# Gliederung

1. Einleitung
2. Motivation
3. Ziel
4. Aufbau und Komponenten
   1. Befehl ausführen
   2. Verhalten aufzeichnen
   3. Verhalten analysieren
   4. Auswertung speichern
5. Funktionen
   1. Daten sammeln
   2. Ausführbare Datei emulieren
6. Fazit & Ausblick
7. Anhang

# Einleitung

Der Prozess des Software Reverse Engineerings beschäftigt sich mit der Rekonstruktion vorhandener Softwareprodukte. Zu diesen Produkten gehören unter anderem Binärdateien, welche Maschinenbefehle enthalten. Da dem Benutzer meist die genaue Funktionsweise und Struktur der Datei vorenthalten wird, ist das Ziel durch Untersuchungen (manuell oder automatisiert) Erkenntnisse über diese zu gewinnen. Zu diesen gehören beispielsweise der Aufbau, der Programmfluss, die Struktur von Objekten und die Funktion von Codesegmenten.

Hierfür können verschiedene Techniken angewandt werden. Die gängigste Vorgehensweise besteht darin, einen Disassemblierer (z.B. *IDA, Hopper, objdump*) zu verwenden, um die Funktion von Codesegmenten anhand der erkannten Maschinenbefehle nachzuvollziehen. Jedoch handelt es sich dabei nur um eine Alternativdarstellung des Binärcodes und deshalb ist die Interpretation von Assemblersprache aufgrund einer großen Anzahl atomarer Befehle schwierig und zeitintensiv. Dieser Vorgang erfordert viel Erfahrung, besonders, wenn aus der zu untersuchenden Software hilfreiche Symboltabellen entfernt wurden, welche die in der Programmiersprache verwendeten Variablen- und Funktionsnamen enthalten können.

Um das Verstehen von Objektdateien zu erleichtern, wird aktuell versucht aus dem schwer lesbaren Maschinencode ein für den Menschen leichter lesbaren (Pseudo-)Code zu generieren. Dieser Vorgang des Dekompilierens befindet sich jedoch aufgrund vieler Schwierigkeiten bei der Umsetzung noch in einer frühen Entwicklungsphase. Das Problem besteht darin, dass zwischen mehreren Maschinenbefehlen, die auch an unterschiedlichen Stellen stehen können, Zusammenhänge erkannt und für den Benutzer ausgegeben leicht verständlich werden müssen. Ebenso können Missverständnisse entstehen, wenn Maschinenbefehle direkt in lesbaren Code umgewandelt oder falsch zusammengefasst werden. Durch kleine Fehler in der Übersetzung kann die Darstellung des Codes verwirrender als der Maschinencode selbst sein oder es können wichtige Schritte des Programms verloren gehen. Bei aktuellen Dekompilierern fällt auf, dass der Maschinencode lediglich Zeile für Zeile übersetzt wird. So werden beispielsweise Sprunganweisungen nicht in eine besser verstehbare *if*-Anweisung umgewandelt, sondern direkt mit der im Binärcode entsprechenden Sprungadresse ausgegeben. Dadurch entsteht kein Mehrwert, weshalb diese einfache Übersetzung als überflüssig angesehen werden kann.

Ein weiterer Bereich, der zum Software Reverse Engineering hinzugezählt werden kann, ist die Emulation. Hierbei wird eine Maschine durch Software nachgebildet und Maschinencode nur virtuell ausgeführt. Meist wird diese Technik benutzt um Maschinencode auf einem nicht mit diesem Code kompatiblen System auszuführen. Dies wird erreicht, indem der Maschinencode geparst wird und anschließend unter Verwendung von Datenbanken, welche die Bedeutung des Maschinencodes enthalten, die entsprechenden Änderungen am nachgebildeten System vorgenommen werden. So kann Software, bzw. Maschinencode, ohne direkten Zugriff auf die Hardware ausgeführt werden. Die Vorteile zeichnen sich zum einen durch die Möglichkeit aus, nicht kompatiblen Bytecode auf einer Maschine laufen zu lassen und zu nutzen. Zum anderen bietet die Emulation die Möglichkeit Software schrittweise auszuführen und die Veränderungen zu beobachten und nachzuvollziehen. Auch Schadcode kann ohne Bedenken untersucht werden, weil die Maschine lediglich durch einen Prozess simuliert wird, welcher mit dem Beenden dieses Prozesses aus dem Speicher verschwindet.

Die schwierigste und zeitintensivste Aufgabe bei der Programmierung eines solchen Emulators besteht aus der Erstellung der benötigten Datenbanken. Es muss für jeden auf der Maschine verfügbaren Befehl ein Datensatz erstellt werden, der das Verhalten der Maschine beschreibt. Hierfür müssen große Dokumente, die sogenannten Befehlssatzarchitekturen, durchgearbeitet und für den Emulator abgespeichert werden. Genau dieses Problem ist der Gegenstand dieser Arbeit und soll durch automatische Generierung der benötigten Daten beseitigt werden.

Aufgrund Mangels an Literatur zu diesem Thema, wurde auf ein Quellenverzeichnis verzichtet. Internetquellen werden in der Fußnote durch einen Verweis an den entsprechenden Stellen angegeben.

# Motivation

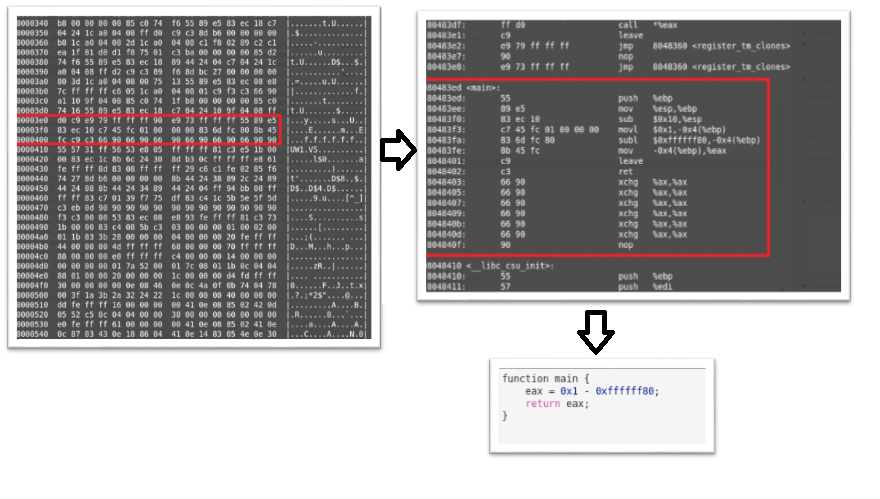
In der vorausgegangenen Projektarbeit wurden verschiedene Software Reverse Engineering Werkzeuge untersucht. Wie in der Einleitung beschrieben, verläuft sich der Rückentwicklungsprozess nicht immer optimal. Die gängigsten Methoden Disassemblieren und Dekompilieren sind in der Abbildung 1 dargestellt.

Abbildung 1 - Disassemblieren und Dekompilieren

Im ersten Fenster sieht man den Inhalt einer ausführbaren Binärdatei. Jede Zeile fängt mit dem Offset in der Datei des jeweiligen Zeichens an. Die Zeichen werden im linken Teil als Hexadezimalzahl dargestellt. Im rechten Teil werden die Zeichen durch die ASCII Darstellung repräsentiert. Ein Punkt gibt in diesem Fall an, dass das Zeichen nicht dargestellt werden kann. Die Verarbeitung des Bytecodes mit einem Disassemblierer liefert die im zweiten Fenster erkennbaren Assemblerbefehle. Aus diesen Assemblerbefehlen die Funktion des Abschnitts zu erkennen ist möglich, wird jedoch bei umfangreicheren Dateien äußerst schwer. Mithilfe des *Hopper* Dekompilierers kann die Main-Funktion auf zwei Zeilen reduziert werden. Bei einfachen Funktionen erfolgt die Vereinfachung ohne Probleme, bei komplexen Funktionen jedoch versagt der Dekompilierer und sorgt für eine weitaus verwirrendere Darstellung wie in Listing 1 zu sehen ist.

function main **{**

esp **=** **(**esp **&** 0xfffffff0**)** **-** 0x30**;**

eax **=** **\***arg1**;**

puts**(**eax**);**

eax **=** time**(**0x0**);**

srand**(**eax**);**

**\***rnd **=** rand**();**

calc\_crossfoot\_rec**(\***rnd**);**

stack**[**2035**]** **=** **\***rnd**;**

**if** **(**calc\_crossfoot**(**stack**[**2035**])** **==** **\*(**esp **+** 0x20**))** **goto** loc\_8049543**;**

**[…]**

**Listing 1: Mit Hopper dekompilierter Code einer komplexen Funktion**

Es werden für das Verständnis der Funktion irrelevante Abschnitte übersetzt, Registernamen werden für Variablen verwendet, es existieren keine Angaben zu Datentypen und bei *if*-Anweisungen wird anstatt mit geschweiften Klammern mit Sprunganweisungen (*goto*) und den Adressen des anzuspringenden Befehls gearbeitet. Beim Vergleich mit dem Disassembly der Main-Funktion fällt auf, dass lediglich jeder Maschinenbefehl in eine andere Darstellung konvertiert wurde, welche das Lesen des Codes jedoch nicht erleichtert.

Zwar bieten Disassemblierer weitere verschiedene Möglichkeiten, wie Kontrollflussgrafen oder Proximity Browser, die Datei zu untersuchen, die Schwierigkeit steigt jedoch proportional zur Größe der zu untersuchenden Software. In der Abbildung 2 ist beispielsweise die Kreisdarstellung des Proximity Browsers für eine ausführbare Datei mit ca. 300 Zeilen Quellcode zu sehen. Durch klicken auf einen der Kreise mit dem Pluszeichen expandiert weitere Knoten, wodurch der schon unübersichtliche Graf fast unlesbar wird.

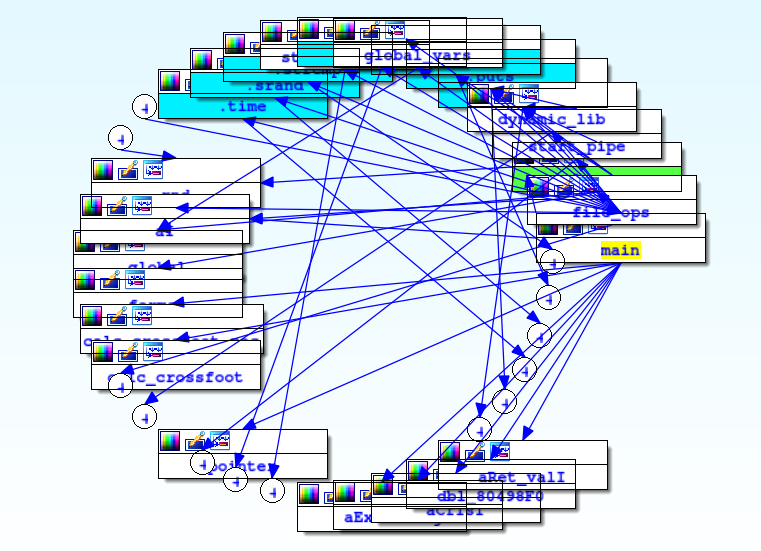


Abbildung 2 - Proximity Browser des IDA für eine ausführbare Datei mit ca. 300 Zeilen Quellcode

Die Werkzeuge können durchaus für große Projekte verwendet werden, erfordern jedoch viel Erfahrung und Zeit. Im Allgemeinen bedienen sich aktuelle Emulatoren den durch die Befehlssatzarchitekturen programmierten Informationen. Bei Assembleranweisung die durch einen Disassemblierer gewonnen wurden und nicht trivial sind, müssen diese Informationen ebenfalls aus den Befehlssatzarchitekturen herausgefunden werden. Um den Software Reverse Engineering Prozess effizienter zu gestalten, sollte das Erlangen dieser Informationen und die sichere Emulation mit für den Benutzer verständlicher Ausgabe automatisiert werden.

# Ziel

Im Vordergrund dieser Arbeit steht die Untersuchung der Möglichkeit Informationen über Maschinenbefehle für einen Emulator automatisch zu generieren und für eine Emulation einer beliebigen ausführbaren Binärdatei zu nutzen. Hierfür müssen vorab die benötigten Komponenten mit ihren Funktionen definiert werden. Anschließend muss festgelegt werden, wie und mit welchen Hilfsmitteln die Implementierung der Funktionen erfolgt.

Zuerst stellt sich die Frage, wie diese Daten ohne weitere Kenntnisse ermittelt werden können. Dies kann erreicht werden, indem auf der zu untersuchenden Maschine alle möglichen Befehle ausgeführt und die Zustände des Systems vor und nach dem jeweiligen Befehl aufgezeichnet und untersucht werden. Nach der Analyse muss die Auswertung für den Emulator aufbereitet und abgespeichert werden. Nach erfolgreicher Abarbeitung aller Befehle soll es möglich sein, eine beliebige ausführbare Binärdatei zu emulieren und jeden einzelnen Schritt nachvollziehen zu können. Im Optimalfall soll der Benutzer über eine Schnittstelle sowohl mit dem Emulator, als auch mit der emulierten Datei kommunizieren können.

Bei der Verwirklichung des Ziels treten verschiedene Probleme auf, die möglicherweise zeitaufwendig lösbar sind. Auch für die Geschwindigkeit des Sammelns der Daten, bzw. der Emulation einer Datei, muss auf eine effiziente Implementierung geachtet werden. Aufgrund der begrenzten Zeit wurde in dieser Arbeit die Priorität auf die Kernfunktionen des Programms gelegt. Genauere Beschreibungen der Probleme und Vernachlässigungen bei der Programmierung finden sich in den folgenden Kapiteln an den jeweiligen Stellen. Ebenso wurde bei der Programmierung nur auf einer 32-Bit x86-Architektur gearbeitet. Das Vorgehen sollte jedoch auf andere Maschinen übertragbar sein.

# Aufbau und Komponenten

## Befehl ausführen

Es gibt mehrere Möglichkeiten einen beliebigen Maschinenbefehl mithilfe des Bytecodes auszuführen. Um das Auftreten unerkannter Abhängigkeiten zu vermeiden, ist es in jedem Fall erforderlich, einen neuen, unberührten Prozess zu erzeugen und zu untersuchen. Wird der Befehl in ein und demselben Prozess ausgeführt und untersucht, so kann sich der Befehl aufgrund zuvor ausgeführter Befehle anders verhalten, als in einer neu initialisierten Prozessumgebung. Beispielsweise wenn ein Register durch eine zuvor ausgeführte nicht deterministische Funktion geändert wurde und von dem untersuchten Befehl verwendet wird, entsteht eine Abhängigkeit, die fast unmöglich zu erkennen ist. Erwähnenswert ist außerdem, dass bestimmte Befehle nur ausgeführt und bestimmte Register (wie die Debug-Register) nur geändert werden können, wenn das Programm sich im Kernel Level befindet. Da eine Realisierung im Kernel Level jedoch sehr zeitintensiv ist, wurde in dieser Arbeit darauf verzichtet.

### Bytecode durch Cast eines Funktionszeigers ausführen

Eine mögliche Lösung ist den Bytecode in einen ausführbaren Speicherbereich zu laden und anschließend auszuführen[[1]](#footnote-1).

/\* copy code to executable buffer \*/

void**\*** buf **=** mmap**(**0**, sizeof(**code**),** PROT\_READ**|**PROT\_WRITE**|**PROT\_EXEC**,**

MAP\_PRIVATE**|**MAP\_ANONYMOUS**, -**1**,** 0**);**

memcpy**(**buf**,** code**,** **sizeof(**code**));**

/\* run code \*/

**((**void **(\*)** **(**void**))**buf**)();**

**Listing 1: Beliebigen Bytecode ausführen**

Hier wird die Variable *code* vom Type *char[]* in einen ausführbaren Bereich kopiert und anschließend ausgeführt. Mit der Funktion *mmap*[[2]](#footnote-2) können Dateien oder Geräte in den dynamischen Speicher geladen werden. Mit den in Listing 1 verwendeten Argumenten gibt die Funktion einen Zeiger auf einen neu allokierten Speicherbereich mit der Länge des Strings zurück, welcher gelesen, beschrieben und ausgeführt werden kann. Die Adresse wird hierbei vom System bestimmt. In Zeile 4 wird mithilfe der Funktion *memcpy*[[3]](#footnote-3) der Bytecode in den Puffer kopiert und schließlich in Zeile 6 ausgeführt. Dies erfolgt durch einen Cast der Variablen *buf* auf einen Funktionszeiger ohne Rückgabewert und ohne Parameter und anschließendem Aufruf der Funktion durch Anhängen zwei runder Klammern.

Diese Vorgehensweise bringt jedoch mehrere Probleme mit sich. Auch wenn der Code in einem neuen Prozess ausgeführt werden würde, sind mehrere Funktionen zur Vorbereitung nötig, die das Ergebnis verfälschen können. Ein weiteres Problem ist der Funktionsaufruf eines dynamisch allokierten Speicherbereichs. Zum einen erfolgen bei einem Funktionsaufruf weitere Änderungen des Systems, die mit aufgezeichnet werden würde. Zum anderen ist die Adresse erst zur Laufzeit bekannt, wodurch Schwierigkeiten bei der Aufzeichnung entstehen können, da nicht klar ist an welchen Stellen der Prozess angehalten werden muss. Ebenfalls wird der Prozess in einen „Segmentation Fault“ hineinlaufen, falls kein Return-Befehl im Bytecode enthalten ist, welcher auch Änderungen nach sich zieht, die nicht zum eigentlich untersuchten Befehl gehören.

### Bytecode durch Modifikation einer Vorlage ausführen

Eine andere und deutlich einfachere Möglichkeit besteht darin, einen Prozess mithilfe einer ausführbaren Datei zu starten, in der der zu untersuchende Befehl einfach in der Main-Funktion hartkodiert vorzufinden ist. Die Adresse des Befehls ist bekannt, wodurch das Anhalten des Prozesses an den gewünschten Stellen vereinfacht wird. Folgendermaßen kann dabei vorgegangen werden: Es wird eine Programmschablone erstellt, indem eine C-Datei mit ausreichend „No Operation“-Befehlen erstellt und kompiliert wird.

int main**(**int argc**,** const char**\*** argv**[])** **{**

asm volatile**(**"nop"**);**

**}**

**Listing 2: Vorlage „template.c“**

Durch Inlineassembly wird in der Main-Funktion Platz geschaffen. Die Main-Funktion enthält (abgesehen von Speicherverwaltungsfunktionen) vorerst nur „No Operation“-Befehle, welche eine Länge von einem Byte haben. Um alle Befehle untersuchen zu können muss die Zeile 2 insgesamt 15-mal vorkommen, da dies der maximalen Länge eines x86-Maschinenbefehls entspricht.

080483eb **<**main**>:**

80483eb**:** 55 **push** %ebp

80483ec**:** 89 e5 **mov** %esp**,%ebp**

80483ee**:** 90 **nop**

...

80483fc**:** 90 **nop**

80483fd**:** b8 00 00 00 00 **mov** $0x0**,%eax**

8048402**:** 5d **pop** %ebp

8048403**:** c3 **ret**

**Listing 3: Disassembly der kompilierten Vorlage**

Nach der Kompilierung erhält man das Ergebnis in Listing 3. Hier wird die Start- und Endadresse des Blocks mit den „No Operation“-Befehlen benötigt, um den gewünschten Befehl in die Datei einzufügen und den Prozess an den entsprechenden Stellen zu stoppen. Hier kann in den Zeilen 4 und 7 die Adressen entnommen werden. Die Startadresse hat den Wert 0x80483ee und die Endadresse 0x80483fd. Die Differenz der beiden Adresse ergibt die erwarteten 15 Bytes.

Um einen beliebigen Befehl jetzt in einer blanken Main-Funktion auszuführen zu können, wird eine Funktion benötigt, welche die Datei entsprechend anpasst.

bool cpu\_x86\_32**::**build\_exe**()** **{**

ifstream templ**;**

ofstream exe**;**

templ**.**open**(**"./template"**,** ios**::**in **|** ios**::**ate **|** ios**::**binary**);**

exe**.**open**(**"./execute"**,** ios**::**out **|** ios**::**trunc **|** ios**::**binary**);**

**if(!**templ**.**is\_open**())** **{**

cerr **<<** "ERROR: Failed to open ./template!" **<<** endl**;**

**return** **false;**

**}**

**if(!**exe**.**is\_open**())** **{**

cerr **<<** "ERROR: Failed to open ./execute!" **<<** endl**;**

**return** **false;**

**}**

int size **=** templ**.**tellg**();**

templ**.**seekg**(**0**);**

char lead**[**INSTR\_POS**];**

templ**.**read**(**lead**,** **sizeof(**lead**));**

exe**.**write**(**lead**,** **sizeof(**lead**));**

templ**.**seekg**((**int**)**templ**.**tellg**()** **+** len**);**

exe**.**write**((**char**\*)**instr**,** len**);**

char trail**[**size **-** templ**.**tellg**()];**

templ**.**read**(**trail**,** **sizeof(**trail**));**

exe**.**write**(**trail**,** **sizeof(**trail**));**

templ**.**close**();**

exe**.**close**();**

**if(**chmod**(**"./execute"**,** S\_IRWXU

**|** S\_IRGRP

**|** S\_IXGRP

**|** S\_IROTH

**|** S\_IXOTH**)** **==** **-**1**)** **{** // Change file permissons to 755

cerr **<<** "ERROR: Failed to change file permissions!" **<<** endl**;**

**return** **false;**

**}**

**return** **true;**

**}**

**Listing 4: Funktion zur Anpassung der Vorlage[[4]](#footnote-4)**

In den Zeilen 2 bis 13 werden die benötigten Dateien vorbereitet, indem zwei Dateiströme erstellt werden[[5]](#footnote-5). Zum einen muss die zuvor erstellte Vorlage gelesen, zum anderen eine neue Datei erstellt, bzw. überschrieben, werden, welche den gewünschten Maschinenbefehl enthalten soll. Beide Dateien werden durch das Flag *ios::binary* byteweise bearbeitet und durch das Flag *ios::ate* wird der Lesezeiger der Vorlage an das Ende der Datei gesetzt. Dadurch kann in Zeile 15 die Größe der Datei bestimmt werden. In Zeile 16 wird der Lesezeiger wieder an den Anfang des Stroms gesetzt. Anschließend wird über den Puffer *lead* alles vor dem zu bearbeitenden Bereich in die neue Datei geschrieben. Das Define *INSTR\_POS* entspricht dem Offset der ersten „No Operation“-Anweisung der Main-Funktion, welcher äquivalent zu den ersten 12 Bits der Startadresse aus Listing 3 ist. Während in Zeile 23 der gewünschte Befehl *instr* in die neue Datei geschrieben wird, wird in Zeile 22 der Lesezeiger der Vorlage um die Länge *len* des Befehls weiterverschoben. In den Zeilen 25 bis 27 wird äquivalent zu den Zeilen 18 bis 20 der Rest der Vorlage in die neue ausführbare Datei übertragen. Nach dem Schließen der Dateien werden in Zeile 32 noch die Zugriffsrechte der neuen Datei „execute“ durch die Funktion *chmod*[[6]](#footnote-6) geändert, damit diese auch im weiteren Verlauf des Prozesses ausgeführt werden kann. Nach Abarbeitung der Funktion *build\_exe* kann nun ein neuer Prozess furch *fork* und *exec* kreiert werden.

Diese Lösung kommt dem Ziel nahe, jedoch ist bei einigen Befehlen eine Vorbereitung des Systems durch andere Befehle erforderlich, weil sie sonst einen „Segmentation Fault“ auslösen. Dadurch können benötigte Befehle nicht untersucht werden, weshalb die Datenbank für die spätere Emulation Lücken aufweisen würde. Um dieses Problem zu lösen, müssten Kombinationen zwischen zwei oder mehreren Befehlen erstellt werden und das Verhalten richtig erkannt werden. Ein stumpfes Kombinieren beliebiger Befehle oder Bytes kann jedoch zu so vielen Möglichkeiten führen, dass der Aufbau der Datenbank nicht in endlicher Zeit möglich ist. Durch Unterscheidung des zurückgegebenen Signalcodes des Prozesses lässt sich erkennen, ob der Prozess aufgrund einer Schutzverletzung oder einer illegalen Instruktion abgestürzt ist. So können zumindest nicht benötigte Bytecodefolgen aussortiert werden.

Ein Beispiel dieses Problems ist der Befehl *pushl -0x4(%ecx)* mit dem Bytecode 0xff71fc. Der Befehl legt den 32-Bit Wert an der Adresse im Register ECX mit dem Offset -0x4 auf dem Stack ab. Schnell wird klar, dass ein „Segmentation Fault“ bei einer ungültigen Adresse im Register ECX auftritt. Bei der Untersuchung des Bytecode 0xff zeigt der Disassembler an, dass es sich um einen nicht existenten Befehl handelt. Das führt dazu, dass der Prozess aufgrund einer „Illegal Instruction“ terminiert.

Dieses Verhalten kann auch in dem üblichen Fall eintreten, dass ein zu untersuchender Befehl kürzer als 15 Bytes ist. Hier kann dennoch durch Überprüfung des Befehlszeigers (EIP-Register) festgestellt werden, ob dennoch etwas ausgeführt wurde. Wie dies funktioniert und auf was zu achten ist wird im nächsten Kapitel deutlich.

## Verhalten aufzeichnen

Wurde nun ein Maschinenbefehl ausgeführt, so wurden an der Maschine Änderungen vorgenommen. Dabei handelt es sich in erster Linie um Änderungen an den Registern und dem Arbeitsspeicher. Um die Werte organisiert zu speichern und zu verwalten, bietet es sich an, eine Klasse zum Erzeugen von Objekten zu erstellen, welche das System und die jeweiligen Zustände abbilden. Zum Speichern der Werte ist es vorteilhaft, vorgefertigte Datenstrukturen wie Listen oder Hash Maps zu verwenden. Die Datei „cpu\_x86\_32.h“ beinhaltet eine mögliche Repräsentation einer CPU. Um die Struktur übersichtlicher zu gestalten, wurde die Verwaltung der Register in die Datei „register\_x86.h“ ausgelagert.

Um alle Zustände der Maschine zu speichern, weisen die Variable für die Register *regs* und die Variable für den Arbeitsspeicher *ram* eine Listenstruktur auf. Ein Zustand des Arbeitsspeichers wird mit einer Mapstruktur abgebildet. Bei dem Schlüssel handelt es sich um einen vorzeichenlosen 32-Bit Integerwert, um den kompletten Speicherbereich abzubilden. Der zugehörige Wert wird ebenfalls durch einen vorzeichenlosen Integerwert dargestellt. Dieser benötigt jedoch lediglich acht Bits, bzw. ein Byte, um ein Zeichen im Arbeitsspeicher darzustellen. Ein Zustand der Register wird durch ein Objekt der Klasse *register\_x86* abgebildet. Diese Klasse enthält neben Verwaltungsfunktionen eine Klassenvariable *regs*, welche ebenso eine Mapstruktur aufweist. Hier wird ein String, welcher den Namen des Registers enthält, auf einen vorzeichenlosen 32-Bit Wert abgebildet, da die Größe der Register einer 32-Bit Maschine vier Byte entspricht und die Variablen in der verwendeten Struktur zum herauslesen der Registerwerte ebenfalls diese Größe aufweisen.

Um einen Prozess zu untersuchen, existieren erneut mehrere Möglichkeiten. Der im Programm verwendete Vorgang erfolgt mithilfe des Systemaufrufs *ptrace*, welches auch in Debuggern wie *gdb* Verwendung findet[[7]](#footnote-7). Eine andere Möglichkeit besteht darin, direkt einen Debugger zu verwenden. Zum Beispiel bietet der *gdb*-Debugger die Option mit einer Maschinenschnittstelle gestartet zu werden. Im Vergleich mit *ptrace* bietet der Debugger deutlich mehr Funktionalität bei der Untersuchung eines Prozesses, unter anderem weil Systeme über eine externe Debugschnittstelle von anderen Maschinen aus erreicht und untersucht werden können.

Eine weitere jedoch komplizierte und aufwendige Möglichkeit die Register und Arbeitsspeicherwerte herauszulesen, kann durch Inlineassembleranweisungen in der Vorlage der zu untersuchenden, ausführbaren Datei realisiert werden. Die Register werden durch Assembleranweisungen, wie beispielsweise *asm volatile("vmovdqu %%ymm0, %0" : "=g" (out))*[[8]](#footnote-8)*,* auf Variablen abgebildet, welche anschließen abgespeichert werden können. Problematisch wird es jedoch, wenn sich bei Assembleranweisungen Werte in Registern oder im Arbeitsspeicher ändern. Da bessere Alternativen existieren, die Realisierung dieses Vorgangs äußerst zeitaufwendig und der Ablauf möglicherweise unvorhersehbar ist, wird auf weitere Untersuchung der Möglichkeit verzichtet.

### Ptrace[[9]](#footnote-9)

Ptrace erlaubt einem Prozess, dem sogenannten Tracer, einen anderen Prozess, den sogenannten Tracee, zu steuern, zu beobachten und zu modifizieren. Die Initialisierung der Verbindung zwischen den beiden Prozessen kann sowohl durch den Tracer, als auch den Tracee erfolgen. Dafür ist ein Aufruf der Funktion *ptrace*[[10]](#footnote-10) mit dem entsprechenden Parameter erforderlich. Die Signatur der Funktion sieht wie folgt aus:

**long ptrace(enum \_\_ptrace\_request** *request***, pid\_t** *pid***, void \****addr***, void \****data***);**

Der erste Parameter *request* erwartet ein Enum, durch welches die auszuführende Aktion bestimmt wird. Durch den Parameter *pid* wird der Prozess gewählt, auf den die Aktion ausgeführt werden soll. Die Aufgabe der anderen Parameter, sowie des Rückgabewerts, richtet sich nach dem übergebenen Enumwert.

Da das Programm zum Sammeln von Daten gezielt einen neuen Kindprozess mit dem *fork*[[11]](#footnote-11) Systemaufruf startet, kann durch einen Funktionsaufruf von *ptrace* mit dem Requestwert *PTRACE\_TRACEME* im Kindprozess die Beobachtung eingeleitet werden. Bei einem anschließenden Aufruf einer Funktion der *exec*[[12]](#footnote-12)-Familie stoppt der Tracee automatisch und wartet auf ein Signal des Tracers um fortzufahren. Eine mögliche Implementierung des Ablaufs ist in Listing 5 zu sehen.

int wstatus**;**

pid **=** fork**();**

**if(**pid **==** 0**)** **{**

ptrace**(**PTRACE\_TRACEME**,** 0**,** **nullptr,** **nullptr);**

**if** **(**execlp**(**"./execute"**,** "./execute"**)** **<** 0**)** **{**

cerr **<<** "ERROR: Execution of file \"execute\" failed!" **<<** endl**;**

**return** **false;**

**}**

**}**

waitpid**(**pid**,** **&**wstatus**,** 0**);**

**if(**wstatus **>>** 8 **!=** SIGTRAP**)** **{**

cerr **<<** "ERROR: Failed to catch exec!" **<<** endl**;**

**return** **false;**

**}**

ptrace**(**PTRACE\_SETOPTIONS**,** pid**,** **nullptr,** PTRACE\_O\_EXITKILL**);**

**Listing 5: Initialisierung der Ptrace-Verbindung[[13]](#footnote-13)**

In Zeile 2 spaltet sich ein Kindprozess vom aktuellen Prozess ab. Der Kindprozess initialisiert sich im darauffolgenden *if*-Block selbst für die *ptrace*-Verbindung. Die Parameter der *ptrace*-Funktion, ausgenommen des Requestparameters, werden ignoriert, weshalb sie den Wert null erhalten. In Zeile 5 wird die aus der Vorlage erstellte, ausführbare Datei in den Prozess geladen. Vor dem Ausführen der Datei stoppt der Prozess und wartet auf den Elternprozess. Dieser erkennt in den Zeilen 10-14 durch die Funktion *waitpid*[[14]](#footnote-14) und durch Überprüfen des Statusflags *wstatus*, dass der Tracee pausiert und bereit ist. In Zeile 15 wird noch zur Absicherung die Option *EXITKILL* gesetzt, sodass beim Beenden des Tracers auch der Tracee beendet wird.

Nach der Initialisierung besteht die nächste Aufgabe darin, die zu untersuchende Stelle in der Main-Funktion zu finden. Dies ist durch die aus der Vorlage bestimmten Speicheradressen möglich. Die Speicheradressen sind in dem Fall durch die Defines *BRACE\_OPEN* und *BRACE\_CLOSE* hartkodiert. Weil bei beiden Adressen die gleichen Funktionen ausgeführt werden müssen, bietet es sich an, die beiden Fälle zusammenzufassen.

**for(**int i **=** 0**;** i **<** 2**;** i**++)** **{**

uint32\_t brace **=** 0**;**

**switch(**i**)** **{**

**case** 0**:** brace **=** BRACE\_OPEN**;** **break;**

**case** 1**:** brace **=** BRACE\_CLOSE**;** **break;**

**}**

**Listing 6: Zusammenfassung der Funktionen durch Schleife mit zwei Fällen[[15]](#footnote-15)**

In Listing 6 ist zu sehen, dass die Schleife zwei Durchgänge erzeugt. Der erste Durchgang wird mit der Startadresse und der zweite mit der Endadresse durchlaufen. Die folgende *while*-Schleife in Listing 7 lässt den Tracee bis zu der gewünschten Adresse mit der Ausführung fortfahren. Zusätzlich müssen mögliche Fehlerquellen durch *if*-Anweisungen abgefangen werden.

uint32\_t eip\_tmp **=** 0**;**

**while(true)** **{**

uint32\_t eip **=** user\_reg**(**PTRACE\_PEEKUSER**,** EIP**,** 0**);**

**if(**eip **==** brace**)** **{**

**break;**

**}** **else** **if(**brace **==** BRACE\_CLOSE **&&** **(**eip **>** BRACE\_CLOSE **||** eip **<** eip\_tmp**))** **{**

**break;**

**}** **else** **if(**eip **==** 0**)** **{**

**return** **false;**

**}** **else** **if(**wstatus **>>** 8 **!=** SIGTRAP**)** **{**

cout **<<** hex **<<** "Terminated with Signal Code 0x" **<<** **(**wstatus **>>** 8**)**

**<<** " in address 0x" **<<** eip **<<** "!" **<<** endl**;**

len **=** eip **-** BRACE\_OPEN**;**

**if(**len **<=** 0**)** **{**

cerr **<<** "ERROR: Length of instruction is 0!" **<<** endl**;**

**return** **false;**

**}**

**if(!**user\_reg**(**PTRACE\_POKEUSER**,**

EFLAGS**,**

user\_reg**(**PTRACE\_PEEKUSER**,** EFLAGS**,** 0**)**

**&** **~(**1 **<<** RESUME\_FLAG**)))** **{**

cerr **<<** "Failed to change EFLAGS register!" **<<** endl**;**

**return** **false;**

**}**

**break;**

**}**

ptrace**(**PTRACE\_SINGLESTEP**,** pid**,** **nullptr,** **nullptr);**

waitpid**(**pid**,** **&**wstatus**,** 0**);**

eip\_tmp **=** eip**;**

**}**

**Listing 7: *While*-Schleife zum Erreichen der Zieladressen[[16]](#footnote-16)**

Durch die geschützte Klassenmethode *user\_reg* in Zeile 3 wird die aktuelle Adresse Befehlszeigers mithilfe des *ptrace*-Systemaufrufs bestimmt. Falls keine der *if*-Anweisungen greift, also die Adresse nicht erreicht und kein Fehler gefunden wurde, so wird in den Zeilen 27-29 der nächste Maschinenbefehl durch den *ptrace*-Request *PTRACE\_SINGLESTEP* vom Tracee ausgeführt. Zeile 29 sichert den Befehlszeiger zur Fehlererkennung im nächsten Durchlauf der *while*-Schleife. Die erste *if*-Anweisung in Zeile 4 prüft, ob der Befehlszeiger die gewünschte Adresse erreicht hat. In Zeile 6 wird der Fall abgefangen, dass gerade ein Maschinenbefehl untersucht wird, der den Befehlszeiger ändert, wie beispielsweise eine Sprunganweisung. Zuerst wird geprüft, ob die Startadresse bereits passiert wurde, anschließend ob die aktuelle Adresse im EIP-Register einen Wert außerhalb des zu untersuchenden Adressbereichs angenommen hat. In diesen beiden Fällen wird die *while*-Schleife mit einem *break* verlassen, um die Registerwerte und den Arbeitsspeicher auszulesen. Zeile 8-9 bricht den Vorgang durch ein *return* mit dem Bool‘schen Rückgabewert *false* ab, falls die Funktion *user\_reg* den Wert null für den Befehlszeiger zurückgibt. Dies deutet auf einen Fehler hin. Eine entsprechende Fehlermeldung wird in der Funktion *user\_reg* ausgegeben. Der letzte Fall in Zeile 10 der eintreten kann, behandelt eine Terminierung des Prozesses mit einem Signalcode, der nicht einen gewöhnlichen Stopp des Prozesses indiziert. Hier wird in den Zeilen 11 und 12 eine Meldung auf der Konsole ausgegeben. Anschließend wird in Zeile 13 die Länge des ausgeführten Bytecodes ermittelt, indem die Differenz zwischen der Startadresse und dem aktuellen Wert des Befehlszeigers berechnet wird. Ist die Länge null, so wurde kein Befehl erfolgreich ausgeführt. In diesem Fall wird derzeit auf eine weitere Analyse des Bytecodes verzichtet. Es kann sich jedoch um einen gültigen Befehl handeln, falls sich der Signalcode vom Define *SIGILL*[[17]](#footnote-17) („Illegal Instruction“) unterscheidet. Dieser Befehl muss zur Untersuchung gegebenenfalls durch andere Befehle vorbereitet werden. Ein Beispiel ist am Ende des vorherigen Teilkapitels „Befehl ausführen“ aufgeführt. Wurde trotzdem festgestellt, dass ein Teil des Bytecodes ausgeführt wurde, so wird in den Zeilen 18-21 das *RESUME\_FLAG* im EFLAGS-Register zurückgesetzt, welches eine fehlerhafte Terminierung eines Prozesses anzeigt. Die Änderung erfolgt durch den *ptrace*-Request *POKE\_USER*. Anschließend wird wie bei den ersten beiden eindeutigen Fällen durch ein *break* mit dem Lesen der Register und des Arbeitsspeichers fortgefahren.

Das Lesen und Speichern der Registerwerte und des Arbeitsspeichers erfolgt durch den Aufruf der beiden geschützten Klassenmethoden *read\_regs* und *read\_ram*. Nach erfolgreichem Ausführen der beiden Methoden wird der Kindprozess mit dem *ptrace*-Request *PTRACE\_KILL* explizit beendet, um Speicherbereich freizugeben.

Die Funktion *read\_regs* verwendet die *ptrace*-Requests *PTRACE\_GETREGS*, *PTRACE\_GETFPREGS* und *PTRACE\_GETFPXREGS*. Als Rückgabeparameter erwartet die *ptrace*-Funktion eine Variable der Struktur *user\_regs\_struct*, *user\_fpregs\_struct*, bzw. *user\_fpxregs\_struct*[[18]](#footnote-18). Beim ersten Request handelt es sich um die „User“-Register, beim zweiten um die „Floating Point“-Register und beim dritten um „Floating Point“-Register mit Streaming SIMD Extensions (also den XMM-Registern). Weil die dritte Struktur alle Werte der zweiten Struktur enthält, kann im Allgemeinen auf diese verzichtet werden. Nachdem die Strukturen durch den *ptrace*-Funktionsaufruf gefüllt wurden, wird ein neues Objekt vom Typ *register\_x86* erstellt, welches die Werte in einer geeigneten Mapstruktur abspeichert. Anschließend wird das Objekt an die Liste *reg* (Klassenvariable der Klasse *cpu\_x86\_32*) angehängt. Ein Nachteil von *ptrace* gegenüber *gdb* ist die unvollständige Struktur *user\_fpxregs\_struct*. Diese umfasst zwar die XMM-Register, es fehlen jedoch die oberen 128 Bits der YMM-Register[[19]](#footnote-19).

Das Auslesen des Arbeitsspeichers gestaltet sich etwas schwieriger, weshalb die Funktion *read\_ram* anhand des Quellcodes in Listing 8 erklärt wird.

bool cpu\_x86\_32**::**read\_ram**()** **{**

ifstream maps\_f**;**

maps\_f**.**open**(**"/proc/" **+** to\_string**(**pid**)** **+** "/maps"**,** ios**::**in**);**

**if(!**maps\_f**.**is\_open**())** **{**

cerr **<<** "ERROR: Failed to open \"maps\"-File \" of process "

**<<** pid **<<** "!" **<<** endl**;**

**return** **false;**

**}**

vector**<**pair**<**uint32\_t**,** uint32\_t**>>** limits\_l**;**

**for(**string line**;** getline**(**maps\_f**,** line**);** **)** **{**

limits\_l**.**push\_back**(**make\_pair**(**stoul**(**line**.**substr**(**0**,** 8**),** **nullptr,** 16**),**

stoul**(**line**.**substr**(**9**,** 8**),** **nullptr,** 16**)));**

**}**

maps\_f**.**close**();**

map**<**uint32\_t**,** uint8\_t**>** map\_m**;**

**for(**pair**<**uint32\_t**,** uint32\_t**>** limits **:** limits\_l**)** **{**

uint32\_t addr **=** limits**.**first**;**

**while(**addr **<** limits**.**second**)** **{**

long val **=** ptrace**(**PTRACE\_PEEKDATA**,** pid**,** addr**,** **nullptr);**

**if(**val **!=** **-**1 **&&** **!**errno**)** **{**

map\_m**.**insert**(**make\_pair**(**addr**,** val**));**

**}** **else** **if(**log**)** **{**

cerr **<<** hex **<<** "Failed to fetch address " **<<** addr **<<** endl**;**

**}**

addr**++;**

**}**

**}**

ram**.**push\_back**(**map\_m**);**

**return** **true;**

**}**

**Listing 8: Funktion *read\_ram* zum Auslesen des Arbeitsspeichers[[20]](#footnote-20)**

Um ein Auslesen des gesamten 4-Gigabyte-Adressraums einer 32-Bit Maschine zu umgehen, müssen die benutzten virtuellen Speicheradressen bekannt sein. Das Mapping der virtuellen Speicheradressen lässt sich aus der Datei „maps“ im *proc*-Dateisystem eines jeden Prozesses herauslesen. Besagte Datei wird in Zeile 3 als Eingangsdateistrom geöffnet. In Zeile 9 wird eine Liste angelegt, welche Paare von zwei Zahlen vom Typ *uint32\_t* enthält. Jedes dieser Paare stellt einen verwendeten Speicherbereich mit einer vorzeichenlosen 32-Bit Start- und Endadresse dar. In den Zeilen 10-13 wird die Datei Zeile für Zeile gelesen. Aus der gelesenen Zeile *line* werden mit der Methode *substr* die geforderten Zahlen extrahiert und durch die Funktion *stoul*[[21]](#footnote-21) in eine vorzeichenlose 32-Bit Zahl umgewandelt. Da die Adresse eine hexadezimale Schreibweise aufweist, muss die Basis der zu konvertierenden Zahl in der Funktion als letzter Parameter angegeben werden. Nachdem das Mapping vollständig ausgelesen wurde, wird der Dateistrom in Zeile 14 geschlossen.

Mit den verwendeten Adressbereichen kann das Auslesen des Arbeitsspeichers erfolgen. In Zeile 16 wird dafür der Behälter (Map mit der Adresse als Schlüssel und dem zugeordneten Zeichen als dazugehörigen Wert) initialisiert. In der darauffolgenden Zeile wird mithilfe einer *for-each*-Schleife über die Liste mit den Adresse-Zeichen-Paaren, welche die Adressbereiche beinhalten, iteriert. Mit einer *while*-Schleife kann durch einen *ptrace*-Funktionsaufruf mit dem Request *PTRACE\_PEEKDATA* der Wert an jeder Adresse des Adressbereichs gelesen werden. Nach erfolgreichem Lesen einer Adresse wird die Adresse mit dem Wert in die Hash Map eingefügt. Zum Schluss wird die vollständige Hash Map *map\_m* an die Liste *ram* (Klassenvariable von *cpu\_x86\_32*) angehängt.

Das Auslesen und die Organisation lässt sich effizienter gestalten. Weil die *ptrace*-Funktion laut Signatur einen 32-Bit Wert zurückgibt, beinhaltet die Variable *val* in Zeile 20 auch die Zeichen der nächsten drei Adressen. Diese werden aufgrund der impliziten Konvertierung in Zeile 22 abgeschnitten. Mit diesem Wissen kann in Zeile 26 die Variable *addr* um den Wert 4 erhöht werden. Das Speichern in der Hash Map kann durch bitweise Verschiebung und Konvertierung erfolgen. Eine weitere mögliche Optimierung der Organisation ließe sich erreichen, wenn der Wert der Hash Map ebenfalls ein 32-Bit Datentyp aufweisen würde.

### GDB/MI[[22]](#footnote-22)

Wie am Anfang des Kapitels beschrieben, bietet der *gdb*-Debugger, welcher auch *ptrace* verwendet, deutlich mehr Funktionalität. Um den Debugger zu starten und mit diesem zu kommunizieren, kann ein Pseudoterminal benutzt werden. Durch das Pseudoterminal werden Dateideskriptoren für beide Prozesse, den Prozess für das Sammeln der Daten und den Debugger-Prozess, erstellt, mit welchen eine Kommunikation über *stdout* und *stdin* des Kindprozesses stattfinden kann. Eine mögliche Implementierung eines Pseudoterminals ist *openpty*[[23]](#footnote-23). Mit der Funktion *forkpty*[[24]](#footnote-24) und den richtigen Parametern kann der *gdb*-Prozess gestartet und mit ihm kommuniziert werden. Beim Aufruf muss als Argument über das Flag *-i* die Schnittstelle *mi* („Machine Interface“) angegeben werden. Die Schnittstelle sorgt für eine für Maschinen bessere Aus- und Eingabe. Zum Beispiel wird bei der Abfrage der Registerwerte ein Query im JSON-Format ausgegeben, welches sich leicht parsen und weiterverarbeiten lässt. Im Gegensatz zu *ptrace* werden auch die vollständigen Werte der YMM-Register auslesbar.

Der Ablauf zum Untersuchen des Maschinenbefehls unterscheidet sich nicht von der Variante mit *ptrace*, er gestaltet sich lediglich etwas einfacher. Zuerst muss der Prozess an der Startadresse des Maschinenbefehls angehalten werden. Anschließend müssen die Register und der Arbeitsspeicher ausgelesen werden. Danach wird der Maschinenbefehl so weit wie möglich ausgeführt. Zum Schluss werden erneut die Register und der Arbeitsspeicher ausgelesen, um die Änderungen feststellen zu können. Im Listing 9 ist eine beispielhafte Kommunikation zwischen den beiden Prozessen zu sehen.

(gdb)

**-break-insert \*0x80483ee**

^done,bkpt={number="1",type="breakpoint",disp="keep",enabled="y",addr="0x080483ee",func="main(int, char const\*\*)",file="template.c",fullname="/home/bachelor/Bachelor/emulator/template.c",line="2",thread-groups=["i1"],times="0",original-location="\*0x80483ee"}

(gdb)

**-exec-run**

=thread-group-started,id="i1",pid="21391"

=thread-created,id="1",group-id="i1"

^running

\*running,thread-id="all"

[…]

(gdb)

**-data-list-register-names**

^done,register-names=["eax","ecx","edx","ebx","esp","ebp","esi","edi","eip","eflags","cs","ss","ds","es","fs","gs","st0","st1","st2","st3","st4","st5","st6","st7","fctrl","fstat","ftag","fiseg","fioff","foseg","fooff","fop","xmm0","xmm1","xmm2","xmm3","xmm4","xmm5","xmm6","xmm7","mxcsr","ymm0h","ymm1h","ymm2h","ymm3h","ymm4h","ymm5h","ymm6h","ymm7h","","","","","","","orig\_eax","al","cl","dl","bl","ah","ch","dh","bh","ax","cx","dx","bx","","bp","si","di","ymm0","ymm1","ymm2","ymm3","ymm4","ymm5","ymm6","ymm7","mm0","mm1","mm2","mm3","mm4","mm5","mm6","mm7"]

(gdb)

**-data-list-register-values x 8**

^done,register-values=[{number="8",value="0x80483ee"}]

(gdb)

**-data-list-register-values x**

^done,register-values=[{number="0",value="0x1"},{number="1",value="0xcf16efe9"}, ... {number="86",value="{uint64 = 0x0, v2\_int32 = {0x0, 0x0}, v4\_int16 = {0x0, 0x0, 0x0, 0x0}, v8\_int8 = {0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0}}"},{number="87",value="{uint64 = 0x0, v2\_int32 = {0x0, 0x0}, v4\_int16 = {0x0, 0x0, 0x0, 0x0}, v8\_int8 = {0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0, 0x0}}"}]

(gdb)

**-data-read-memory-bytes 0x08048000 4**

^done,memory=[{begin="0x08048000",offset="0x00000000",end="0x08048004",contents="7f454c46"}]

(gdb)

**-exec-step-instruction**

^running

\*running,thread-id="1"

(gdb)

\*stopped,reason="end-stepping-range",frame={addr="0x080483ef",func="main",args=[{name="argc",value="1"},{name="argv",value="0xffffd164"}],file="template.c",fullname="/home/bachelor/Bachelor/emulator/template.c",line="3"},thread-id="1",stopped-threads="all",core="0"

(gdb)

**-data-list-register-values x 8**

^done,register-values=[{number="8",value="0x80483ef"}]

(gdb)

**-exec-step-instruction**

^running

\*running,thread-id="all"

(gdb)

\*stopped,reason="signal-received",signal-name="SIGSEGV",signal-meaning="Segmentation fault",frame={addr="0x080483ef",func="main",args=[{name="argc",value="1"},{name="argv",value="0xffffd164"}],file="template.c",fullname="/home/bachelor/Bachelor/emulator/template.c",line="3"},thread-id="1",stopped-threads="all",core="0"

(gdb)

[…]

**Listing 9: Eine Debug-Sitzung mit dem Machine Interface[[25]](#footnote-25)**

Aufgrund vieler Informationen in der Ausgabe ist es schwer die Übersicht zu behalten, weshalb das Protokoll an einigen Stellen gekürzt wurde. Die entsprechenden Stellen sind mit drei aufeinanderfolgenden Punkten markiert. Anders als bei *ptrace* kann sofort, wie in Zeile 2 zu sehen, ein Breakpoint für die gewünschte Adresse angelegt werden. Durch den Befehl *-exec-run* wird der Prozess bis zu der Adresse ausgeführt. Mit dem Befehl in Zeile 15 *-data-list-register-names* können die Namen der Register ausgegeben werden. Die Reihenfolge der Registername in der Ausgabe entspricht der Nummerierung der Register. So wird in Zeile 24 durch *-data-list-register-values x 8* der Wert des an der achten Stelle befindlichen EIP-Registers ausgegeben, welcher dem Wert der Startadresse entspricht. Wird der gleiche Befehl, wie in Zeile 27, ohne zweiten Parameter verwendet, werden alle verfügbaren Register formatiert ausgegeben. Durch den Befehl *-data-read-memory-bytes* kann der Arbeitsspeicher ausgelesen werden. Der erste Parameter gibt in dem Fall die Adresse an, der zweite Parameter die Länge der zu lesenden Zeichen. Dadurch kann auf einen Schlag der gesamte Adressbereich ausgelesen werden. Der Befehl in Zeile 37 *-exec-step-instruction* entspricht dem *ptrace*-Request *PTRACE\_SINGLESTEP*. Das Auslesen des EIP-Registers in Zeile 46 indiziert, dass ein Byte ausgeführt wurde. Auch in den Zeile 42 und 43 kann sowohl der Grund des Haltens des Prozesses sowie der Wert des aktuellen Befehlszeigers entnommen werden. Nach erneuter Einzelschrittausführung eines Maschinenbefehls in Zeile 49 terminiert der Prozess aufgrund eines „Segmentation Faults“, wie der Zeile 53 zu entnehmen ist. Anschließend können mit den üblichen Mitteln erneut die Register und der Arbeitsspeicher ausgelesen werden.

Genauso wie bei *ptrace* muss hier im Falle eines Sprungbefehls das EIP-Register nach Erreichen der Startadresse in jedem Einzelschritt überprüft werden. Neben dem Auslesen aller Register kann auch der Befehl *-data-list-changed-registers* in Betracht gezogen werden, um nur die Registerwerte aufzulisten, die sich verändert haben. Für den Anfang und um Bytecode, der in einem „Segmentation Fault“ landet, zu untersuchen, ist es jedoch sinnvoll, alle Registerwerte aufzunehmen.

## Verhalten analysieren

Es gibt verschiedene Befehlstypen, die zuverlässig erkannt werden müssen. Um jeden Befehl kategorisieren zu können, müsste er mehrfach mit unterschiedlichen vorangestellten Befehlen untersucht werden, da die Abhängigkeit beispielsweise von Registerwerten nicht bei erstem Mal erkannt werden kann. Die Vorbereitung des Befehls muss deshalb gründlich überdacht werden und möglichst alle Fälle abdecken. Grundlegende Maschinenbefehle können in folgende Kategorien unterteilt werden: arithmetische Operationen, logische Operationen, bitorientierte Operationen, Speicheroperationen, Vergleichsoperationen, Steueroperationen, Datenkonvertierungsoperationen. Einige Befehle treten auch als eine Kombination mehrerer Kategorien auf[[26]](#footnote-26). Auch die vom jeweiligen Befehl verwendete Adressierungsform muss zuverlässig erkannt werden. Die automatische Kategorisierung erweist sich als äußerst schwer und ist wahrscheinlich am besten mit Vorgehensweisen aus der künstlichen Intelligenz, wie zum Beispiel Entscheidungsbäumen, lösbar.

In dieser Arbeit kommt die tatsächliche Analyse eines Maschinenbefehls aufgrund der Schwierigkeit dieser Aufgabe und der begrenzten Zeit nicht ausführlich vor. Es wird im Weiteren davon ausgegangen, dass es sich bei jedem Befehl um eine arithmetische Operation als Addition oder Subtraktion handelt.

Eine Addition oder Subtraktion zu erkennen ist mit einfachen Mitteln möglich. Ändert sich ein Register oder ein Speicherbereich kann durch die Bildung der Differenz des geänderten Wertes und des vorherigen Wertes der Summand, bzw. der Subtrahend, bestimmt werden. Zusätzlich muss erkannt werden, ob die Operation ein Register oder den Arbeitsspeicher als Ziel hat und ob sie mit einem hartkodierten Wert oder einem Wert im Speicherbereich des Prozesses erfolgt. Erfolgt die Operation im Arbeitsspeicher, so muss der Wert des Stapelzeigers berücksichtigt werden, da sich dieser bei jedem neu erzeugten Prozess eines Programmes unterscheiden kann. Falls kein hartkodierter Wert benutzt wird, ist die Erkennung der verwendeten Speicheradresse, bzw. des Registers, erforderlich. Wird der Wert dem Arbeitsspeicher entnommen, muss darauf geachtet werden, ob die verwendete Speicheradresse absolut oder relativ zu einem festen oder variablen Wert ist.

Alle diese Informationen müssen für einen effizienten Emulator erkannt, aufbereitet und geeignet abgespeichert werden. Das Erkennen weiterer benötigter Informationen und das Erstellen einer geeigneten Datenbank ist in diesem Kontext ein kontinuierlicher Prozess, weil bei komplexen Prozessoren mit vielen verschiedenen Befehlen Verhaltensweisen existieren, die nicht immer eindeutig sind. Deshalb wird in dem für diese Arbeit erstellten Programm lediglich der Unterschied der Zustände vor und nach dem untersuchten Maschinenbefehl abgespeichert. Der Versuch, einen kompletten Dump der CPU-nachbildenden Objekte mit allen Informationen und Zuständen zu sichern, führte schnell zu einem vollen Festplattenspeicher, weshalb diese Option verworfen wurde.

## Auswertung speichern und laden

### Auswertung speichern

Die nächste Frage besteht darin, wie die Datensätze dargestellt und gespeichert werden. Da die Art des Speicherns, bzw. des Speichermediums, (z.B. *SQL*, Textdatei) und des Formats (z.B. Trennzeichen, *JSON*) nur schnittstellenentscheidend und deshalb nicht relevant für die Kernfunktion des Programms ist, wird hier nicht weiter darauf eingegangen. In der programmatischen Ausarbeitung werden, wie im vorigen Teilkapitel „Verhalten analysieren“ beschrieben, nur die Unterschiede zwischen den Zuständen gesichert. Hierbei gehen entscheidende Informationen, wie zum Beispiels ein verwendeter Registerwert, verloren. Um die Informationen zu erkennen und zu behalten, müsste vor dem Speichern das CPU-emulierende Objekt intensiver untersucht werden. Für das Ziel dieser Bachelorarbeit, unter der Prämisse, dass nur arithmetische Befehle untersucht werden, reicht es, die Änderung der Werte passend zu speichern. Die Speicherung erfolgt dabei in üblichen Textdateien mit bestimmter Syntax.

**$ cat ff75f4**

EIP: 0x080483f0 1-> 0x080483f3

ESP: 0xffae65e8 1-> 0xffae65e4

0xffae65e5: 0x00 1-> 0x70

0xffae65e6: 0x00 1-> 0x71

0xffae65e7: 0x00 1-> 0xf7

**Listing 10: Informationen des Befehls *pushl -0xc(%ebp)* mit dem Bytecode 0xff75f4**

Im Listing 10 ist der Inhalt der Datei „ff75f4“ zu sehen. Im verwendeten Datenmodell erhalten die Dateien als Namen den Bytecode, über welchen sie die Informationen enthalten. In der Datei sind im ersten Block alle Register und im zweiten alle Speicheradressen aufgelistet, die sich nach Ausführung des Prozesses geändert haben. Die Blöcke sind dabei durch eine freie Zeile getrennt. Der Anfang jeder Zeile enthält den Namen des Registers, bzw. den Wert der Speicheradresse, gefolgt von einem Doppelpunkt. Anschließend folgt der Anfangswert des betrachteten Speicherelements. Darauf folgt jeweils ein Pfeil „->“ mit dem neuen Wert. Die Notierung der Werte und Adressen erfolgt in hexadezimaler Schreibweise mit dem Präfix „0x“ und zwischen allen Syntaxelementen befindet sich ein Tabulatorzeichen. Die Zahl vor dem Pfeil gibt an, in welchem Schritt die Änderung erfolgt. Da im vorgestellten Teil des Programms bisher immer nur ein Maschinenbefehl untersucht wurde, beträgt der Wert dieser Angabe immer 1 und ist somit vorerst irrelevant. Sollen jedoch wie beabsichtigt mehrere Maschinenbefehle kombiniert untersucht werden, so ist diese Angabe äußerst wichtig. Außerdem wurde diese Darstellung für die Kompatibilität zwischen Datensammler und Emulator gewählt, da bei der Emulation der Benutzer dadurch nachvollziehen kann, nach welchem Maschinenbefehl die Änderung stattgefunden hat.

Betrachtet man die gespeicherten Daten des Bytecodes und vergleicht diese mit den Informationen, die für den entsprechenden Assemblerbefehl *pushl -0xc(%ebp)* benötigt werden, so wird das zuvor erwähnte Problem klar. Mit dem Befehl wird ein 32-Bit Wert aus dem Arbeitsspeicher auf den Stapel gelegt. Die Speicheradresse des Werts im Arbeitsspeicher wird in diesem Fall mithilfe des EBP-Registers („Base Pointer“) bestimmt. Ohne die Information scheint der Befehl lediglich den festen Wert 0xf7717000 auf dem Stack abzulegen. Interpretiert man diesen Befehl arithmetisch, so wird der Befehlszeiger mit der Länge des Befehls 3 summiert und der Stapelzeiger um die Größe eines Wertes vom Typ *long* 4 erniedrigt. Bei den geänderten Adressen relativ zum Stapelzeiger wird jeweils der Endwert aufsummiert.

Die entsprechende Funktion zum Speichern der Änderungen befindet sich in der Datei „cpu\_x86\_32.cpp“ und lautet *dump\_change*. In dieser wird, falls noch keine Datei für den untersuchten Maschinenbefehl existiert, eine neue Datei angelegt und durch die Hilfsklassenmethoden *diff\_ram* der Klasse *cpu\_x86\_32* und *diff\_change* der Klasse *register\_x86* mit den Informationen in der angegebenen Syntax gefüllt.

### Auswertung laden

Für die Emulation ist es erforderlich, die gespeicherten Daten entsprechend wieder laden und verwenden zu können. Beim Laden müssen, abhängig vom Typ des Maschinenbefehls und dem aktuellen Zustand des Objekts zur Abbildung der CPU, unterschiedliche Schritte ausgeführt werden. Anhand des folgenden Listings wird exemplarisch die Vorgehensweise für das Laden der arithmetischen Operationen Addition und Subtraktion erklärt.

bool cpu\_x86\_32**::**load\_change**(**uint8\_t**\*** instr**,** int len**)** **{**

**this->**instr **=** instr**;**

**this->**len **=** len**;**

string path **=** "data/" **+** get\_instr**();**

**if(!**data\_exist**(**path**))** **{**

**return** **false;**

**}**

string line**;**

ifstream file**(**path**,** ios**::**in**);**

regex rx\_reg**(**"(.+):\t0x([0-9a-f]{8})\t[0-9]+->\t0x([0-9a-f]{8})"**);**

register\_x86 new\_reg **=** regs**.**back**();**

**while(**getline**(**file**,** line**))** **{**

**if(**line **==** ""**)** **break;**

smatch sm**;**

**if(!**regex\_search**(**line**,** sm**,** rx\_reg**))** **{**

cerr **<<** "ERROR: regex\_search returned false!"**;**

**return** **false;**

**}**

string reg\_name**;**

unsigned long vals**[**2**];**

reg\_name **=** sm**[**1**];**

vals**[**0**]** **=** strtoul**(**string**(**sm**[**2**]).**c\_str**(),** **nullptr,** 16**);**

vals**[**1**]** **=** strtoul**(**string**(**sm**[**3**]).**c\_str**(),** **nullptr,** 16**);**

new\_reg**.**regs**[**reg\_name**]** **+=** vals**[**1**]-** vals**[**0**];**

**}**

regs**.**push\_back**(**new\_reg**);**

regex rx\_ram**(**"0x([0-9a-f]+):\t0x([0-9a-f]{2})\t[0-9]+->\t0x([0-9a-f]{2})"**);**

map**<**uint32\_t**,** uint8\_t**>** new\_ram**;**

new\_ram**.**insert**(**ram**.**back**().**begin**(),** ram**.**back**().**end**());**

int counter **=** 0**;**

**while(**getline**(**file**,** line**))** **{**

**if(**line **==** ""**)** **break;**

smatch sm**;**

**if(!**regex\_search**(**line**,** sm**,** rx\_ram**))** **{**

cerr **<<** "ERROR: regex\_search returned false!"**;**

**return** **false;**

**}**

// uint32\_t address;

unsigned long vals**[**2**];**

// address = strtoul(string(sm[1]).c\_str(), nullptr, 16);

vals**[**0**]** **=** strtoul**(**string**(**sm**[**2**]).**c\_str**(),** **nullptr,** 16**);**

vals**[**1**]** **=** strtoul**(**string**(**sm**[**3**]).**c\_str**(),** **nullptr,** 16**);**

new\_ram**[**regs**.**back**().**regs**[**"ESP"**]+**counter**]** **+=** vals**.**back**()** **-** vals**.**front**();**

counter**++;**

**}**

ram**.**push\_back**(**new\_ram**);**

file**.**close**();**

**return** **true;**

**}**

**Listing 11: Funktion zum Laden von Additions- und Subtraktionsbefehlen[[27]](#footnote-27)**

Die Klassenmethode *load\_change* erwartet einen Maschinenbefehl und dessen Länge als Übergabeparameter. Durch die gespeicherten Auswertungen wird das Objekt vom Typ *cpu\_x86\_32* um den Zustand erweitert, der durch den Maschinenbefehl angenommen wird. In den Zeilen 2 und 3 werden die Klassenvariablen auf den Wert der Übergabeparameter gesetzt, um die Information später auf der Konsole ausgeben zu können. Anschließend werden die benötigten Informationen aus der Datenbasis geladen. Mit der Hilfsmethode *get\_instr* wird der Befehl entsprechend des Namens der Datei, in der die Informationen gespeichert sind, formatiert. In den darauffolgenden Zeilen wird durch den in Zeile 5 gebildeten Pfad überprüft, ob die Datei existiert. Ist dies nicht der Fall, so kann der Befehl nicht emuliert werden und die Funktion gibt den Bool’schen Wert *false* zurück. Ist dies nicht der Fall, dann fährt die Methode fort und öffnet in Zeile 11 die Datei in einem Eingangsdateistrom.

Für das Lesen der benötigten Werte aus der Datei werden reguläre Ausdrücke[[28]](#footnote-28) verwendet. Die Ausdrücke werden in den Zeilen 13 und 36 initialisiert. Die regulären Ausdrücke bilden die im vorherigen Teilkapitel „Auswertung speichern“ definierte Syntax ab. Eingeklammerte Teilausdrücke in den Zeichenketten werden beim *regex*-Suchprozess gefangen und in einer Struktur zum Auslesen gespeichert. Zu den Ausdrücken gehört der Name des Registers, bzw. die Adresse des Arbeitsspeichers, der initiale Wert und der Wert nach Ausführung des Befehls.

In Zeile 14 wird das letzte Elemente der Liste *regs* mit den Zuständen der Register in eine neue Variable geladen. Die Variable vom Typ *register\_x86* wird mithilfe der Informationen aus dem Eingangsdateistrom modifiziert und als neue Instanz an die Liste *regs* angehangen. Nach dem *regex*-Suchprozess in Zeile 19 werden in den Zeilen 27-29 die benötigten Daten aus der *regex*-Struktur ausgelesen und gegebenenfalls konvertiert.

Die Zeile 31 ist ausschlaggebend für den Typ des Befehls. Hier könnte durch Informationen, die bei der Untersuchung eines Maschinenbefehls gewonnen werden können, und einer *select-case*-Struktur die entsprechende Operation gewählt werden. In diesem Beispiel wird durch Aufsummierung der Differenz zwischen dem Wert vor und nach Ausführung des Maschinenbefehls auf das jeweilige Register die arithmetische Operation Addition, bzw. bei einer negativen Differenz die Subtraktion, abgebildet.

Die Vorgehensweise bei der Erstellung des neuen Zustands des Arbeitsspeichers ist, abgesehen von einem kleinen Unterschied, identisch. In den Zeilen 37 und 38 wird ebenfalls ein neues Objekt mit den Werten des letzten Zustands für den Arbeitsspeicher angelegt. Anschließend werden die Daten aus dem Eingangsdateistrom herausgelesen. Jedoch wird hier die Adresse im Arbeitsspeicher nicht benötigt, da die verwendeten Adressen meist nicht absolut, sondern relativ zum aktuellen Wert des Stack Pointers sind. So wird in Zeile 56 auch die Differenz des Wertes vor und nach Ausführung des Maschinenbefehls auf den Wert des vorherigen Zustands aufsummiert. Die verwendete Adresse wird jedoch durch den aktuellen Stapelzeiger (ESP-Register) und einen Zähler *counter*, der bei jedem Durchgang der Schleife in Zeile 57 erhöht wird, berechnet. Dies führt bei absoluter Adressierung zu fehlerhafter Interpretation, wird aber in dieser Arbeit vorerst vernachlässigt. Es müssten bei der Analyse weitere Informationen bereitgestellt werden, um den Maschinenbefehl in jedem Fall richtig zu emulieren.

# Funktionen

In diesem Kapitel soll die entwickelte Software beispielhaft an einer Testdatei vorgeführt werden. Dabei werden explizit die Informationen über die in der kompilierten Testdatei verwendeten Maschinenbefehle von der Maschine gesammelt. Anschließend wird versucht das Verhalten des Testprogramms durch eine Emulation nachzuvollziehen. Einfachheitshalber werden in der Testdatei hauptsächlich arithmetische Operationen durchgeführt.

#include <stdio.h>

int main**()** **{**

int i **=** 0x77**;**

i **=** 0x33**;**

i **+=** 0x55**;**

int j **=** 0x99**;**

j **-=** 0x44**;**

i **=** i **-** 0x11 **-** 0x11 **+** j **+** 0x11**;**

j **=** i **+** 0x22**;**

printf**(**"%i\n"**,** i**);**

**return** i**;**

**}**

**Listing 11: Datei für ein Beispiel der Emulation**

Um das komplette Programm inklusive der Testdatei zu kompilieren wird ein Makefile[[29]](#footnote-29) bereitgestellt.

TARGET **=** collect\_data emulate\_file

ADDITIONAL **=** template test/test

LIBS **=**

CC **=** g++-4.9 # because of regex

CFLAGS **=** -g -Wall -m32 -std=c++11

.PHONY**:** default all clean

default**:** $(TARGET) $(ADDITIONAL)

all**:** default

OBJECTS **=** $(patsubst %.cpp, %.o, $(wildcard \*.cpp))

HEADERS **=** $(wildcard \*.h)

%.o**:** %.cpp $(HEADERS)

$(CC) $(CFLAGS) -c $< -o $@

collect\_data**:** $(OBJECTS)

$(CC) $(filter-out emulate\_file.o,$(OBJECTS)) -Wall -m32 $(LIBS) -o $@

emulate\_file**:** $(OBJECTS)

$(CC) $(filter-out collect\_data.o,$(OBJECTS)) -Wall -m32 $(LIBS) -o $@

%**:** %.c

$(CC) $(CFLAGS) $< -o $@

clean**:**

rm -f $(TARGET) $(ADDITIONAL) execute \*.o

**Listing 12: Makefile zur Kompilierung der entwickelten Software**

In Zeile 1 enthält die Variable *TARGET* die zu kompilierenden Teilprogramme der Software. Das Programm „collect\_data“ ist für das Sammeln der Informationen über Maschinenbefehle zuständig. Mit dem Programm „emulate\_file“ sollen ausführbare Dateien emuliert werden. In Zeile 2 sind zwei weitere Programme beschrieben, die kompiliert werden müssen. Zum einen die Vorlage „template“ zur Erstellung der zu untersuchenden ausführbaren Datei mit dem jeweiligen Bytecode. Zum anderen die im zweiten Schritt zu emulierende Testdatei „test/test“. In Zeile 4 ist der zu verwendende Kompilierer spezifiziert. Es wird der *gcc* der Version 4.9 verwendet, weil ältere Kompilierer die regulären Ausdrücke nicht richtig umsetzten. Dabei werden die Flags *CFLAGS* in Zeile 5 als Parameter übergeben. Wichtig sind hierbei die Flags *-m32*, für eine Kompilierung für 32-Bit Maschinen, und *-std=c++11*, aufgrund der im Quellcode verwendeten Syntax und Funktionen. Im restlichen Teil des Makefiles wurden die entsprechenden Regeln für die Erstellung der benötigten Dateien angegeben.

## Daten sammeln

Im Allgemeinen kann durch einfaches Ausführen der Datei „collect\_data“ der Prozess zum Sammeln der Informationen über Maschinenbefehle gestartet werden. Im Moment werden nur acht Byte lange Maschinenbefehle untersucht, da die wichtigsten Maschinenbefehle diese Länge nicht überschreiten. Wird der Prozess ohne Argumente gestartet, so werden die Daten startend vom Bytecode 0x00000000 aus gesammelt. Dabei wird immer das erste Byte im nächsten Durchgang um eins erhöht, sodass beim zweiten Durchlauf der Bytecode 0x01000000 untersucht wird. Die andere Option besteht daraus, als Argument eine hexadezimalkodierte Zahl ohne das Präfix „0x0“ zu übergeben. In dem Fall startet der Prozess mit diesem Bytecode und fährt ebenfalls wie angegeben fort.

Um die Beispieldatei zu emulieren, müssen die verwendeten Maschinenbefehle gelernt werden. Um den Lernvorgang zu beschleunigen werden die Maschinenbefehle aus der Beispieldatei explizit gelernt, indem jeder Bytecode des Testprogramms dem Programm „collect\_data“ als Argument übergeben wird.

0804841b **<**main**>:**

804841b**:** 8d 4c 24 04 **lea** 0x4**(%esp),%ecx**

804841f**:** 83 e4 f0 **and** $0xfffffff0**,%esp**

8048422**:** ff 71 fc pushl **-**0x4**(%ecx)**

8048425**:** 55 **push** %ebp

8048426**:** 89 e5 **mov** %esp**,%ebp**

8048428**:** 51 **push** %ecx

8048429**:** 83 ec 14 **sub** $0x14**,%esp**

804842c**:** c7 45 f0 77 00 00 00 movl $0x77**,-**0x10**(%ebp)**

8048433**:** c7 45 f0 33 00 00 00 movl $0x33**,-**0x10**(%ebp)**

804843a**:** 83 45 f0 55 addl $0x55**,-**0x10**(%ebp)**

804843e**:** c7 45 f4 99 00 00 00 movl $0x99**,-**0xc**(%ebp)**

8048445**:** 83 6d f4 44 subl $0x44**,-**0xc**(%ebp)**

8048449**:** 8b 45 f0 **mov** **-**0x10**(%ebp),%eax**

804844c**:** 8d 50 de **lea** **-**0x22**(%eax),%edx**

804844f**:** 8b 45 f4 **mov** **-**0xc**(%ebp),%eax**

8048452**:** 01 d0 **add** %edx**,%eax**

8048454**:** 83 c0 11 **add** $0x11**,%eax**

8048457**:** 89 45 f0 **mov** %eax**,-**0x10**(%ebp)**

804845a**:** 8b 45 f0 **mov** **-**0x10**(%ebp),%eax**

804845d**:** 83 c0 22 **add** $0x22**,%eax**

8048460**:** 89 45 f4 **mov** %eax**,-**0xc**(%ebp)**

8048463**:** 83 ec 08 **sub** $0x8**,%esp**

8048466**:** ff 75 f0 pushl **-**0x10**(%ebp)**

8048469**:** 68 20 85 04 08 **push** $0x8048520

804846e**:** e8 7d fe ff ff **call** 80482f0 **<**printf@plt**>**

8048473**:** 83 c4 10 **add** $0x10**,%esp**

8048476**:** 8b 45 f0 **mov** **-**0x10**(%ebp),%eax**

8048479**:** 8b 4d fc **mov** **-**0x4**(%ebp),%ecx**

804847c**:** c9 **leave**

804847d**:** 8d 61 fc **lea** **-**0x4**(%ecx),%esp**

8048480**:** c3 **ret**

**Listing 13: Disassembly der Main-Funktion der Beispieldatei**

Um nur den wesentlichen Teil der Beispieldatei zu emulieren, werden die Maschinenbefehle zur Verwaltung des Speichers und für den Aufruf der *printf*-Funktion ausgelassen, sodass nur die Zeilen 8-22 gelernt werden müssen. Das Lernen der Maschinenbefehle wird beispielhaft an Zeile 8 dargestellt und wird für die restlichen Zeilen auf dieselbe Weise wiederholt. Der Prozess zum Sammeln der Daten für den Maschinenbefehl in Zeile 8 wird durch folgenden Konsolenbefehl gestartet: *./collect\_data 83ec14*. Falls noch keine Datei für den Maschinenbefehl existiert, wird die Datei erzeugt und mit Informationen angereichert. Dabei wird in der Konsole de Text in Listing 14 ausgegeben. Andernfalls wird die Meldung „Data already exist!“ ausgegeben und mit dem nächsten Bytecode fortgefahren.

$ **./collect\_data 83ec14**

0x83 0xec 0x14 0x00 0x00 0x00 0x00 0x00

Terminated with Signal Code 0xb in address 0x80483f1!

Following registers changed:

EFLAGS: 0x00000246 1-> 0x00000282

EIP: 0x080483ee 1-> 0x080483f1

ESP: 0xffa3d4d8 1-> 0xffa3d4c4

Following ram address values changed:

SUCCESS!

**Listing 14: Ausgabe des Aufrufs *./collect\_data 83ec14***

In der ersten Zeile der Ausgabe wird der zu untersuchende Bytecode ausgegeben. Zeile 3 gibt an, dass das Programm aufgrund eines „Segmentation Faults“ (Signalcode 0xb) an der Adresse 0x80483f1 abgestürzt ist. Dabei handelt es sich um die Position, an der sich der irrelevante Bytecode 0x00, der auf unseren eigentlichen Bytecode folgt, befindet. Anschließend wird zusätzlich ausgegeben, welche Änderungen stattgefunden haben. Die Zeilen 5-6 werden unverändert in die für die Emulation erzeugte Datei „83ec14“ im Ordner „data“ eingefügt. Falls Änderungen im Arbeitsspeicher stattgefunden hätten, so hätte die Ausgabe diesbezüglich auf Zeile 8 gefolgt.

## Ausführbare Datei emulieren

Für die Emulation von Maschinencode aus einer ausführbaren Datei erfolgt ein Aufruf des Programms „emulate\_file“. Als Argument erwartet das Programm den Pfad der zu emulierenden Datei. Der Pfad zu der Beispieldatei aus Listing 11 kann in dem Fall relativ durch „test/test“ angegeben werden. Für die Emulation wird ein initialer Zustand der Maschine zum Anfang der Main-Funktion benötigt. Hierfür wird ein komplettes Objekt der Klasse *cpu\_x86\_32* mit dem Zustand der Maschine zum Anfang der Main-Funktion erstellt und komplett in der Datenbank im Ordner „data“ unter dem Dateinamen „blank“ abgespeichert, falls die Datei nicht bereits existiert. Anschließend kann der Zustand durch die Klassenmethode *load\_blank*, auch zu späteren Zeitpunkten der Ausführung des Emulationsprogramms, geladen werden. Die Emulation erfolgt durch byteweises Lesen der Programmdatei mithilfe eines Eingangsdateistroms. Um eine gezielte Funktion, bzw. einen gezielten Bereich, dieser Datei zu emulieren, müssen Start- und Endadresse bekannt sein. Für dieses Beispiel wurden die Werte im Emulationsprogramm hartkodiert. Die Werte können aus dem Disassembly der Beispieldatei im Listing 13 herausgelesen werden. Wie bereits beschrieben soll für diesen Test versucht werden ausschließlich die Zeilen 8-22 zu emulieren. Dementsprechend lautet die Adresse der Startanweisung in Zeile 8 0x8048429. Die Endadresse wird aus Zeile 23 herausgelesen und lautet 0x8048463, damit die Zeile 22 noch einschließlich emuliert und erst in Zeile 23 abgebrochen wird. Die Adressen geben jedoch die Position des Maschinenbefehls im Arbeitsspeicher während der Ausführung des Programms an. Der Offset des Maschinenbefehls in der ausführbaren Datei wird durch die letzten 12 Bits der genannten Speicheradressen dargestellt. In dem Fall lautet die Startadresse 0x429 und die Endadresse 0x463.

Nachdem die Vorbereitungen getroffen wurden, beginnt die Emulation durch das Lesen der Testdatei. Es wird immer ein Zeichen in einen Puffer geschrieben. Anschließend wird überprüft, ob zu der Zeichenkette im Puffer ein verfügbarer Datensatz vorliegt. Ist dies der Fall, so wird der Datensatz geladen und die Zustandsänderungen des maschinenabbildenden Objekts der Klasse *cpu\_x86\_32* werden auf der Konsole ausgegeben. Andernfalls wird das nächste Zeichen aus der Testdatei in den Puffer geladen und erneut das Vorhandensein des Datensatzes überprüft. Nach erfolgreicher Ausgabe wird der Benutzer durch die Ausgabe „Finished!“ benachrichtigt, dass die Emulation fortgesetzt werden kann. Durch die Eingabe von „Enter“ in der Konsole wird die Emulation auf die gleiche Weise fortgesetzt. Die Emulation wird beendet, falls die Endadresse erreicht wurde oder der Puffer mehr als acht Zeichen erhält, weil im Moment die maximale Länge eines gelernten Maschinenbefehls acht Zeichen beträgt. So ergibt der Aufruf *./emulate\_file test/test* die in Listing 15 zu sehenden Ausgaben. Aufgrund des großen Umfangs wurde die Ausgabe gekürzt und es werden exemplarisch vier Emulationsschritte analysiert. Die gekürzten Stellen wurden mit drei aufeinanderfolgenden Punkten markiert.

$ **./emulate\_file test/test**

Data data/83 does not exist!

Data data/83ec does not exist!

Instruction: 83ec14

Following registers changed:

EFLAGS: 0x00000246 1-> 0x00000282

EIP: 0x080483f0 1-> 0x080483f3

ESP: 0xff873c38 1-> 0xff873c24

Following ram address values changed:

Finished!

[…]

Instruction: c745f077000000

Following registers changed:

EFLAGS: 0x00000246 1-> 0x00000282

EIP: 0x080483f0 1-> 0x080483f3 2-> 0x080483fa

ESP: 0xff873c38 1-> 0xff873c24

Following ram address values changed:

0xff873c24: 0x00 2-> 0x5c

0xff873c25: 0x01 2-> 0x7d

0xff873c26: 0x01 2-> 0xfd

0xff873c27: 0x02 2-> 0xfa

Finished!

[…]

Instruction: 8345f055

Following registers changed:

EFLAGS: 0x00000246 1-> 0x00000282 4-> 0x0000024e

EIP: 0x080483f0 1-> 0x080483f3 2-> 0x080483fa 3-> 0x08048401 4-> 0x08048405

ESP: 0xff873c38 1-> 0xff873c24

Following ram address values changed:

0xff873c24: 0x00 2-> 0x5c 3-> 0x74 4-> 0xc9

0xff873c25: 0x01 2-> 0x7d 3-> 0xf9

0xff873c26: 0x01 2-> 0xfd 3-> 0xf9

0xff873c27: 0x02 2-> 0xfa 3-> 0xf2

Finished!

[…]

Instruction: 83c022

Following registers changed:

EAX: 0x00000001 7-> 0x0804841b 9-> 0xff7e741a 10-> 0xff05b3fe 11-> 0xff05b40f 13-> 0x070a3829 14-> 0x070a384b

EDX: 0xff873c64 8-> 0xffbe0bff

EFLAGS: 0x00000246 1-> 0x00000282 4-> 0x0000024e 6-> 0x0000029a 10-> 0x000002d6 11-> 0x00000296 14-> 0x00000252

EIP: 0x080483f0 1-> 0x080483f3 2-> 0x080483fa 3-> 0x08048401 4-> 0x08048405 5-> 0x0804840c 6-> 0x08048410 7-> 0x08048413 8-> 0x08048416 9-> 0x08048419 10-> 0x0804841b 11-> 0x0804841e 12-> 0x08048421 13-> 0x08048424 14-> 0x08048427

ESP: 0xff873c38 1-> 0xff873c24

Following ram address values changed:

0xff873c24: 0x00 2-> 0x5c 3-> 0x74 4-> 0xc9 5-> 0x62 6-> 0x1e 12-> 0x04

0xff873c25: 0x01 2-> 0x7d 3-> 0xf9 5-> 0x39 6-> 0x38 12-> 0xb4

0xff873c26: 0x01 2-> 0xfd 3-> 0xf9 5-> 0x89 12-> 0x85

0xff873c27: 0x02 2-> 0xfa 3-> 0xf2 5-> 0xfb 12-> 0xf3

Finished

[…]

**Listing 15: Ausgabe des Emulators[[30]](#footnote-30)**

In den Zeilen 2 und 3 ist zu erkennen, wie der Puffer Byte für Byte gefüllt wird und nach der entsprechenden Datei mit den benötigten Informationen gesucht wird. In Zeile 4 wurde der erste Maschinenbefehl mit dem Bytecode 0x83ec14 erkannt. Dieser entspricht der Assembleranweisung *sub $0x14,%esp*, welche den Stapelzeiger (ESP-Register) um den festen Wert 0x14 erniedrigt. Die Operation auf dem ESP-Register wird in Zeile 8 erfolgreich angezeigt. Der Befehlszeiger (EIP-Register) wird ebenfalls korrekt um die Länge des Bytecodes erhöht. Hier tritt jedoch schon die erste Fehlinterpretation beim EFLAGS-Register auf. Hier wird auch eine arithmetische Operation ausgeführt, jedoch stellt das EFLAGS-Register Informationen durch gesetzte Bits über die letzte ausgeführte Operation bereit. In diesem Fall wurde das Parity Flag fälschlicherweise auf den Wert null gesetzt. Das Parity Flag gibt an, ob das niederwertigste Byte des Ergebnisses eine gerade Anzahl gesetzter Bits enthält[[31]](#footnote-31). Das Ergebnis ist in dem Fall der neu berechnete Wert im ESP-Register 0xff873c24 und das niederwertigste Byte 0x24, welches dem Binärcode „0010 0100“ entspricht und mit zwei gesetzten Bits eine gerade Bitanzahl besitzt. Dieses Problem tritt auf, weil bei der Untersuchung des Maschinenbefehls der Stapelzeiger einen anderen initialen Wert hatte, dessen niederwertigstes Byte nach der Subtraktion eine ungerade Anzahl gesetzter Bits aufwies.

Nachdem durch die vorausgegangene Operation der Stapel vorbereitet wurde, wird nun in Zeile 13 mit dem Bytecode 0xc745f077000000, welcher der Assembleranweisung *movl $0x77,-0x10(%ebp)* entspricht, der Wert 0x77 auf diesem abgelegt. Die Operation entspricht in dem Fall der Initialisierung der Variablen *i* in der Testdatei in Zeile 4. Anhand der Assembleranweisung erkennt man, dass die Adresse, an der der Wert abgelegt wird, relativ zum Wert im EBP-Register berechnet wird. Da diese Information beim Datensammeln nicht erkannt wurde, wird der Wert an der Adresse im aktuellen Stapelzeiger abgelegt. Auf dieser Ebene wird der Befehl trotzdem korrekt emuliert, weil der Wert des EBP-Registers dem Wert des ESP-Register vor der Subtraktion mit dem Wert 0x14 entspricht. Der Integerwert mit einer Größe von vier Byte wird dementsprechend an der Adresse im EBP-Register mit dem Offset -0x10 abgelegt, an welcher vier Bytes Platz bis zum Stapelzeiger für den Wert zur Verfügung stehen. Das Problem liegt jedoch an der Art der Operation. Ebenso wie in der Emulation hatten auch bei der Untersuchung des Bytecodes die jeweiligen Adressen einen von null unterschiedlichen Wert. Aufgrund der Annahme, dass nur arithmetische Operationen ausgeführt werden, wurden falsche Summanden aus der Datenbank berechnet und auf den aktuellen Wert an der Adresse addiert.

In der Zeile 26 wird durch den Bytecode eine Addition auf derselben Adresse wie in Zeile 13 ausgeführt. Der Bytecode entspricht dem Assemblerbefehl *addl $0x55,-0x10(%ebp)* und addiert den festen Wert 0x55, wie in der Testdatei in Zeile 6 zu sehen, auf eine Adresse im Arbeitsspeicher. Nach dieser arithmetischen Operation wird erneut das EFLAGS-Register modifiziert. Da jedoch der neue Wert nicht einfach gesetzt, sondern Differenz zum vorherigen Wert aufaddiert wird, erhält das EFLAGS-Register einen ungültigen Wert, da das reservierte Bit an dritter Stelle verändert wurde. In Zeile 33 ist jedoch zu erkennen, dass die Addition erfolgreich emuliert wurde. Auf den (durch Fehler in den vorausgegangenen Emulationsschritten) falschen Wert wurde der Wert 0x55 erfolgreich addiert.

Der Bytecode in Zeile 40 entspricht der der Assembleranweisung *add $0x22,%eax*, welcher den Wert 0x22 auf den Wert im EAX-Register addiert. Der Maschinenbefehl ist ein Teil der in der Testdatei in Zeile 10 ausgeführten Operation. Die Zeile 10 der Testdatei umfasst drei Maschinenbefehl. Zuerst wird die Variable *i* in das EAX-Register geladen, anschließend wird der in diesem Abschnitt beschriebene Befehl ausgeführt und zum Schluss wird der errechnete Wert aus dem EAX-Register auf den Stapel an den für die Variable *j* reservierten Platz abgelegt. Der Addierbefehl findet im 14ten Emulationsschritt statt, wie in Zeile 43 zu erkennen ist. Die Zahl 0x22 wird erfolgreich auf den Wert im vorigen Emulationsschritt 0x070a3829 aufaddiert. Auch das EFLAGS-Register wird in diesem Emulationsschritt durch eine falsche Operation modifiziert und das EIP-Register wird wie gewohnt um die Länge des Bytecodes erhöht.

# Fazit & Ausblick

Abschließend lässt sich sagen, dass die Erstellung eines selbstgenerierenden Emulators möglich ist, jedoch wie die Programmierung von einem Emulator mithilfe der Befehlssatzarchitekturen viel Zeit beansprucht, sodass abgewogen werden muss, ob dadurch tatsächlich ein Gewinn entsteht. Das Programm dieser Bachelorarbeit steht, wie die meisten öffentlichen Dekompilierer-Projekte, ganz am Anfang der Entwicklung und ist nur ansatzweise zu gebrauchen. Der erste und einfachste Schritt in Richtung einer fehlerlosen Emulation wäre es, mithilfe einer *switch-case*-Anweisung beim Laden einer Anweisung zwischen gewissen Registern zu unterscheiden und bei Operationen auf dem Arbeitsspeicher zu überprüfen, ob die Adressierung durch ein Register erfolgt. Durch weitere kleine Verbesserungen sollte es schließlich möglich sein, ein funktionsfähiges Reverse Engineering Werkzeug zu erstellen.

Das Wichtigste für eine richtige Emulation ist die Bereitstellung der Datenbank. Durch Techniken der künstlichen Intelligenz kann es gelingen weitere wichtige Informationen über Maschinenbefehle zu erlangen. Es wird beispielsweise vorteilhaft sein, Maschinenbefehle mit Parametern zu erkennen und nur einen Datensatz mit Platzhalter abzuspeichern, anstatt für jedes Argument Informationen zu speichern. Auch die Struktur der Datenbank muss für eine effiziente Emulation entsprechend angepasst werden. Im Programm wird es sinnvoll sein, die Klassen zur Abbildung der Maschine anzupassen. Beispielsweise wird derzeit für jeden Zustand der Maschine eine neue Map an eine Liste angehängt, in der sowohl der Name des Registers, als auch der jeweilige Wert gespeichert werden. Durch Umstellen der Struktur von Maps in Listen zu einer Map, bei der der Schlüssel dem Namen des Registers entspricht, und der zugehörige Wert eine Liste mit den Werten zu jedem Zustand der Maschine ist, kann das redundante Speichern des Registernamens vermieden werden.

Des Weiteren sollte die weitaus stärkere GDB Maschinenschnittstelle zur Untersuchung eines Prozesses und somit des Bytecodes implementiert werden. Um eine Maschine vollständig untersuchen zu können, muss es möglich sein, beliebigen Bytecode auf Kernel Level auszuführen und die Debug und Control Register auszulesen. Um für die Emulation eines kleinen Programms nicht zuerst komplett alle Maschinenbefehle auf der Maschine auszuführen und zu untersuchen, was einige Zeit beansprucht, ist es vorteilhaft, die Maschinenbefehle des Programms parallel zur Emulation zu lernen.

In dieser Arbeit wurde das Programm nur für eine x86 Maschine mit 32-Bit entwickelt. Um andere Maschinen zu untersuchen, müssen meist kleine Änderungen vorgenommen werden. In diesem Sinn bietet es sich an, abstrakte Klassen zu verwenden, welche maschinenunabhängige Funktionen implementieren und virtuelle Funktionen zum Überschreiben anbieten. Beispielsweise müssten Methoden zum Auslesen der Register überschrieben werden. Methoden zur Ausgabe der Änderungen der Register- und Arbeitsspeicherwerte auf der Konsole dürften hingegen ohne Probleme auf verschiedenen Maschinen laufen. Jedoch sollten diese Methoden effizient gestaltet werden. Im Moment wird im kompletten Arbeitsspeicher nach Änderungen gesucht, bevor diese auf der Konsole ausgegeben werden. Zur Beschleunigung der Emulation können beispielsweise Adressen, bei denen sich der Wert geändert hat, markiert werden, sodass die relevanten Stellen schneller gefunden werden können.

# Anhang

## collect\_data.cpp

## cpu\_x86\_32.cpp

## cpu\_x86\_32.h

## emulate\_file.cpp

## register\_x86.cpp

## register\_x86.h

## GDB/MI Sitzungsprotokoll

## Ausgabe des Emulators

1. <http://stackoverflow.com/questions/9960721/how-to-get-c-code-to-execute-hex-bytecode>, 27.06.2016 [↑](#footnote-ref-1)
2. Deklariert in „sys/mman.h“ [↑](#footnote-ref-2)
3. Deklariert in „string.h“ [↑](#footnote-ref-3)
4. Kompletter Quellcode im Anhang unter „cpu\_x86\_32.cpp“ [↑](#footnote-ref-4)
5. Deklariert in „fstream“ [↑](#footnote-ref-5)
6. Deklariert in „sys/stat.h“ [↑](#footnote-ref-6)
7. <https://en.wikipedia.org/wiki/Ptrace>, 05.06.2016 [↑](#footnote-ref-7)
8. Laden eines *unaligned double-quadword floating-point* Wertes aus dem YMM0-Register in die Variable *out* [↑](#footnote-ref-8)
9. <http://man7.org/linux/man-pages/man2/ptrace.2.html>, 06.06.2016 [↑](#footnote-ref-9)
10. Deklariert in „sys/ptrace.h“ [↑](#footnote-ref-10)
11. Deklariert in „unistd.h“ [↑](#footnote-ref-11)
12. Deklariert in „unistd.h“ [↑](#footnote-ref-12)
13. Kompletter Quellcode im Anhang unter „cpu\_x86\_32.cpp“ [↑](#footnote-ref-13)
14. Deklariert in „sys/wait.h“ [↑](#footnote-ref-14)
15. Kompletter Quellcode im Anhang unter „cpu\_x86\_32.cpp“ [↑](#footnote-ref-15)
16. Kompletter Quellcode im Anhang unter „cpu\_x86\_32.cpp“ [↑](#footnote-ref-16)
17. Definiert in „signal.h“ [↑](#footnote-ref-17)
18. Definiert in „sys/user.h“ [↑](#footnote-ref-18)
19. <https://de.wikipedia.org/wiki/Advanced_Vector_Extensions>, 07.06.2016 [↑](#footnote-ref-19)
20. Kompletter Quellcode im Anhang unter „cpu\_x86\_32.cpp“ [↑](#footnote-ref-20)
21. Deklariert in „string“ [↑](#footnote-ref-21)
22. <https://sourceware.org/gdb/onlinedocs/gdb/GDB_002fMI.html>, 07.06.2016 [↑](#footnote-ref-22)
23. <http://man7.org/linux/man-pages/man3/openpty.3.html>, 07.06.2016 [↑](#footnote-ref-23)
24. Deklariert in „pty.h“ [↑](#footnote-ref-24)
25. Komplettes Protokoll im Anhang unter „GDB/MI Sitzungsprotokoll“ [↑](#footnote-ref-25)
26. <https://de.wikipedia.org/wiki/Maschinensprache>, 07.06.2016 [↑](#footnote-ref-26)
27. Kompletter Quellcode im Anhang unter „cpu\_x86\_32.cpp“ [↑](#footnote-ref-27)
28. Deklariert in „regex“ [↑](#footnote-ref-28)
29. Vorlage: <http://stackoverflow.com/questions/1484817/how-do-i-make-a-simple-makefile-gcc-unix>, 27.06.2016 [↑](#footnote-ref-29)
30. Komplette Ausgabe im Anhang unter „Ausgabe des Emulators“ [↑](#footnote-ref-30)
31. <https://www.lowlevel.eu/wiki/EFLAGS>, 27.06.2016 [↑](#footnote-ref-31)