Maturaarbeit

Compiler Construction

Fabio Stalder

Betreut durch
Thomas Jampen

1. September 2024

Gymnasium Kirchenfeld
Abteilung MN

Inhaltsverzeichnis

1 Introduction

In der Informatik beschreibt Compiler ein Programm, das Code aus einer Programmiersprache in eine andere übersetzt. In dieser Hinsicht gleichen Compiler Übersetzern für Menschensprache. Jedoch unterscheidet sich ein Compiler grundsätzlich von Übersetzern in der Erwartungshaltung, die an sie gestellt wird. Menschensprache ist sehr komplex und [...]

I have Idea

2 Traditioneller Compiler

Ein Compiler ist traditionell nach folgendem Schema aufgebaut.

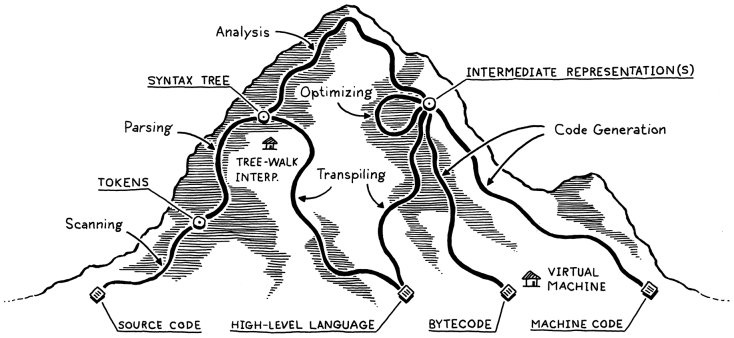

Abbildung 2.1: Schritte, die ein Compiler durchläuft [1]

In dieser Arbeit werde ich mich nur auf die im unteren Schema dargestellten Schritte fokussieren.

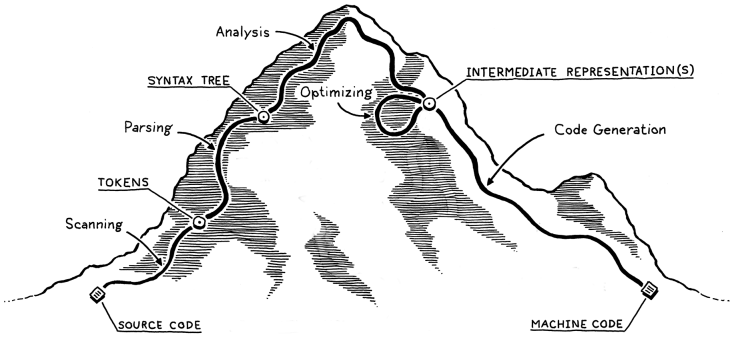

Abbildung 2.2: Schritte, die in dieser Arbeit behandelt werden (Basierend auf Figure 2.1)

2.1 Lexical Analysis

Meist werden Programme so geschrieben, dass wir Menschen es lesen und verstehen können. Dafür verwendet man Buchstaben und Zahlen, Zeichen, wie +, *, oder Klammern, und Whitespaces, wie Leerzeichen oder Absätze. Diese Zeichen sind jedoch für den Computer unverständlich. Der erste Schritt beim compilieren ist daher die Lexical Analysis. Dies wird von einem Teil des Compilers, dem Lexer, durchgeführt. Die Aufgabe dieses Lexers ist es den Input File zu scannen und die gescannten Zeichen in sogenannte Tokens zu verwandeln. Diese Tokens sind Datenstrukturen, die der Compiler kennt und mit denen er weiterarbeiten kann.

Als Beispiel:

Listing 2.1: C code vor Lexical Analysis

```
int foo()
{
    if (bar == 0)
    {
        return 0;
    }

    return 1;
}
```

Würde hierbei zu einem Array von Token Objekten umgewandelt werden:

Listing 2.2: Tokens nach Lexical Analysis

```
Keyword         (keyword="int")
Identifier      (id="foo")
LParenthesis
RParenthesis
Keyword         (keyword="if")
LParenthesis
Identifier      (id="bar")
Operator        (operator=ComparisonEqual)
LiteralInt      (value=0)
[...]
```

Der Lexer legt hierbei fest welche Zeichen die Input-Programmiersprache enthalten darf und welche Bedeutung ihnen Zugesprochen wird. So ist zum Beispiel im Lexer festgelegt, dass ein + Zeichen als Addition interpretiert wird. Genauso wie im Listing 2.2 'if' als KeywordToken gesehen wird, lässt sich im Lexer auch bestimmen, dass ein Wort wie 'print' als Keyword angesehen werden soll.

2.2 Syntax Analysis

Nun versteht der Compiler was mit den Zeichen im Input File gemeint ist, jedoch fehlt noch etwas bis tatsächlich in einer andere Programmiersprache übersetzt werden kann. Und das ist Verständnis für Syntax. Die meisten High-Level Programmiersprachen weisen Syntaxregeln auf. Diese beinhalten, wie Funktionen und Variablen definiert werden oder mit welchen Punktvorstrich-Regeln Expressions evaluiert werden. Die bei der Lexical Analysis gefundenen Tokens werden nun ineinander verschachtelt und in einen sogenannten Abstract Syntax Tree (AST) überführt.

Ein AST enthält somit nicht nur Informationen über die Tokens, sondern über die gesamte Struktur die sich aus den Tokens ergibt. Variabel- und Funktionsdefinitionen oder komplexe Statements wie 'if' oder 'for' sind hierbei im AST enthalten. [...]

2.3 Semantic Analysis

Wie auch bei den meisten Menschensprachen gibt es auch für Programmiersprachen eine Semantik und diese muss natürlich vom Compiler verstanden werden. [...]

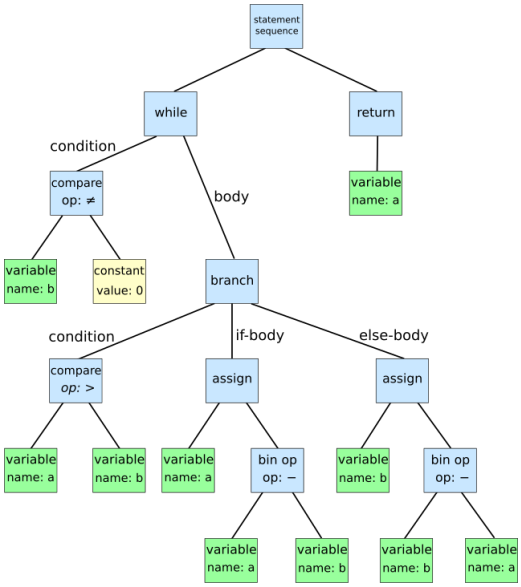

Abbildung 2.3: Abstract Syntax Tree zum Euklidischen Algorithmus [2]

2.4 Code Generation

Code Generation ist der finale und oft auch komplexeste Schritt, der ein Compiler ausführen muss. Nun da unser Input-Code nicht mehr nur als Textfile, sondern als Intermediate Representation vorliegt, kann endlich Output-Code generiert werden. Jedoch lässt sich über diesen Schritt fast am wenigsten sagen, da er je nach Output-Sprache sehr unterschiedlich aussehen kann.

2.5 Optimization

Code Generation ist zwar der letzte Schritt beim Compilieren, trotzdem wurde eine wichtige Aufgabe des Compilers noch nicht betrachtet. Optimization ist ein Sprache die zwischen jedem der genannten Schritte geschiet. Dabei geht es darum den Output-Code so effizient wie möglich zu machen. Effizient kann hierbei jedoch viel Verschiedenes bedeuten. Der Output-Code muss so schnell wie möglich ausgeführt werden können, Memory sparsam verwenden und am besten auch noch ein kleiner File sein. Optimization reicht vom Entfernen der Kommentare beim Scannen oder umstellen von mathematischen Operationen bis zu entfernen von ungebrauchten Variablen und Deadstores. Es muss von CPU Registern profitiert, mit Heap-Memory umgegangen und von inline Funktionen Gebrauch gemacht werden. Compiler Optimization ist somit ein sehr vielseitiges Problem, dass hierbei nicht weiter thematisiert werden sollte.

3 Meine Idee

Ein Compiler ein äusserts komplexes Programm, mit vielen verschiedenen Schritten. Jedoch ist die zugrundeliegende Aufgabe gar nicht so kompliziert. Man braucht ja nur, ein Dokument mit Text der bestimmten Regeln folgt, in Text mit anderen Regeln verwandeln. Natürlich ist dies etwas salopp ausgedrückt, trotzdem fragte ich mich, ob es nicht möglich sei einen viel einfacheren Compiler zu schreiben. Während meinen **Nachforschungen** zum Thema Compiler, stiess ich auf den Begriff *Macro Expansion*. Dabei handelt es sich um eine Methode zur Code Generation, bei der eine Struktur des AST (z.B. ein If-Statement) mit einem *Macro* ersetzt wird. Bei diesem Macro kann es sich um so ziemlich alles, meist jedoch ein Stuck Assembly oder Object Code, handeln. Somit wird nach und nach der gesamte AST ersetzt, bis nur noch Macros übrig sind. Inspiriert von dieser Methode der Code Generation, überlegte ich mir ein Compiler, der möglichst stark dieses Konzept der Macros verwendet. Ein Compiler, der sowohl Lexical als auch Syntax Analysis überspringen kann. Ein Compiler, der es ermöglicht in einem Programm weitere Macros, und somit eine Art eigene Programmiersprache, selbst zu definieren.

3.1 Vergleich der Compiler

Um die Leistung meiner Idee zu testen, werden folgende der Compiler verglichen.

Der *QHS Compiler* ist ein von mir nach meiner Idee entwickelter Compiler. Der QHS Compiler ist in C++ geschrieben und generiert x86 Assembly nach NASM Syntax aus meiner Sprache QHS.

Der GCC Compiler ist der gebräuchlichste Compiler für die Programmiersprache C. Veröffentlicht im Jahre 1987 wird GCC bis heute weiterentwickelt [...]

Der *THS Compiler* repräsentiert in diesem Vergleich einen traditionellen von mir geschrieben Compiler. Im Gegensatz zu GCC ist der THS Compiler deutlich simpler und kleiner. Er dient daher als realistische Konkurrenz zum QHS Compiler und wird verwendet, um zu testen, wie viel meine Idee taugt. Er folgt dem theoretischen Aufbau eines Compilers und besteht aus Lexer, Parser und Code Generator. Als Parser wird ein Predictive Descent Parser verwendet. Optimization wird spezifisch keine durchgeführt. Der Code Generator arbeitet auf dem Abstract Syntax Tree mithilfe eines Visitor Patterns. Die Semantic Analysis wird während der Code Generation durchgeführt. Geschrieben ist der Compiler in C++ und liefert x86 Assembly nach NASM Syntax.

3.1.1 Kriterien des Vergleichs

Die Compiler werden nach folgenden Kriterien bewertet und verglichen.

Geschwindigkeit des Output-Codes	Wie schnell wird der Output-Code ausgeführt?
Geschwindigkeit der Compilation	Wie lange dauert die Compilation von Code?
Benutzerfreundlichkeit	Wie einfach ist die Verwendung des Compilers?
Möglichkeit für Erweiterung	Wie einfach ist die Input-Sprache zu erweitern?

3.1.2 Anforderungen an die Compiler

Um einen **fairen** Vergleich zu ermöglichen, müssen die Compiler folgende Anforderungen erfüllen.

Output als Assembly Code	Die Output-Sprache muss Assembly Code sein
C-like Syntax	Die Input-Sprache muss einen C-ähnlichen Syntax aufweisen
Variablen und Funktionen	Lokale und globale Variablen sowie Funktionen müssen unterstützt werden
Branching und Loops	If-Statements sowie While-Loops müssen umsetzbar sein

4 QHS Compiler

Der Compilation von QHS steht ein einfacher Zyklus zugrunde, dessen Vorbild der Von-Neumann Zyklus ist.

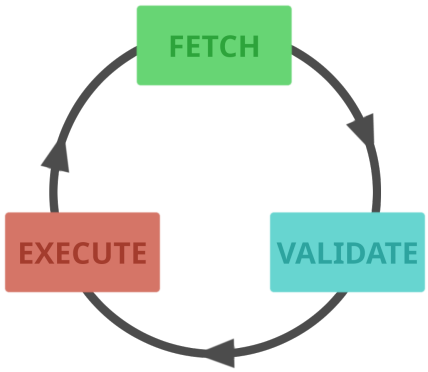

Abbildung 4.1: Zyklus der QHS Compilation

TALK ABOUT ORDERS AND THE MEANING OF IDENTIFIERS, INSTRUCTIONS AND LITERAL CODE

4.1 Fetch

Der QHS-Zyklus beginnt mit dem ersten Fetch. Dabei wird die erste Order aus dem Inputfile extrahiert. Eine Order weist einen der drei Typen Identifier, Instruction oder Literal-Code auf. Diese sind mit folgenden RegEx definiert.

identifier	<identiferChar>*
instruction	# <identiferChar>*
literalCode	".*"

<identiferChar> = [^# "<whitespace>]
<whitespace> = SPACE | NEWLINE | TAB

Es ist hierbei möglich bestimmte Orders **voraus zu stellen**, die anstelle der nächsten Order im Inputfile gefetched werden. Dies geschieht mit Hilfe des Fetch-Stacks auf den eine Liste an Orders gepushed werden kann. Dieser Fetch-Stack folgt Last-In First-Out und **auf ihn** kann während jeder der drei Schritte des Zyklus, meist jedoch während Execute, gepushed werden. Die Hauptanwendung des Fetch-Stacks wird im Abschnitt 4.3 ausgeführt.

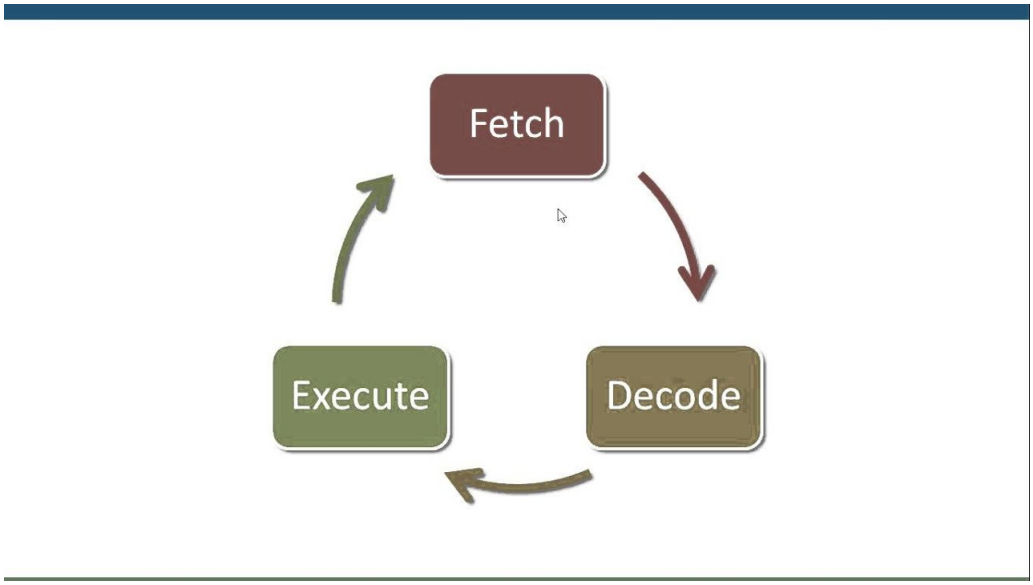

Abbildung 4.2: TEMP! Struktur des Fetch-Stacks

Wenn eine Liste an Orders komplett gefetched wurde, wird diese vom Stack gelöscht. Der Inputfile befindet sich auf dem letzten Platz des Fetch-Stacks und wird somit nur verwendet, wenn der Stack ansonsten komplett leer ist. Die Compilation wird beendet, sobald keine Order mehr auf dem Fetch-Stack übrig ist.

4.2 Decode

Nachdem eine Order gefetched wurde, wird diese an Decode weitergegeben. **Während Decode werden zwei Aufgaben durchgeführt**.

Während dem Decode Schritt kommt die Order-Queue ins Spiel. Hierbei handelt es sich um eine Liste an Orders, der Form First-In First-Out. Die Anwendung Aufgabe der Order-Queue ist das Speichern und spätere Ausführen von Orders. Die Order-Queue kann mit Hilfe von Instructions, die im Abschnitt 4.3 weiter ausgeführt werden, aktiviert und deaktiviert werden. Wenn nun eine Order in den Decode Schritt gelangt und die Order-Queue aktiviert ist, wird diese Order der Order-Queue hinzugefügt. Der Execute Schritt wird danach übersprungen und der Zyklus beginnt von neuem bei Fetch. Die Order wurde ohne ausgeführt zu werden auf der Order-Queue gespeichert. Später ist es nun möglich diese Order mit Hilfe von Instructions, die im Abschnitt 4.3 weiter thematisiert werden, von der Order-Queue zu entfernen und auszuführen. Bestimmte Instructions und Identifiers können jedoch Order-Queue-Proof, also immun gegen die Order-Queue, gemacht werden. Diese werden, auch wenn die Order-Queue aktiv ist, normal an Execute weitergegeben. Dies ist zum Beispiel besonders bei der Instruction, die die Order-Queue wieder deaktiviert, wichtig. Da diese Instruction sonst nicht ausgeführt und somit die Order-Queue nie mehr deaktiviert wird. LiteralCode kann nicht Code-Queue-Proof sein.

MAYBE CODESTACK FIGURE

Ist die Order-Queue deaktiviert oder die Order Code-Queue-Proof wird diese an den letzen Schritt Execute weitergegeben.

4.3 Execute

Execute ist der letzte Schritt des Zyklus. Und hier wird nun auch endlich der tatsächliche Assembly Code generiert. Je nach Typ der Order, Identifier, Instruction oder Literal-Code, läuft Execute sehr unterschiedlich ab.

4.3.1 Identifier

Ein Identifier ist eine Zusammenfassung von mehreren Orders. Diese sind in einem Environment definiert. Hierbei handelt es sich um eine einfach Map (Dictionary), **die einen Identifier als string mit einer Liste an Orders verknüpft**. Wenn nun ein Identifier in den Execute Schritt kommt, werden die dazugehörige Liste an Orders auf den Fetch-Stack aus Abschnitt 4.1 gepushed. Beim nächsten Fetch werden nun die zum Identifier gehörenden Orders zurückgegeben. Um Grunde wird der Identifier mit seinen Orders ersetzt.

Environments sind hierbei in einer Linked-List gespeichert. Somit können neue Environments zu dieser Liste hinzugefügt und von der Liste entfernt werden. Das unterste Element der Liste ist hierbei das älteste und das oberste Element das neuste. Bei der Definition eines Identifiers wird dieser immer zum obersten Environment hinzugefügt. Definitionen des gleichen Identifiers in älteren Environments werden nicht überschrieben oder gelöscht. Bei der Abfrage nach einem Identifier wird immer die neuste vorhandene Definition zurückgegeben. Ist keine vorhanden, wird ein Error ausgegeben.

4.3.2 Literal-Code

Literal-Code ist der Weg wie der QHS-Compiler Assembly Code generiert. Dieser ist sehr simpel. Wenn Literal-Code in den Execute Schritt gelangt, wird alles was zwischen den SZeichen steht in das Output-Dokument geschrieben.

4.3.3 Instructions

Instructions sind die komplexeste Order für den Execute Schritt. Für jede Instruction ist im QHS-Compiler eine Funktion definiert, die ausgeführt wird, wenn diese Instruction in den Execute Schritt gelangt. Diese Funktionen können Variabeln im QHS-Compiler speichern, den Order-Queue aktivieren, Identifier definieren und noch viel mehr. Folgend sind ein paar der wichtigsten Instructions aufgelistet. Für manche Instructions ist nach ihrem Namen ein Zeichen als Abkürzung angegeben. Diese Abkürzungen sind in QHS als Identifier definiert, die für diese Instruction stehen. Diese Abkürzungen werden in weitern QHS Beispielen verwendet.

#enterOrderQueue [	Aktiviert die Order-Queue.
#exitOrderQueue]	Deaktiviert die Order-Queue.
#assign »	Das erste Element der Order-Queue muss ein Identifier sein. Der Rest der Orders auf Order-Queue wird als Definition für diesen Identifier festgelegt.
#assignToOne ->	Wie #assign, jedoch wird nach dem Identifier nur eine weitere Order von der Order-Queue genommen und als Definition für den Identifier verwendet.
#force !	Die nächste Order wird nach Fetch sofort an Execute weitergegeben. Überspringt Decode und somit die Order-Queue.
#orderEnqueue	Die nächste Order wird sofort der Order-Queue hinzugefügt, auch wenn diese Order Order-Queue-Proof wäre. Execute wird übersprungen.
#orderFrontEnqueue	Ähnlich wie #orderEnqueue. Die Order wird jedoch auf den obersten Platz der Order-Queue gesetzt.
#deepFetch	Wird mit der ersten Order der zweitobersten Liste an Order auf dem Fetch-Stack. Ermöglicht den Zugriff auf den Inputfile innerhalb einer Identifier-Definition.
#pushEnv	Ein neues Environment wird der Environment Linked-List hinzugefügt.
#popEnv	Das neuste Environment der Environment Linked-List wird gelöscht.

Der QHScompiler umfässt **28** Instructions, wobei **5** dieser nur für Debugging des Compilers dienen.

4.4 Bringing it all together

Und das war's. Dies ist der gesamte QHScompiler. Im Vergleich zu einem traditionellen Compiler wirkt der QHScompiler fast schon **armselig**. Und dies hat einen einfachen Grund. Der QHScompiler ist zwar **vollendet**, die dazugehörige Sprache QHS jedoch noch lange nicht. Es ist zwar grundsätzlich durch LiteralCode möglich jedes Programm QHS zu schreiben, jedoch handelt es sich dann dabei einfach nur um Assembly Code. Doch der Aufbau des QHScompilers ermöglicht es mithilfe von Identifiern eine komplexere Programmiersprache zu definieren.

COMMENTS PSEUDO CODE IN EXAMPLES

4.4.1 Shortcuts

Shortcuts like ; and stuff

4.4.2 Identifier Parameters and Return

Parameter and Return through OrderQueue or PutInFront lateArg» ?

4.4.3 Variablen

Die Umsetzung von Variablen in QHS ist simpel. Zuerst soll die Grösse der Variabel vom Stack abgezogen werden. Dann wird ein für den Identifier ein Variabelnamen definiert, der zu der Position der Variabel auf dem Stack zeigt. Mit nur LiteralCode in QHS lässt sich dies wie folgt ausdrücken:

Listing 4.1: THING

```
"sub 4" ;
[ a "[rbp-4]" ] >> /* a = [rbp-4] */

"add " a ", 5" ;
```

Dieser QHS Code wird zu folgendem compiled:

Listing 4.2: THING

```
sub 4
add [rbp-4], 5
```

Jedoch ist dies noch nicht besonders angenehm. Weiter lässt sich zum Beispiel ein Var Identifier definieren, der die Grösse der Variabel als Argument annimmt OrderQueue. Um die in vielen Programmiersprachen geläufige Syntax einer Variabel Definition beizubehalten, wird der Variabel Name als lateArg» **angegeben**. RBP-OFFSET

EXAMPLE

Zuletzt lässt sich das umständliche Hinzufügen der Grösse der Variable sowie der Var Identifier unter einem Identifier zusammenfassen. Dies wäre passenderweise die bekannte Bezeichnung für den Variabel Typen.

EXAMPLE

So sieht eine Variabel Definition genau so aus, wie es in anderen Programmiersprachen gebräuchlich ist. Der Shortcut ; ist hierbei optional.

4.4.4 Funktionen

Funktionen sind im Vergleich zu Variablen deutlich komplizierter. **Daher soll hierbei nur auf zwei der Problematiken an Funktionen behandelt werden.**

Zum Schluss sollte eine Funktionsdefinition wie folgt aussehen:

EXAMPLE

Hier lässt sich bereits ein erstes Problem feststellen. Im vorherigen Abschnitt 4.4.3 wurde der Int Identifier für eine Variabel Definition verwendet. REF-TO-EXAMPLE würde vom QHScompiler also als eine Variabel Definition verstehen. Der Unterschied zwischen Funktions und Variabel Definition besteht hierbei in den Klammern die auf den **Namen** folgen. Der QHScompiler müsste also beim Int Identifier nach vorne

schauen, ob sich eine Klammer nach dem **Namen** befindet, und folglich eine Variabel oder Funktionsdefinition ausführen. Dies ist jedoch aufgrund des einfachen Designs des QHScompilers nicht möglich. Er kann bloss Orders ausführen, nicht jedoch überprüfen, ob eine Order vorhanden ist. Glücklicherweise lässt sich dieses Problem jedoch lösen, ohne eine Änderung am QHScompiler vorzunehmen. Die Lösung basiert darauf beim Int Identifier sowohl eine Variabel als auch eine Funktionsdefinition vorzubereiten, aber keine der beiden bereits auszuführen. Weiter wird nun eine Klammer als Identifier für eine Funktionsdefinition definiert. Sowie ein Semikolon als Variabeldefinition. Befindet sich nach dem **Namen** also eine Klammer, wird eine Funktionsdefinition ausgeführt. Ist dort aber ein Semikolon wird eine Variabeldefinition durchgeführt. Dieses Konzept wird weiterführend als DelayedExecute beschrieben. Das Ganze könnte dann wie folgt aussehen:

EXAMPLE

Das zweite Problem sind die Parameter eine Funktionsdefinition. Diese sehen genau gleich aus wie eine Variabeldefinition, sollen jedoch vom QHScompiler anders ausgeführt werden. Erstens sollte bei einer Parameterdefiniton nicht der LiteralCode zur Subtraktion vom RSP hinzugefügt werden. Zweitens verwendet eine Parameterdefintion einen anderen RBP-Offset. Die Lösung hierzu liegt in einer Umdefinition des Definition Identifiers. Dieser ist momentan für sowohl für Variabel als auch Funktionsdefinitionen verantwortlich. Bei der Anfangsklammer der Funktionsdefinition wird der Definition Identifier neu definiert, so dass er eine Parameterdefinition ausführt. Die vorherige Definition geht Dank der #pushEnv Instruction nicht verloren. Bei der schliessenden Klammer wird #popEnv durchgeführt, und der Definition Identifier ist wieder für Variablen und Funktionen zuständig. Diese Lösung wird im folgenden TempAssign genannt. Das ganze lässt sich im QHS wie folgt umsetzten:

EXAMPLE

Nun fehlt nur noch eine Sache der Funktionsdefiniton, der Funktionsbody. Dieser ist vergleichsweise simpel. Beim Öffnen des Bodies wird das ein Assembly Label des Funktionnamens gesetzt. Der gesammte Code innerhalb des Bodies wird nun einfach ganz normal vom QHScompiler ausgeführt und an den Output Assembly File angehenkt. So lässt sich die Funktion durch einen einfachen Assembly Call ausführen. **Wie das Callen von Funktionen aussieht, soll hierbei nicht weiter betrachtet werden.**

Abbildungsverzeichnis

Literaturverzeichnis

[1] Bob Nystrom. A map of territory (mountain.png), 2021. [Online; accessed 2024-08-05].

[2] Wikipedia. Abstract syntax tree for euclidean algorithm. [Online; accessed 2024-08-05].