stack is not the same as the regular system stack.

4.5.2 Calling Conventions

With the addition of the floating-point stack, there is an entirely new dimension for passing parameters and returning values. We will examine our calling conventions here, and see how they are affected by the presence of floating-point numbers. These are the functions that we will be assembling, using both GCC, and cl.exe:

__cdecl double MyFunction1(double x, double y, float z) { return (x + 1.0) * (y + 2.0) * (z + 3.0); } __fastcall double MyFunction2(double x, double y, float z) { return (x + 1.0) * (y + 2.0) * (z + 3.0); } __stdcall double MyFunction3(double x, double y, float z) { return (x + 1.0) * (y + 2.0) * (z + 3.0); }

CDECL

Here is the cl.exe assembly listing for MyFunction1:

PUBLIC _MyFunction1 PUBLIC __real@3ff0000000000000 PUBLIC __real@4000000000000000 PUBLIC __real@4008000000000000 EXTRN __fltused:NEAR ; COMDAT __real@3ff0000000000000 CONST SEGMENT __real@3ff0000000000000 DQ 03ff0000000000000r ; 1 CONST ENDS ; COMDAT __real@4000000000000000 CONST SEGMENT __real@4000000000000000 DQ 04000000000000000r ; 2 CONST ENDS ; COMDAT __real@4008000000000000 CONST SEGMENT __real@4008000000000000 DQ 04008000000000000r ; 3 CONST ENDS _TEXT SEGMENT _x$ = 8 ; size = 8 _y$ = 16 ; size = 8 _z$ = 24 ; size = 4 _MyFunction1 PROC NEAR ; Line 2 push ebp mov ebp, esp ; Line 3 fld QWORD PTR _x$[ebp] fadd QWORD PTR __real@3ff0000000000000 fld QWORD PTR _y$[ebp] fadd QWORD PTR __real@4000000000000000 fmulp ST(1), ST(0) fld DWORD PTR _z$[ebp] fadd QWORD PTR __real@4008000000000000 fmulp ST(1), ST(0) ; Line 4 pop ebp ret 0 _MyFunction1 ENDP _TEXT ENDS

Our first question is this: are the parameters passed on the stack, or on the floating-point register stack, or some place different entirely? Key to this question, and to this function is a knowledge of what **fld** and **fstp** do. fld (Floating-point Load) pushes a floating point value onto the FPU stack, while fstp (Floating-Point Store and Pop) moves a floating point value from ST0 to the specified location, and then pops the value from ST0 off the stack entirely. Remember that **double** values in cl.exe are treated as 8-byte storage locations (QWORD), while floats are only stored as 4-byte quantities (DWORD). It is also important to remember that floating point numbers are not stored in a human-readable form in memory, even if the reader has a solid knowledge of binary. Remember, these aren't integers. Unfortunately, the exact format of floating point numbers is well beyond the scope of this chapter.

x is offset +8, y is offset +16, and z is offset +24 from ebp. Therefore, z is pushed first, x is pushed last, and the parameters are passed right-to-left on the *regular stack* not the floating point stack. To understand how a value is returned however, we need to understand what **fmulp** does. fmulp is the "Floating-Point Multiply and Pop" instruction. It performs the instructions:

ST1 := ST1 * ST0 FPU POP ST0

This multiplies ST(1) and ST(0) and stores the result in ST(1). Then, ST(0) is marked empty and stack pointer is incremented. Thus, contents of ST(1) are on the top of the stack. So the top 2 values are multiplied together, and the result is stored on the top of the stack. Therefore, in our instruction above, "fmulp ST(1), ST(0)", which is also the last instruction of the function, we can see that the last result is stored in ST0. Therefore, floating point parameters are passed on the regular stack, but floating point results are passed on the FPU stack.

One final note is that MyFunction2 cleans its own stack, as referenced by the **ret 20** command at the end of the listing. Because none of the parameters were passed in registers, this function appears to be exactly what we would expect an STDCALL function would look like: parameters passed on the stack from right-to-left, and the function cleans its own stack. We will see below that this is actually a correct assumption.

For comparison, here is the GCC listing:

LC1: .long 0 .long 1073741824 .align 8 LC2: .long 0 .long 1074266112 .globl _MyFunction1 .def _MyFunction1; .scl 2; .type 32; .endef _MyFunction1: pushl %ebp movl %esp, %ebp subl $16, %esp fldl 8(%ebp) fstpl −8(%ebp) fldl 16(%ebp) fstpl −16(%ebp) fldl −8(%ebp) fld1 faddp %st, %st(1) fldl −16(%ebp) fldl LC1 faddp %st, %st(1) fmulp %st, %st(1) flds 24(%ebp) fldl LC2 faddp %st, %st(1) fmulp %st, %st(1) leave ret .align 8

This is a very difficult listing, so we will step through it (albeit quickly). 16 bytes of extra space is allocated on the stack. Then, using a combination of fldl and fstpl instructions, the first 2 parameters are moved from offsets +8 and +16, to offsets −8 and −16 from ebp. Seems

like a waste of time, but remember, optimizations are off. **fld1** loads the floating point value 1.0 onto the FPU stack. **faddp** then adds the top of the stack (1.0), to the value in ST1 ([ebp - 8], originally [ebp + 8]).

FASTCALL

Here is the cl.exe listing for MyFunction2:

PUBLIC @MyFunction2@20 PUBLIC __real@3ff0000000000000 PUBLIC __real@4000000000000000 PUBLIC __real@4008000000000000 EXTRN __fltused:NEAR ; COMDAT __real@3ff0000000000000 CONST SEGMENT __real@3ff0000000000000 DQ 03ff0000000000000r ; 1 CONST ENDS ; COMDAT __real@4000000000000000 CONST SEGMENT __real@4000000000000000 DQ 04000000000000000r ; 2 CONST ENDS ; COMDAT __real@4008000000000000 CONST SEGMENT __real@4008000000000000 DQ 04008000000000000r ; 3 CONST ENDS _TEXT SEGMENT _x$ = 8 ; size = 8 _y$ = 16 ; size = 8 _z$ = 24 ; size = 4 @MyFunction2@20 PROC NEAR ; Line 7 push ebp mov ebp, esp ; Line 8 fld QWORD PTR _x$[ebp] fadd QWORD PTR __real@3ff0000000000000 fld QWORD PTR _y$[ebp] fadd QWORD PTR __real@4000000000000000 fmulp ST(1), ST(0) fld DWORD PTR _z$[ebp] fadd QWORD PTR __real@4008000000000000 fmulp ST(1), ST(0) ; Line 9 pop ebp ret 20 ; 00000014H @MyFunction2@20 ENDP _TEXT ENDS

We can see that this function is taking 20 bytes worth of parameters, because of the @20 decoration at the end of the function name. This makes sense, because the function is taking two **double** parameters (8 bytes each), and one **float** parameter (4 bytes each). This is a grand total of 20 bytes. We can notice at a first glance, without having to actually analyze or understand any of the code, that there is only one register being accessed here: **ebp**. This seems strange, considering that FASTCALL passes its regular 32-bit arguments in registers. However, that is not the case here: all the floating-point parameters (even z, which is a 32-bit float) are passed on the stack. We know this, because by looking at the code, there is no other place where the parameters could be coming from.

Notice also that **fmulp** is the last instruction performed again, as it was in the CDECL example. We can infer then, without investigating too deeply, that the result is passed at the top of the floating-point stack.

Notice also that x (offset [ebp + 8]), y (offset [ebp + 16]) and z (offset [ebp + 24]) are pushed in reverse order: z is first, x is last. This means that floating point parameters are passed in right-to-left order, on the stack. This is exactly the same as CDECL code, although only because we are using floating-point values.

Here is the GCC assembly listing for MyFunction2:

.align 8 LC5: .long 0 .long 1073741824 .align 8 LC6: .long 0 .long 1074266112 .globl @MyFunction2@20 .def @MyFunction2@20; .scl 2; .type 32; .endef @MyFunction2@20: pushl %ebp movl %esp, %ebp subl $16, %esp fldl 8(%ebp) fstpl −8(%ebp) fldl 16(%ebp) fstpl −16(%ebp) fldl −8(%ebp) fld1 faddp %st, %st(1) fldl −16(%ebp) fldl LC5 faddp %st, %st(1) fmulp %st, %st(1) flds 24(%ebp) fldl LC6 faddp %st, %st(1) fmulp %st, %st(1) leave ret $20

This is a tricky piece of code, but luckily we don't need to read it very close to find what we are looking for. First off, notice that no other registers are accessed besides **ebp**. Again, GCC passes all floating point values (even the 32-bit float, z) on the stack. Also, the floating point result value is passed on the top of the floating point stack.

We can see again that GCC is doing something strange at the beginning, taking the values on the stack from [ebp + 8] and [ebp + 16], and moving them to locations [ebp - 8] and [ebp - 16], respectively. Immediately after being moved, these values are loaded onto the floating point stack and arithmetic is performed. z isn't loaded till later, and isn't ever moved to [ebp - 24], despite the pattern.

LC5 and LC6 are constant values, that most likely represent floating point values (because the numbers themselves, 1073741824 and 1074266112 don't make any sense in the context of our example functions. Notice though that both LC5 and LC6 contain two **.long** data items, for a total of 8 bytes of storage? They are therefore most definitely **double** values.

STDCALL

Here is the cl.exe listing for MyFunction3:

PUBLIC _MyFunction3@20 PUBLIC __real@3ff0000000000000 PUBLIC __real@4000000000000000 PUBLIC __real@4008000000000000 EXTRN __fltused:NEAR ; COMDAT __real@3ff0000000000000 CONST SEGMENT __real@3ff0000000000000 DQ 03ff0000000000000r ; 1 CONST ENDS ; COMDAT __real@4000000000000000 CONST SEGMENT __real@4000000000000000 DQ 04000000000000000r ; 2 CONST ENDS ; COMDAT __real@4008000000000000 CONST SEGMENT __real@4008000000000000 DQ 04008000000000000r ; 3 CONST ENDS _TEXT SEGMENT _x$ = 8 ; size = 8 _y$ = 16 ; size = 8 _z$ = 24 ; size = 4 _MyFunction3@20 PROC NEAR ; Line 12 push ebp mov ebp, esp ; Line 13 fld QWORD PTR _x$[ebp] fadd QWORD PTR __real@3ff0000000000000 fld QWORD PTR _y$[ebp] fadd QWORD PTR __real@4000000000000000 fmulp ST(1), ST(0) fld DWORD PTR _z$[ebp] fadd QWORD PTR __real@4008000000000000 fmulp ST(1), ST(0) ;

Line 14 pop ebp ret 20 ; 00000014H _MyFunction3@20 ENDP _TEXT ENDS END

x is the highest on the stack, and z is the lowest, therefore these parameters are passed from right-to-left. We can tell this because x has the smallest offset (offset [ebp + 8]), while z has the largest offset (offset [ebp + 24]). We see also from the final fmulp instruction that the return value is passed on the FPU stack. This function also cleans the stack itself, as noticed by the call '*ret 20*. It is cleaning exactly 20 bytes off the stack which is, incidentally, the total amount that we passed to begin with. We can also notice that the implementation of this function looks exactly like the FASTCALL version of this function. This is true because FASTCALL only passes DWORD-sized parameters in registers, and floating point numbers do not qualify. This means that our assumption above was correct.

Here is the GCC listing for MyFunction3:

.align 8 LC9: .long 0 .long 1073741824 .align 8 LC10: .long 0 .long 1074266112 .globl @MyFunction3@20 .def @MyFunction3@20; .scl 2; .type 32; .endef @MyFunction3@20: pushl %ebp movl %esp, %ebp subl $16, %esp fldl 8(%ebp) fstpl −8(%ebp) fldl 16(%ebp) fstpl −16(%ebp) fldl −8(%ebp) fld1 faddp %st, %st(1) fldl −16(%ebp) fldl LC9 faddp %st, %st(1) fmulp %st, %st(1) flds 24(%ebp) fldl LC10 faddp %st, %st(1) fmulp %st, %st(1) leave ret $20

Here we can also see, after all the opening nonsense, that [ebp - 8] (originally [ebp + 8]) is value x, and that [ebp - 24] (originally [ebp - 24]) is value z. These parameters are therefore passed right-to-left. Also, we can deduce from the final fmulp instruction that the result is passed in ST0. Again, the STDCALL function cleans its own stack, as we would expect.

Conclusions

Floating point values are passed as parameters on the stack, and are passed on the FPU stack as results. Floating point values do not get put into the general-purpose integer registers (eax, ebx, etc...), so FASTCALL functions that only have floating point parameters collapse into STDCALL functions instead. **double** values are 8-bytes wide, and therefore will take up 8-bytes on the stack. **float** values however, are only 4-bytes wide.

4.5.3 Float to Int Conversions

4.5.4 FPU Compares and Jumps

4.6 x86 Disassembly/Floating Point Examples

4.6.1 Example: Floating Point Arithmetic

Here is the C source code, and the GCC assembly listing of a simple C language function that performs simple floating-point arithmetic. Can you determine what the numerical values of LC5 and LC6 are?

__fastcall double MyFunction2(double x, double y, float z) { return (x + 1.0) * (y + 2.0) * (z + 3.0); }
.align 8 LC5: .long 0 .long 1073741824 .align 8 LC6: .long 0 .long 1074266112 .globl @MyFunction2@20 .def @MyFunction2@20; .scl 2; .type 32; .endef @MyFunction2@20: pushl %ebp movl %esp, %ebp subl $16, %esp fldl 8(%ebp) fstpl −8(%ebp) fldl 16(%ebp) fstpl −16(%ebp) fldl −8(%ebp) fld1 faddp %st, %st(1) fldl −16(%ebp) fldl LC5 faddp %st, %st(1) fmulp %st, %st(1) flds 24(%ebp) fldl LC6 faddp %st, %st(1) fmulp %st, %st(1) leave ret $20

For this, we don't even need a floating-point number calculator, although you are free to use one if you wish (and if you can find a good one). LC5 is added to [ebp - 16], which we know to be y, and LC6 is added to [ebp - 24], which we know to be z. Therefore, LC5 is the number “2.0”, and LC6 is the number “3.0”. Notice that the **fld1** instruction automatically loads the top of the floating-point stack with the constant value “1.0”.

Chapter 5

Difficulties

5.1 x86 Disassembly/Code Optimization

5.1.1 Code Optimization

An **optimizing compiler** is perhaps one of the most complicated, most powerful, and most interesting programs in existence. This chapter will talk about optimizations, although this chapter will not include a table of common optimizations.

5.1.2 Stages of Optimizations

There are two times when a compiler can perform optimizations: first, in the intermediate representation, and second, during the code generation.

Intermediate Representation Optimizations

While in the intermediate representation, a compiler can perform various optimizations, often based on dataflow analysis techniques. For example, consider the following code fragment:

x = 5; if(x != 5) { //loop body }

An optimizing compiler might notice that at the point of "if (x != 5)", the value of x is always the constant "5". This allows substituting "5" for x resulting in "5 != 5". Then the compiler notices that the resulting expression operates entirely on constants, so the value can be calculated now instead of at run time, resulting in optimizing the conditional to "if (false)". Finally the compiler sees that this means the body of the if conditional will never be executed, so it can omit the entire body of the if conditional altogether.

Consider the reverse case:

x = 5; if(x == 5) { //loop body }

In this case, the optimizing compiler would notice that the IF conditional will always be true, and it won't even bother writing code to test x.

Control Flow Optimizations

Another set of optimization which can be performed either at the intermediate or at the code generation level are control flow optimizations. Most of these optimizations deal with the elimination of useless branches. Consider the following code:

if(A) { if(B) { C; } else { D; } end_B: } else { E; } end_A:

In this code, a simplistic compiler would generate a jump from the C block to end_B, and then another jump from end_B to end_A (to get around the E statements). Clearly jumping to a jump is inefficient, so optimizing compilers will generate a direct jump from block C to end_A.

This unfortunately will make the code more confused and will prevent a nice recovery of the original code. For complex functions, it's possible that one will have to consider the code made of only if()-goto; sequences, without being able to identify higher level statements like if-else or loops.

The process of identifying high level statement hierarchies is called "code structuring".

Code Generation Optimizations

Once the compiler has sifted through all the logical inefficiencies in your code, the code generator takes over. Often the code generator will replace certain slow machine instructions with faster machine instructions.

For instance, the instruction:

beginning: ... loopnz beginning

operates *much* slower than the equivalent instruction set:

beginning: ... dec ecx jne beginning

So then why would a compiler ever use a loopxx instruc-

tion? The answer is that most optimizing compilers never use a loopxx instruction, and therefore as a reverser, you will probably never see one used in real code.

What about the instruction:

mov eax, 0

The mov instruction is relatively quick, but a faster part of the processor is the arithmetic unit. Therefore, it makes more sense to use the following instruction:

xor eax, eax

because xor operates in very few processor cycles (and saves a byte or two at the same time), and is therefore faster than a "mov eax, 0". The only drawback of a xor instruction is that it changes the processor flags, so it cannot be used between a comparison instruction and the corresponding conditional jump.

5.1.3 Loop Unwinding

When a loop needs to run for a small, but definite number of iterations, it is often better to **unwind the loop** in order to reduce the number of jump instructions performed, and in many cases prevent the processor's branch predictor from failing. Consider the following C loop, which calls the function MyFunction() 5 times:

for(x = 0; x < 5; x++) { MyFunction(); }

Converting to assembly, we see that this becomes, roughly:

mov eax, 0 loop_top: cmp eax, 5 jge loop_end call _MyFunction inc eax jmp loop_top

Each loop iteration requires the following operations to be performed:

1. Compare the value in eax (the variable "x") to 5, and jump to the end if greater then or equal
2. Increment eax
3. Jump back to the top of the loop.

Notice that we remove all these instructions if we manually repeat our call to MyFunction():

call _MyFunction call _MyFunction call _MyFunction call _MyFunction call _MyFunction

This new version not only takes up less disk space because it uses fewer instructions, but also runs faster because fewer instructions are executed. This process is called **Loop Unwinding**.

5.1.4 Inline Functions

The C and C++ languages allow the definition of an inline type of function. Inline functions are functions which are treated similarly to macros. During compilation, calls to an inline function are replaced with the body of that function, instead of performing a call instruction. In addition to using the inline keyword to declare an inline function, optimizing compilers may decide to make other functions inline as well.

Function inlining works similarly to loop unwinding for increasing code performance. A non-inline function requires a call instruction, several instructions to create a stack frame, and then several more instructions to destroy the stack frame and return from the function. By copying the body of the function instead of making a call, the size of the machine code increases, but the execution time *decreases*.

It is not necessarily possible to determine whether identical portions of code were created originally as macros, inline functions, or were simply copy and pasted. However, when disassembling it can make your work easier to separate these blocks out into separate inline functions, to help keep the code straight.

5.2 x86 Disassembly/Optimization Examples

5.2.1 Example: Optimized vs Non-Optimized Code

The following example is adapted from an algorithm presented in Knuth(vol 1, chapt 1) used to find the greatest common denominator of 2 integers. Compare the listing file of this function when compiler optimizations are turned on and off.

/*line 1*/ int EuclidsGCD(int m, int n) /*we want to find the GCD of m and n*/ { int q, r; /*q is the quotient, r is the remainder*/ while(1) { q = m / n; /*find q and r*/ r = m % n; if(r == 0) /*if r is 0, return our n value*/ { return n; } m = n; /*set m to the current n value*/ n = r; /*set n to our current remainder value*/ } /*repeat*/ }

Compiling with the Microsoft C compiler, we generate a listing file using no optimization:

PUBLIC _EuclidsGCD _TEXT SEGMENT _r$ = −8 ; size = 4 _q$ = −4 ; size = 4 _m$ = 8 ; size = 4 _n$ = 12 ; size = 4 _EuclidsGCD PROC NEAR ; Line 2 push ebp mov ebp, esp sub esp, 8 $L477: ; Line 4 mov eax, 1 test eax, eax je SHORT $L473 ; Line 6 mov eax, DWORD PTR _m$[ebp] cdq idiv DWORD PTR

_n$[ebp] mov DWORD PTR _q$[ebp], eax ; Line 7 mov eax, DWORD PTR _m$[ebp] cdq idiv DWORD PTR _n$[ebp] mov DWORD PTR _r$[ebp], edx ; Line 8 cmp DWORD PTR _r$[ebp], 0 jne SHORT $L479 ; Line 10 mov eax, DWORD PTR _n$[ebp] jmp SHORT $L473 $L479: ; Line 12 mov ecx, DWORD PTR _n$[ebp] mov DWORD PTR _m$[ebp], ecx ; Line 13 mov edx, DWORD PTR _r$[ebp] mov DWORD PTR _n$[ebp], edx ; Line 14 jmp SHORT $L477 $L473: ; Line 15 mov esp, ebp pop ebp ret 0 _EuclidsGCD ENDP _TEXT ENDS END

Notice how there is a very clear correspondence between the lines of C code, and the lines of the ASM code. the addition of the "; line x" directives is very helpful in that respect.

Next, we compile the same function using a series of optimizations to stress speed over size:

cl.exe /Tceuclids.c /Fa /Ogt2

and we produce the following listing:

PUBLIC _EuclidsGCD _TEXT SEGMENT _m$ = 8 ; size = 4 _n$ = 12 ; size = 4 _EuclidsGCD PROC NEAR ; Line 7 mov eax, DWORD PTR _m$[esp-4] push esi mov esi, DWORD PTR _n$[esp] cdq idiv esi mov ecx, edx ; Line 8 test ecx, ecx je SHORT $L563 $L547: ; Line 12 mov eax, esi cdq idiv ecx ; Line 13 mov esi, ecx mov ecx, edx test ecx, ecx jne SHORT $L547 $L563: ; Line 10 mov eax, esi pop esi ; Line 15 ret 0 _EuclidsGCD ENDP _TEXT ENDS END

As you can see, the optimized version is significantly shorter then the non-optimized version. Some of the key differences include:

- The optimized version does not prepare a standard stack frame. This is important to note, because many times new reversers assume that functions always start and end with proper stack frames, and this is clearly not the case. EBP isnt being used, ESP isnt being altered (because the local variables are kept in registers, and not put on the stack), and no subfunctions are called. 5 instructions are cut by this.

- The “test EAX, EAX” series of instructions in the non-optimized output, under ";line 4” is all unnecessary. The while-loop is defined by “while(1)" and therefore the loop always continues. this extra code is safely cut out. Notice also that there is no unconditional jump in the loop like would be expected: the “if(r == 0) return n;" instruction has become the new loop condition.

- The structure of the function is altered greatly: the division of m and n to produce q and r is performed in this function twice: once at the beginning of the function to initialize, and once at the end of the loop. Also, the value of r is tested twice, in the same places. The compiler is very liberal with how it assigns storage in the function, and readily discards values that are not needed.

5.2.2 Example: Manual Optimization

The following lines of assembly code are not optimized, but they can be optimized very easily. Can you find a way to optimize these lines?

mov eax, 1 test eax, eax je SHORT $L473

The code in this line is the code generated for the “while(1)" C code, to be exact, it represents the loop break condition. Because this is an infinite loop, we can assume that these lines are unnecessary.

“mov eax, 1” initializes eax.

the test immediately afterwards tests the value of eax to ensure that it is nonzero. because eax will always be nonzero (eax = 1) at this point, the conditional jump can be removed along whith the “mov” and the “test”.

The assembly is actually checking whether 1 equals 1. Another fact is, that the C code for an infinite **FOR** loop:

for(; ;) { ... }

would not create such a meaningless assembly code to begin with, and is logically the same as “while(1)".

5.2.3 Example: Trace Variables

Here are the C code and the optimized assembly listing from the EuclidGCD function, from the example above. Can you determine which registers contain the variables **r** and **q**?

/*line 1*/ int EuclidsGCD(int m, int n) /*we want to find the GCD of m and n*/ { int q, r; /*q is the quotient, r is the remainder*/ while(1) { q = m / n; /*find q and r*/ r = m % n; if(r == 0) /*if r is 0, return our n value*/ { return n; } m = n; /*set m to the current n value*/ n = r; /*set n to our current remainder value*/ } /*repeat*/ } PUBLIC _EuclidsGCD _TEXT SEGMENT _m$ = 8 ; size = 4 _n$ = 12 ; size = 4 _EuclidsGCD PROC NEAR ; Line 7 mov eax, DWORD PTR _m$[esp-4] push esi mov esi, DWORD PTR _n$[esp] cdq idiv esi mov ecx, edx ; Line 8 test ecx, ecx je SHORT $L563 $L547: ; Line 12 mov eax, esi cdq idiv ecx ; Line 13 mov esi, ecx mov ecx, edx test ecx, ecx jne SHORT $L547 $L563: ; Line 10 mov eax, esi pop esi ; Line 15 ret 0 _EuclidsGCD ENDP _TEXT ENDS END

At the beginning of the function, **eax** contains m, and **esi** contains n. When the instruction “idiv esi” is executed, eax contains the quotient (q), and edx contains the re-

mainder (r). The instruction “mov ecx, edx” moves r into **ecx**, while q is not used for the rest of the loop, and is therefore discarded.

5.2.4 Example: Decompile Optimized Code

Below is the optimized listing file of the EuclidGCD function, presented in the examples above. Can you decompile this assembly code listing into equivalent “optimized” C code? How is the optimized version different in structure from the non-optimized version?

PUBLIC _EuclidsGCD _TEXT SEGMENT _m$ = 8 ; size = 4 _n$ = 12 ; size = 4 _EuclidsGCD PROC NEAR ; Line 7 mov eax, DWORD PTR _m$[esp-4] push esi mov esi, DWORD PTR _n$[esp] cdq idiv esi mov ecx, edx ; Line 8 test ecx, ecx je SHORT $L563 $L547: ; Line 12 mov eax, esi cdq idiv ecx ; Line 13 mov esi, ecx mov ecx, edx test ecx, ecx jne SHORT $L547 $L563: ; Line 10 mov eax, esi pop esi ; Line 15 ret 0 _EuclidsGCD ENDP _TEXT ENDS END

Altering the conditions to maintain the same structure gives us:

int EuclidsGCD(int m, int n) { int r; r = m % n; if(r != 0) { do { m = n; r = m % r; n = r; }while(r != 0) } return n; }

It is up to the reader to compile this new “optimized” C code, and determine if there is any performance increase. Try compiling this new code without optimizations first, and then with optimizations. Compare the new assembly listings to the previous ones.

5.2.5 Example: Instruction Pairings

Q Why does the **dec/jne** combo operate faster than the equivalent **loopnz**?

A The **dec/jnz** pair operates faster then a **loopsz** for several reasons. First, **dec** and **jnz** pair up in the different modules of the netburst pipeline, so they can be executed simultaneously. Top that off with the fact that **dec** and **jnz** both require few cycles to execute, while the **loopnz** (and all the loop instructions, for that matter) instruction takes more cycles to complete. loop instructions are rarely seen output by good compilers.

5.2.6 Example: Avoiding Branches

Below is an assembly version of the expression c ? d : 0. There is no branching in the code, so how does it work?

; ecx = c and edx = d ; eax will contain c ? d : 0 (eax = d if c is not zero, otherwise eax = 0) neg ecx sbb eax, eax and eax, edx ret

This is an example of using various arithmetic instructions to avoid branching. The **neg** instruction sets the carry flag if *c* is not zero; otherwise, it clears the carry flag. The next line depends on this. If the carry flag is set, then **sbb** results in eax = eax - eax - 1 = 0xffffffff. Otherwise, eax = eax - eax = 0. Finally, performing an **and** on this result ensures that if **ecx** was not zero in the first place, **eax** will contain **edx**, and zero otherwise.

5.2.7 Example: Duff’s Device

What does the following C code function do? Is it useful? Why or why not?

void MyFunction(int *arrayA, int *arrayB, int cnt) { switch(cnt % 6) { while(cnt != 0) { case 0: arrayA[--cnt] = arrayB[cnt]; case 5: arrayA[--cnt] = arrayB[cnt]; case 4: arrayA[--cnt] = arrayB[cnt]; case 3: arrayA[--cnt] = arrayB[cnt]; case 2: arrayA[--cnt] = arrayB[cnt]; case 1: arrayA[--cnt] = arrayB[cnt]; } } }

This piece of code is known as a **Duff’s device** or “Duff’s machine”. It is used to partially unwind a loop for efficiency. Notice the strange way that the while() is nested inside the switch statement? Two arrays of integers are passed to the function, and at each iteration of the while loop, 6 consecutive elements are copied from arrayB to arrayA. The switch statement, since it is outside the while loop, only occurs at the beginning of the function. The modulo is taken of the variable cnt with respect to 6. If cnt is not evenly divisible by 6, then the modulo statement is going to start the loop off somewhere in the middle of the rotation, thus preventing the loop from causing a buffer overflow without having to test the current count after each iteration.

Duff’s Device is considered one of the more efficient general-purpose methods for copying strings, arrays, or data streams.

5.3 x86 Disassembly/Code Obfuscation

5.3.1 Code Obfuscation

Code Obfuscation is the act of making the assembly code or machine code of a program more difficult to disassemble or decompile. The term “obfuscation” is typically used to suggest a deliberate attempt to add difficulty,

but many other practices will cause code to be obfuscated without that being the intention. Software vendors may attempt to obfuscate or even encrypt code to prevent reverse engineering efforts. There are many different types of obfuscations. Notice that many code optimizations (discussed in the previous chapter) have the side-effect of making code more difficult to read, and therefore optimizations act as obfuscations.

5.3.2 What is Code Obfuscation?

There are many things that obfuscation could be:

- Encrypted code that is decrypted prior to runtime.
- Compressed code that is decompressed prior to runtime.
- Executables that contain Encrypted sections, and a simple decrypter.
- Code instructions that are put in a hard-to read order.
- Code instructions which are used in a non-obvious way.

This chapter will try to examine some common methods of obfuscating code, but will not necessarily delve into methods to break the obfuscation.

5.3.3 Interleaving

Optimizing Compilers will engage in a process called **interleaving** to try and maximize parallelism in pipelined processors. This technique is based on two premises:

1. That certain instructions can be executed out of order and still maintain the correct output
2. That processors can perform certain pairs of tasks simultaneously.

x86 NetBurst Architecture

The Intel **NetBurst Architecture** divides an x86 processor into 2 distinct parts: the supporting hardware, and the primitive core processor. The primitive core of a processor contains the ability to perform some calculations blindingly fast, but not the instructions that you or I am familiar with. The processor first converts the code instructions into a form called "micro-ops" that are then handled by the primitive core processor.

The processor can also be broken down into 4 components, or modules, each of which is capable of performing certain tasks. Since each module can operate separately, up to 4 separate tasks can be handled *simultaneously* by the processor core, so long as those tasks can be performed by each of the 4 modules:

Port0 Double-speed integer arithmetic, floating point load, memory store

Port1 Double-speed integer arithmetic, floating point arithmetic

Port2 memory read

Port3 memory write (writes to address bus)

So for instance, the processor can simultaneously perform 2 integer arithmetic instructions in both Port0 and Port1, so a compiler will frequently go to great lengths to put arithmetic instructions close to each other. If the timing is just right, up to 4 arithmetic instructions can be executed in a single instruction period.

Notice however that writing to memory is particularly slow (requiring the address to be sent by Port3, and the data itself to be written by Port0). Floating point numbers need to be loaded to the FPU before they can be operated on, so a floating point load and a floating point arithmetic instruction cannot operate on a single value in a single instruction cycle. Therefore, it is not uncommon to see floating point values loaded, integer values be manipulated, and then the floating point value be operated on.

5.3.4 Non-Intuitive Instructions

Optimizing compilers frequently will use instructions that are not intuitive. Some instructions can perform tasks for which they were not designed, typically as a helpful side effect. Sometimes, one instruction can perform a task more quickly than other specialized instructions can.

The only way to know that one instruction is faster than another is to consult the processor documentation. However, knowing some of the most common substitutions is very useful to the reverser.

Here are some examples. The code in the first box operates more quickly than the one in the second, but performs exactly the same tasks.

Example 1

Fast

```
xor eax, eax
```

Slow

```
mov eax, 0
```

Example 2

Fast

```
shl eax, 3
```

Slow

mul eax, 8

Sometimes such transformations could be made to make the analysis more difficult:

Example 3

Fast

push $next_instr jmp $some_function $next_instr:...

Slow

call $some_function

Example 4

Fast

pop eax jmp eax

Slow

retn

Common Instruction Substitutions

lea The lea instruction has the following form:

lea dest, (XS:)[reg1 + reg2 * x]

Where XS is a segment register (SS, DS, CS, etc...), reg1 is the base address, reg2 is a variable offset, and x is a multiplicative scaling factor. What lea does, essentially, is load the memory address being pointed to in the second argument, into the first argument. Look at the following example:

mov eax, 1 lea ecx, [eax + 4]

Now, what is the value of ecx? The answer is that ecx has the value of (eax + 4), which is 5. In essence, lea is used to do addition and multiplication of a register and a constant that is a byte or less (-128 to +127).

Now, consider:

mov eax, 1 lea ecx, [eax+eax*2]

Now, ecx equals 3.

The difference is that lea is quick (because it only adds a register and a small constant), whereas the **add** and **mul** instructions are more versatile, but slower. lea is used for arithmetic in this fashion very frequently, even when compilers are not actively optimizing the code.

xor The xor instruction performs the bit-wise exclusive-or operation on two operands. Consider then, the following example:

mov al, 0xAA xor al, al

What does this do? Lets take a look at the binary:

10101010 ;10101010 = 0xAA xor 10101010 -------- 00000000

The answer is that "xor reg, reg" sets the register to 0. More importantly, however, is that "xor eax, eax" sets eax to 0 *faster* (and the generated code instruction is smaller) than an equivalent "mov eax, 0".

mov edi, edi On a 64-bit x86 system, this instruction clears the high 32-bits of the rdi register.

shl, shr left-shifting, in binary arithmetic, is equivalent to multiplying the operand by 2. Right-shifting is also equivalent to integer division by 2, although the lowest bit is dropped. in general, left-shifting by N spaces multiplies the operand by 2^N , and right shifting by N spaces is the same as dividing by 2^N . One important fact is that resulting number is an integer with no fractional part present. For example:

mov al, 31 ; 00011111 shr al, 1 ; 00001111 = 15, not 15.5

xchg xchg exchanges the contents of two registers, or a register and a memory address. A noteworthy point is the fact that xchg operates faster than a move instruction. For this reason, xchg will be used to move a value from a source to a destination, when the value in the source no longer needs to be saved.

As an example, consider this code:

mov ebx, eax mov eax, 0

Here, the value in eax is stored in ebx, and then eax is loaded with the value zero. We can perform the same operation, but using xchg and xor instead:

xchg eax, ebx xor eax, eax

It may surprise you to learn that the second code example operates significantly faster than the first one does.

5.3.5 Obfuscators

There are a number of tools on the market that will automate the process of code obfuscation. These products will use a number of transformations to turn a code snippet into a less-readable form, although it will not affect the program flow itself (although the transformations may increase code size or execution time).

5.3.6 Code Transformations

Code transformations are a way of reordering code so that it performs exactly the same task but becomes more difficult to trace and disassemble. We can best demonstrate this technique by example. Let's say that we have 2 functions, FunctionA and FunctionB. Both of these two functions are comprised of 3 separate parts, which are performed in order. We can break this down as such:

FunctionA() { FuncAPart1(); FuncAPart2(); FuncAPart3(); } FunctionB() { FuncBPart1(); FuncBPart2(); FuncBPart3(); }

And we have our main program, that executes the two functions:

main() { FunctionA(); FunctionB(); }

Now, we can rearrange these snippets to a form that is much more complicated (in assembly):

main: jmp FAP1 FBP3: call FuncBPart3 jmp end FBP1: call FuncBPart1 jmp FBP2 FAP2: call FuncAPart2 jmp FAP3 FBP2: call FuncBPart2 jmp FBP3 FAP1: call FuncAPart1 jmp FAP2 FAP3: call FuncAPart3 jmp FBP1 end:

As you can see, this is much harder to read, although it perfectly preserves the program flow of the original code. This code is much harder for a human to read, although it isn't hard at all for an automated debugging tool (such as IDA Pro) to read.

5.3.7 Opaque Predicates

An **Opaque Predicate** is a predicate inside the code, that cannot be evaluated during static analysis. This forces the attacker to perform a dynamic analysis to understand the result of the line. Typically this is related to a branch instruction that is used to prevent in static analysis the understanding which code path is taken.

5.3.8 Code Encryption

Code can be encrypted, just like any other type of data, except that code can also work to encrypt and decrypt *itself*. Encrypted programs cannot be directly disassembled. However, such a program can also not be run directly because the encrypted opcodes cannot be interpreted properly by the CPU. For this reason, an encrypted program must contain some sort of method for decrypting itself prior to operation.

The most basic method is to include a small stub program that decrypts the remainder of the executable, and then passes control to the decrypted routines.

Disassembling Encrypted Code

To disassemble an encrypted executable, you must first determine how the code is being decrypted. Code can be decrypted in one of two primary ways:

1. All at once. The entire code portion is decrypted in a single pass, and left decrypted during execution. Using a debugger, allow the decryption routine to run completely, and then dump the decrypted code into a file for further analysis.

2. By Block. The code is encrypted in separate blocks, where each block may have a separate encryption key. Blocks may be decrypted before use, and re-encrypted again after use. Using a debugger, you can attempt to capture all the decryption keys and then use those keys to decrypt the entire program at once later, or you can wait for the blocks to be decrypted, and then dump the blocks individually to a separate file for analysis.

5.4 x86 Disassembly/Debugger Detectors

5.4.1 Detecting Debuggers

It may come as a surprise that a running program can actually detect the presence of an attached user-mode debugger. Also, there are methods available to detect kernel-mode debuggers, although the methods used depend in large part on which debugger is trying to be detected.

This subject is peripheral to the narrative of this book, and the section should be considered an optional one for most readers.

5.4.2 IsDebuggerPresent API

The Win32 API contains a function called "IsDebuggerPresent", which will return a boolean true if the program is being debugged. The following code snippet will detail a general usage of this function:

if(IsDebuggerPresent()) { TerminateProcess(GetCurrentProcess(), 1); }

Of course, it is easy to spot uses of the IsDebuggerPresent() function in the disassembled code, and a skilled reverser will simply patch the code to remove this line. For OllyDbg, there are many plugins available which hide the debugger from this and many other APIs.

5.4.3 PEB Debugger Check

The Process Environment Block stores the value that IsDebuggerPresent queries to determine its return value. To avoid suspicion, some programmers access the value directly from the PEB instead of calling the API function. The following code snippet shows how to access the value:

mov eax, fs:[30h] mov eax, byte [eax+2] test eax, eax jne @DebuggerDetected

5.4.4 Timeouts

Debuggers can put break points in the code, and can therefore stop program execution. A program can detect this, by monitoring the system clock. If too much time has elapsed between instructions, it can be determined that the program is being stopped and analyzed (although this is not always the case). If a program is taking too much time, the program can terminate.

Notice that on preemptive multithreading systems, such as modern Windows or Linux systems will switch away from your program to run other programs. This is called thread switching. If the system has many threads to run, or if some threads are hogging processor time, your program may detect a long delay and may falsely determine that the program is being debugged.

5.4.5 Detecting SoftICE

SoftICE is a local kernel debugger, and as such, it can't be detected as easily as a user-mode debugger can be. The IsDebuggerPresent API function will not detect the presence of SoftICE.

To detect SoftICE, there are a number of techniques that can be used:

1. Search for the SoftICE install directory. If SoftICE is installed, the user is probably a hacker or a reverser.
2. Detect the presence of **int 1**. SoftICE uses interrupt 1 to debug, so if interrupt 1 is installed, SoftICE is running.

5.4.6 Detecting OllyDbg

OllyDbg is a popular 32-bit usermode debugger. Unfortunately, the last few releases, including the latest version (v1.10) contain a vulnerability in the handling of the Win32 API function OutputDebugString(). A programmer trying to prevent his program from being debugged by OllyDbg could exploit this vulnerability in order to make the debugger crash. The author has never released a fix, however there are unofficial versions and plugins available to protect OllyDbg from being exploited using this vulnerability.

Chapter 6

Resources and Licensing

6.1 x86 Disassembly/Resources

6.1.1 Wikimedia Resources

Wikibooks

- X86 Assembly
- Subject:Assembly languages
- Compiler Construction
- Floating Point
- C Programming
- C++ Programming

Wikipedia

6.1.2 External Resources

External Links

- The MASM Project: http://www.masm32.com/
- Randall Hyde's Homepage: http://www.cs.ucr.edu/~{}rhyde/
- Borland Turbo Assembler: http://info.borland.com/borlandcpp/cppcomp/tasmfact.html
- NASM Project Homepage: http://nasm.sourceforge.net/wakka.php?wakka=HomePage
- FASM Homepage: http://flatassembler.net/
- DCC Decompiler:
- Boomerang Decompiler Project:
- Microsoft debugging tools main page:

  http://www.microsoft.com/whdc/devtools/debugging/default.mspx
- Solaris observation and debugging tools main page:

  http://www.opensolaris.org/os/community/dtrace/

  http://www.opensolaris.org/os/community/mdb/
- Free Debugging Tools, Static Source Code Analysis Tools, Bug Trackers
- Microsoft Developers Network (MSDN): http://msdn.microsoft.com
- Gareth Williams: http://gareththegeek.ga.funpic.de/
- B. Luevelsmeyer "PE Format Description":http://www.cs.bilkent.edu.tr/~{}hozgur/PE.TXT PE format description
- TheirCorp "The Unofficial TypeLib Data Format Specification":http://theircorp.byethost11.com/index.php?vw=TypeLib
- MSDN Calling Convention page:
- Dictionary of Algorithms and Data Structures
- Charles Petzold's Homepage: http://www.charlespetzold.com/
- Donald Knuth's Homepage: http://www-cs-faculty.stanford.edu/~{}knuth/
- "THE ISA AND PC/104 BUS" by Mark Sokos 2000
- "Practically Reversing CRC" by Bas Westerbaan 2005
- "CRC and how to Reverse it" by anarchriz 1999
- *"Reverse Engineering is a Way of Life"* by Matthew Russotto
- *"the Reverse and Reengineering Wiki"*
- F-Secure Khallenge III: 2008 Reverse Engineering competition (is this an annual challenge?)

- "Breaking Eggs And Making Omelettes: Topics On Multimedia Technology and Reverse Engineering"
- "Reverse Engineering Stack Exchange"

Books

- Yurichev, Dennis, "An Introduction To Reverse Engineering for Beginners". Online book: http://yurichev.com/writings/RE_for_beginners-en.pdf
- Eilam, Eldad. "Reversing: Secrets of Reverse Engineering." 2005. Wiley Publishing Inc. ISBN 0764574817
- Hyde, Randall. "The Art of Assembly Language," No Starch, 2003 ISBN 1886411972
- Aho, Alfred V. et al. "Compilers: Principles, Techniques and Tools," Addison Wesley, 1986. ISBN: 0321428900
- Steven Muchnick, "Advanced Compiler Design & Implementation," Morgan Kaufmann Publishers, 1997. ISBN 1-55860-320-4
- Kernighan and Ritchie, "The C Programming Language", 2nd Edition, 1988, Prentice Hall.
- Petzold, Charles. "Programming Windows, Fifth Edition," Microsoft Press, 1999
- Hart, Johnson M. "Win32 System Programming, Second Edition," Addison Wesley, 2001
- Gordon, Alan. "COM and COM+ Programming Primer," Prentice Hall, 2000
- Nebbett, Gary. "Windows NT/2000 Native API Reference," Macmillan, 2000
- Levine, John R. "Linkers and Loaders," Morgan-Kauffman, 2000
- Knuth, Donald E. "The Art of Computer Programming," Vol 1, 1997, Addison Wesley.
- *MALWARE: Fighting Malicious Code,* by Ed Skoudis; Prentice Hall, 2004
- *Maximum Linux Security, Second Edition,* by Anonymous; Sams, 2001

6.2 x86 Disassembly/Licensing

6.2.1 Licensing

6.3 x86 Disassembly/Manual of Style

6.3.1 Global Stylesheet

This book has a global stylesheet that can be loaded for you. Go to the **Gadgets** tab at Special:Preferences, and activate the "**Per-book Javascript and Stylesheets**" gadget.

Chapter 7

Text and image sources, contributors, and licenses

7.1 Text

- **Wikibooks:Collections Preface** *Source:* http://en.wikibooks.org/wiki/Wikibooks%3ACollections%20Preface?oldid=2347851 *Contributors:* RobinH, Whiteknight, Jomegat, Mike.lifeguard, Martin Kraus, Adrignola and Magesha
- **X86 Disassembly/Cover** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Cover?oldid=2595883 *Contributors:* Whiteknight, Icktoofay and Anonymous: 2
- **X86 Disassembly/Introduction** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Introduction?oldid=2370674 *Contributors:* DavidCary and Whiteknight
- **X86 Disassembly/Assemblers and Compilers** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Assemblers%20and%20Compilers?oldid=2748744 *Contributors:* DavidCary, Panic2k4, AlbertCahalan, Whiteknight, Az1568, Gcaprino, Scientes, Sigma 7, Adrignola, Jfmantis, EleoTager, Arthur200000 and Anonymous: 21
- **X86 Disassembly/Disassemblers and Decompilers** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Disassemblers%20and%20Decompilers?oldid=2778732 *Contributors:* DavidCary, Mshonle, Panic2k4, AlbertCahalan, Quoth, Whiteknight, Mike Van Emmerik, Koavf, Mdupont, 0xf001, MichaelFrey, Svdb, Herbythyme, Macpunk, C1de0x, Ysangkok, Phatom87, Gannalech, SamB, Spongebob88, QuiteUnusual, Afog, Adrignola, Duplode, JamesCrook, Voomoo, Jfmantis, EleoTager, Arthur200000, Chip Wildon Forster, C4Decompiler, Aquynh and Anonymous: 90
- **X86 Disassembly/Disassembly Examples** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Disassembly%20Examples?oldid=1232569 *Contributors:* Whiteknight and Anonymous: 1
- **X86 Disassembly/Analysis Tools** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Analysis%20Tools?oldid=2759797 *Contributors:* Utcursch, Panic2k4, Marcika, AlbertCahalan, Quoth, Whiteknight, Jomegat, Kaosone, Perpetuum, Hagindaz, Wikimoder, Dr Dnar, Macpunk, Frozen dude, AnthonyD, Spongebob88, MohammadEbrahim, QuiteUnusual, Jodell1, Adrignola, Jfmantis, KenMacD, Rohitab, Rotlink, Arthur200000, IamMe3141 and Anonymous: 61
- **X86 Disassembly/Microsoft Windows** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Microsoft%20Windows?oldid=2838197 *Contributors:* Panic2k4, Quoth, Whiteknight, Hexed321, Chazz, Mantis, Wj32, Gcaprino, Adrignola, Dennis714 and Anonymous: 35
- **X86 Disassembly/Windows Executable Files** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Windows%20Executable%20Files?oldid=2768125 *Contributors:* Quoth, Whiteknight, Shokuku, Barthax, Hexed321, Dr Dnar, Gcaprino, Chris.digiamo, Van der Hoorn, Adrignola, LaZ0r, EroCarrera, Self, CallumPoole and Anonymous: 25
- **X86 Disassembly/Linux** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Linux?oldid=2027237 *Contributors:* Whiteknight, Dr Dnar, Gcaprino, Recent Runes, MohammadEbrahim, Adrignola, Swatnio and Anonymous: 10
- **X86 Disassembly/Linux Executable Files** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Linux%20Executable%20Files?oldid=2748762 *Contributors:* Orderud, Whiteknight, Ddouthitt, Gcaprino, ChrisR, Ulf Abrahamsson, Arthur200000 and Anonymous: 2
- **X86 Disassembly/The Stack** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/The%20Stack?oldid=2622875 *Contributors:* Whiteknight, Dr Dnar, Swift, Mantis, Gcaprino, Gannalech, Jsvcycling, Jfmantis, X-Fi6 and Anonymous: 17
- **X86 Disassembly/Functions and Stack Frames** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Functions%20and%20Stack%20Frames?oldid=2749425 *Contributors:* Whiteknight, Hagindaz, Mantis, Gcaprino, Gannalech, Svick, Jfmantis and Anonymous: 23
- **X86 Disassembly/Functions and Stack Frame Examples** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Functions%20and%20Stack%20Frame%20Examples?oldid=2759822 *Contributors:* Whiteknight, NipplesMeCool, Jfmantis and Anonymous: 2
- **X86 Disassembly/Calling Conventions** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Calling%20Conventions?oldid=2839884 *Contributors:* DavidCary, Whiteknight, Mantis, Gcaprino, Sigma 7, Timjr, Crazy Ivan, Jfmantis and Anonymous: 19
- **X86 Disassembly/Calling Convention Examples** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Calling%20Convention%20Examples?oldid=2699639 *Contributors:* Cspurrier, Whiteknight, Spongebob88, NipplesMeCool and Anonymous: 13

- **X86 Disassembly/Branches** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Branches?oldid=2739113 *Contributors:* Whiteknight, Chazz, Mantis, Leonus, Gcaprino, Gannalech, Spongebob88, Adrignola and Anonymous: 11
- **X86 Disassembly/Branch Examples** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Branch%20Examples?oldid=1791851 *Contributors:* Whiteknight, NipplesMeCool, Jason Lee and Anonymous: 3
- **X86 Disassembly/Loops** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Loops?oldid=2538348 *Contributors:* Whiteknight, Mantis, Gcaprino and Anonymous: 5
- **X86 Disassembly/Loop Examples** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Loop%20Examples?oldid=1975904 *Contributors:* Whiteknight, Sz and Anonymous: 4
- **X86 Disassembly/Variables** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Variables?oldid=1358131 *Contributors:* Whiteknight, Mantis, Gcaprino, Shnizzedy, Spongebob88 and Anonymous: 5
- **X86 Disassembly/Variable Examples** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Variable%20Examples?oldid=1480358 *Contributors:* Whiteknight and NipplesMeCool
- **X86 Disassembly/Data Structures** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Data%20Structures?oldid=2501053 *Contributors:* Whiteknight, Mantis, Gcaprino, Dirk Hünniger and Anonymous: 3
- **X86 Disassembly/Objects and Classes** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Objects%20and%20Classes?oldid=2501049 *Contributors:* Whiteknight, Mantis, Isaiah.v, Dirk Hünniger and Anonymous: 6
- **X86 Disassembly/Floating Point Numbers** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Floating%20Point%20Numbers?oldid=2214827 *Contributors:* Whiteknight, Gcaprino, Spongebob88, Abhi166 and Jfmantis
- **X86 Disassembly/Floating Point Examples** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Floating%20Point%20Examples?oldid=1076115 *Contributors:* Whiteknight
- **X86 Disassembly/Code Optimization** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Code%20Optimization?oldid=1554141 *Contributors:* Whiteknight, Gcaprino and Anonymous: 7
- **X86 Disassembly/Optimization Examples** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Optimization%20Examples?oldid=2676907 *Contributors:* Whiteknight, Wj32, I-VANN and Anonymous: 5
- **X86 Disassembly/Code Obfuscation** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Code%20Obfuscation?oldid=2501052 *Contributors:* DavidCary, AlbertCahalan, Whiteknight, Wj32, Gcaprino, Adrignola, Dirk Hünniger, JamesCrook and Anonymous: 17
- **X86 Disassembly/Debugger Detectors** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Debugger%20Detectors?oldid=1138375 *Contributors:* Orderud, Whiteknight, D0gg, Chris.digiamo and Anonymous: 3
- **X86 Disassembly/Resources** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Resources?oldid=2578714 *Contributors:* DavidCary, Whiteknight, Adrignola, Cognoscent and Anonymous: 3
- **X86 Disassembly/Licensing** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Licensing?oldid=1075890 *Contributors:* Whiteknight
- **X86 Disassembly/Manual of Style** *Source:* http://en.wikibooks.org/wiki/X86%20Disassembly/Manual%20of%20Style?oldid=1076917 *Contributors:* Whiteknight

7.2 Images

- **File:1Fh_01.png** *Source:* http://upload.wikimedia.org/wikibooks/en/a/af/1Fh_01.png *License:* Fair use *Contributors:* ? *Original artist:* ?
- **File:C_language_building_steps.png** *Source:* http://upload.wikimedia.org/wikipedia/commons/b/b3/C_language_building_steps.png *License:* CC-BY-SA-3.0 *Contributors:* ? *Original artist:* ?
- **File:C_language_do_while.png** *Source:* http://upload.wikimedia.org/wikipedia/commons/2/21/C_language_do_while.png *License:* CC BY 3.0 *Contributors:* Own work *Original artist:* Thedsadude
- **File:C_language_for.png** *Source:* http://upload.wikimedia.org/wikipedia/commons/5/51/C_language_for.png *License:* CC BY 3.0 *Contributors:* Own work *Original artist:* Thedsadude
- **File:C_language_if_else.png** *Source:* http://upload.wikimedia.org/wikipedia/commons/a/ac/C_language_if_else.png *License:* CC BY 3.0 *Contributors:* Own work *Original artist:* Thedsadude
- **File:C_language_linked_list.png** *Source:* http://upload.wikimedia.org/wikipedia/commons/1/1b/C_language_linked_list.png *License:* CC BY 3.0 *Contributors:* Own work *Original artist:* Thedsadude
- **File:Data_stack.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/2/29/Data_stack.svg *License:* Public domain *Contributors:* made in Inkscape, by myself User:Boivie. Based on Image:Stack-sv.png, originally uploaded to the Swedish Wikipedia in 2004 by sv:User:Shrimp *Original artist:* User:Boivie
- **File:Elf-layout--en.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/7/77/Elf-layout--en.svg *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* Surueña
- **File:Heckert_GNU_white.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/2/22/Heckert_GNU_white.svg *License:* CC BY-SA 2.0 *Contributors:* gnu.org *Original artist:* Aurelio A. Heckert <aurium@gmail.com>
- **File:Information_icon.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/3/35/Information_icon.svg *License:* Public domain *Contributors:* en:Image:Information icon.svg *Original artist:* El T
- **File:Kernel-exo.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/8/8f/Kernel-exo.svg *License:* CC-BY-SA-3.0 *Contributors:* self-made, based on Image:Kernel-exo.png by User:Aholstenson *Original artist:* Surachit
- **File:RevEngDosHead.JPG** *Source:* http://upload.wikimedia.org/wikipedia/commons/2/2f/RevEngDosHead.JPG *License:* Public domain *Contributors:* Transferred from en.wikibooks; transferred to Commons by User:Adrignola using CommonsHelper. *Original artist:* Original uploader was Whiteknight at en.wikibooks

- **File:RevEngPEFile.JPG** *Source:* http://upload.wikimedia.org/wikipedia/commons/e/ea/RevEngPEFile.JPG *License:* Public domain *Contributors:* Transferred from en.wikibooks; transferred to Commons by User:Adrignola using CommonsHelper. *Original artist:* Original uploader was Whiteknight at en.wikibooks
- **File:RevEngPeSig.JPG** *Source:* http://upload.wikimedia.org/wikipedia/commons/c/ca/RevEngPeSig.JPG *License:* Public domain *Contributors:* Transferred from en.wikibooks; transferred to Commons by User:Adrignola using CommonsHelper. *Original artist:* Original uploader was Whiteknight at en.wikibooks
- **File:ReverseEngineeringPop.JPG** *Source:* http://upload.wikimedia.org/wikibooks/en/8/8c/ReverseEngineeringPop.JPG *License:* Public domain *Contributors:* ? *Original artist:* ?
- **File:ReverseEngineeringPush.JPG** *Source:* http://upload.wikimedia.org/wikibooks/en/9/98/ReverseEngineeringPush.JPG *License:* Public domain *Contributors:* ? *Original artist:* ?
- **File:Tree-data-structure.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/4/45/Tree-data-structure.svg *License:* GFDL 1.2 *Contributors:* ? *Original artist:* ?
- **File:Wikibooks-logo-en-noslogan.svg** *Source:* http://upload.wikimedia.org/wikipedia/commons/d/df/Wikibooks-logo-en-noslogan.svg *License:* CC BY-SA 3.0 *Contributors:* Own work *Original artist:* User:Bastique, User:Ramac et al.
- **File:Wikipedia-logo.png** *Source:* http://upload.wikimedia.org/wikipedia/commons/6/63/Wikipedia-logo.png *License:* GFDL *Contributors:* based on the first version of the Wikipedia logo, by Nohat. *Original artist:* version 1 by Nohat (concept by Paullusmagnus);
- **File:_C_language_if.png** *Source:* http://upload.wikimedia.org/wikipedia/commons/f/fb/C_language_if.png *License:* CC BY 3.0 *Contributors:* Own work *Original artist:* Thedsadude

7.3 Content license

- Creative Commons Attribution-Share Alike 3.0