Creating a Concurrent Compiler in Haksell

Luke Jackson

April 24, 2015

Abstract

Creating a compiler and executer in Haksell for Goal, a simple concurrent programming language of my own design based on Google Go. This paper can be treated as a introduction to implementing a compiler in Haskell, but a basic understanding of Haskell and Monads is assumed.

Contents

Chapter 1

Introduction & Overview

1.1 Introduction To Project

For my final year project I have created a compiler in Haskell for a simple programming language of my own design, which I based on Google's relatively new language Go. I called this language Goal, and it is a simplified version of Go but contains all of the features I was interested in implementing in my compiler, most importantly allowing concurrent programming.

The main focus of this project was to implement a parser, compiler and executor for a modern concurrent programming language, not on designing a completely new language. Though I will briefly discuss the reasons behind creating Goal, it is important to understand the focus of this project is on compiler implementation not programming language design.

This document is split into five main sections; Designing Goal, Parsing, Code Generation & Intermediate Representations, Code Execution and Testing. In each of these sections I will go into more detail about how I approached each of these problems.

These next few pages will give a brief introduction to compilers and programming languages, also giving more information about Go and the motivation behind the project. I will also briefly discuss any development methodologies used and give a justification for certain technical decisions.

1.1.1 Introduction to Compilers & Executors

To understand what a compiler is, you need to first understand what a program is and, more importantly, what a programming language is. Programs have become a fundamental part of human existence. From performing simple calculations to creating applications that allow me to create documents such as this, they exist in every aspect of our life. The simplest definition of what a program is, is to think of a program as a recipe.

You have a list of resources that you need, followed by a series of instructions telling you what to do with these resources. Then, to run your program, you

grab all the resources you need and follow the instructions.

Programs that run on computers have to be represented in someway that both humans and eventually computers can understand. Therefore programs are written in programming languages, a good definition for programming languages can be found in [Aho et al.(2007)Aho, Lam, Sethi, and Ullman, p, 1] where they state;

> Programming Languages are notations for describing computations to people and machines ... But, before a program (in this format) can be run it first must be translated into a form which can be executed by a computer.

So this brings us nicely to being able to define what a compiler is, a compiler is something that takes programs written in one programming language and translates it into another programming language. Often taking a program written in a high level language, that humans write code in, and translating them into lower level languages that computers can more easily understand and then execute.

Finally an executor is something that executes code written in a given programming language. You will usually only see executors created to execute low level languages given the smaller, and more precisely defined, instruction set. As is the case with the executor I have created. This highlights the need for compilers, so that they can be used to translate high level languages into more easily executable lower level languages, that are much harder for humans to create programs in.

Looking at these explanations it is clear to see the importance of compilers in the modern world. They allow people to create complicated programs and applications efficiently without having to concern themselves with low level details.

Introduction to Compiler Structure

1.1.2 Introduction to Goal

Goal is the language I have created my compiler for. It is a language I have created specifically for this project as a means of exploring compilers. It draws heavy inspiration from Go and uses the same basic syntax where possible. The main features of Goal are the ability to perform function recursion and to create and run concurrent programs.

The idea for Goal came when i was picking features from Go I was interested in compiling then decided it would be a nice idea to bring them together to create a more complete language. In my opinion during the course of this project Goal has evolved from a means to explore compilers and language design into a simple standalone language with it's own uses. I will discuss more about the design and functionality of Goal in chapter 2.

1.1.3 Introduction to Go

Go is an object oriented programming language created by Google relatively recently in 2007. It is used by Google for many different applications, most notably it powers dl.google.com, a service which contains the source for Chrome and The Android SDK. A good place to find out more about the history of Go and it's development is in the documentation section on golang.org.

In many ways Go is similar to C in terms of syntax and that it is also statically typed. There are however some interesting features thrown in, such as how Go handles concurrency and even some newer ideas such as splices, which we will talk about later.

In summary I would say Go (also referred to as golang) is a new language with quite a bit of potential, which can be seen by it's growing popularity. Although it is not revolutionary, it's mixture of simplistic syntax and more exciting features makes it a good language to emulate.

1.1.4 Introduction to Concurrency

Concurrency is the primary feature I was interested in implementing in my compiler. It is a key part of this project and also a fundamental part of modern programming.

The basic idea behind concurrent programming is you can do many things at one time. If we first define a sequential program as a sequence of operations carried out one at a time. We can then define a concurrent program as a program that contains a set of sequential processes executing in parallel [Terry(1997), p. 414].

The importance of concurrent programming can be seen now more than ever with rising popularity of online services and cloud computing. Lets take one of the most popular websites in the world Google.com, with around 40,000 request a second there is a high chance that when you hit that search button, so are thousands of other people at exactly the same time. But instead of queuing every request and doing it sequentially a server will handle the requests concurrently, ensuring that many requests can be handled at the same time, and you don't have to wait for the thousands of people who clicked a few milliseconds ahead of you till you get your results.

There are plenty more examples of the importance of concurrency in modern technology, I'm sure if you think about most pieces of technology you use it would be easy to list numerous process that have to run concurrently. This is why I was interested in implementing concurrency into my compiler, because of it's usefulness and importance in modern programming languages.

1.2 Motivation

The biggest motivation behind this project was a desire to learn more about compiling and executing modern object oriented programming languages. I was

also curious about programming language design and wanted to take the chance to really look at a programming language analytically.

1.2.1 Why Haskell

I chose to implement my compiler using Haskell. Haskell is a purely functional language with strong static typing. Functional languages are often seen as good platforms to use when creating a compiler because some of their features make it easier to handle tree data structures, which can be important during parsing and creating an intermediate representation of language. Also the use of pattern matching and efficiently being able to recurse makes Haskell a good choice for implementing a compiler and a virtual machine.

Another reason I chose Haskell was quite simply that it is a language I enjoy working with, and more importantly for a project like this, would like to learn more about. I was keen with this project to not only create something interesting but also learn more about functional programming. I hoped to use this project to explore some unique functional approaches to some of the problems creating a compiler can throw up. A good example of this is how I handled parsing and my use of Monadic Parser Combinators.

1.2.2 Why Create A New Language

There are so many great programming languages out there with fantastic supporting documentation that would be excellent choices to make compilers for. Which really begs the question, why did I go through the hassle of designing my own language to compile? The answer to this is pretty simple, with a project like this if I was to create a compiler for C++, for example, I would never have the time to create something that covers all of C++'s functionality within 9 months. Therefore no matter how well I had done with the project it would always feel a little incomplete, but more importantly if someone was to use my compiler they could very well end up trying to use features my compiler didn't support which would be frustrating for me and anyone who wishes to use my compiler in the future.

Instead of having this problem I decided I would create a new language that I could provide supporting documentation for and strictly define what is and isn't possible with in it. Because the focus of this project was not on language design my approach to creating a new programming language was finding a language I liked, then picking key features I was interested in exploring. Before finally adding some extra functionality so that this new language could be useful on its own.

The language I liked the look of was Go and that is where most of the features for my language came from, with the key feature I was interested in exploring being concurrency. I decided to call the language I had created Goal.

1.2.3 Why Base This Language on Go

Go is not a language I was overly familiar with before the start of this project, but as soon as I looked into it I very much liked what I saw.

At the start of this project I was exploring different possibilities of languages I could use as inspiration for Goal. When I came across Go I discovered it had a very clean and easy to understand way of creating and running concurrent process, which was one of the key features I was interested in putting into Goal. From there I began looking into Go's design and discovered not only did it have a wealth of interesting features to look at but also a clean and understandable syntax, which would work nicely with the simplistic easy to use nature I wanted to create in Goal.

The most important factors where defiantly the way Go handle concurrency and channels but also there was an aspect of using this project as a way of familiarizing myself with another language. It also helped that Go came with such a wealth of easy to navigate online resources which was perfect for when I needed to check the exact functionality of certain expressions.

Overall I felt Go provided me with a simple syntax and good implementations of the main features I was interested in implementing. Which made it a good choice to use as the basis of Goal.

1.3 Development Process

The main method I used for development was Test Driven Development (TDD). This is a simple approach where you write your test cases first, then write your code so that is passes all your tests. I found this to be quite an appropriate approach due to the iterative nature of my development process.

I initially started work by focusing on being able to compile and execute a small subset of simple features such as if statements, variable assignments and simple arithmetic. Then using TDD I wrote tests for each new piece of functionality I wanted to add, expanding the subset of features I was able to handle. Due to the nature of TDD I was able to do this without worrying about breaking earlier features as I could just ensure all my original tests still passed.

As was mentioned before, I split the implementations of my compiler into three main sections;

- Parsing
- Code Generation & Intermediate Representations
- Execution, using a Stack Based Virtual Machine

While developing my compiler and executer I often found I would implement each feature in two steps; First ensuring I could create a suitable intermediate representation and updating my virtual machine to ensure it could handle the new feature. Then, once this was complete, I would update my parser to handle this new command.

Often when working on multiple features at a time, it would be more efficient to update the code generation and virtual machine first for all the new features. Then update the parser second, rather than adding each new piece of functionality one at a time.

I will go into more detail about the creation and running of my tests in Chapter 5.

Chapter 2

Designing Goal

This chapter deals with how I approached designing a new language and how I went about choosing the features I wished for Goal to contain. I will discuss some reasoning behind why certain features differ from how they are handled in Go and also talk about how I tried to make the language as complete as possible.

I think it is also important to note here that the main focus of this project is not on the design and creation of a new language but rather the implementation of the compiler itself.

2.1 Picking Features

As I mentioned before, rather than designing Goal from scratch I chose to instead choose features from Go I was interested in, then find a way of putting them into Goal. I did not create Goal with a target audience or with potential uses in mind, hence the lack of a specification or market research. Instead I created Goal as a means to let me implement a compiler that could handle a number of different interesting features.

Therefore you will find the features Goal can handle my seem quite varied and not completely complimentary of each other, but I do feel that Goal still has many uses.

As much as Goal was created simply as a tool for creating an interesting compiler, I did also attempt to make it as user friendly as possible. I have created supporting documentation for Goal that has full examples of the syntax it uses and detailed explanations of how to use each of it's features.

A good summary of my approach to creating Goal is that I filled it with features I wanted to explore, then added extra functionality to try and make the language as easy to use, complete and useful as possible.

2.1.1 Syntax

A language's syntax can be said to describe the form that commands and expressions in a language must take [Terry(1997), p. 72]. In the case of programming languages it defines how you must write your code so that it can perform the computations you wish.

The syntax for Goal is pretty simple and almost identical to that of Go, with some minor differences. I decided to follow Go's syntax rules not only for simplicity but also so it was easy to see where Goal got it's functionality from. I felt it would make sense to give it syntax rules close to the language it was emulating.

A key part of Goal's syntax is that each command must end with a semicolon, including if statements, functions and for loops. A more detailed outline and examples of valid Goal syntax can be seen in the accompanying documentation for Goal with examples of syntactically correct Goal [Jackson(2015)].

2.1.2 Types and Scope

A type system defines what type of variables are allowed to be used in a language. A variables type can be defined by saying [Cooper and Torczan(2012), p. 164];

> The type (of a variable) specifies a set of properties held in common by all values of that type ... For example, an integer might be any whole number i in the range $-2^{31} \leq i < 2^{31}$.

Go has a very broad type system and static typing. Static typing is where any type errors are checked during compiling as opposed to dynamic typing where type errors are checked at run time. By saying Go has a broad type system I'm saying it allows for a lot of different types.

In Goal I decided not to focus on allowing lots of different types and instead only allow 3 main types; Integers, Booleans and Strings. This was mainly so I could spend more time working on other features and not worrying about creating an extensive type checking system, whilst still having enough types to be able to create some interesting programs.

This also lead to allowing very basic variable declarations and assignments. You do not need to declare a type when initially declaring a variable and variables can be created on the fly. For example if you were to write;

```
i = 4;
i = True;
i = "hello";
```

There would be no compiler or run time errors in this code and your variable i would have the value "hello" as each new assignment writes over the old value. This is the biggest difference Goal and Go, it is actually much closer to how a language like Python handles variable declarations and puts more emphasis on the user keeping track of their variables.

A variable's scope defines where it can be accessed from. In Goal I decided to keep this very simple by only allowing variables to have two different scopes; global or local. A global variable is defined by using the keyword global before the definition and once declared can be accessed anywhere in your code. A local variable is just declared as shown above with no key word and can only be accessed with in the function you declared it in. More information about how variable scope works in Goal can be found in the accompanying documentation.

2.1.3 Basic Commands

By basic commands I mean the set of statements you expect to find in most object orientated languages. In Goal this includes;

- If Statements
- While Loops
- For Loops
- Output Commands
- Return Statements

All of these almost exactly follow the same syntax rules and have the same functionality that you find in Go. There are some small differences such as in Goal you are required to hold your conditional expression in brackets in each command but other than that there is not much difference. Again for more clarification on how these statements work you can find examples in the Goal documentation.

2.1.4 Functions

Functions are treated very similarly in Goal as they are in Go. The main difference is that you can only have functions that either return Booleans or Integers, or your functions can be void. Function recursion is allowed as are nested function calls.

2.1.5 Concurrency

For concurrency I wanted to use the same simple technique that allowed you to run any void function as a concurrent process with the use of a keyword. Meaning if you wanted to run two functions $funA()$ and $funB()$ concurrently all you needed to do was write;

```
go funA();
go funB();
```

This code would start both the processes as independent subroutines and would execute them in parallel. I also added some useful library functions that I felt would be helpful for creating concurrent programs. These are not found in Go but are important to the way Goal works. The two most important ones are $Wait()$, which will halt a program until all subroutines have finished executing, and $Kill()$ which will immediately stop all running subroutines. There are a few other library functions I have included and again more information can be found in Goal's Documentation.

I also needed to implement a way to pass information between subroutines, although you could use global variables it is safer to have a more controlled method. This is why I implemented channels. These are also present in Go. You can declare a channel anywhere and it has global scope. They behave like stacks so you can push a value onto a channel in one subroutine then pop it out in another subroutine. Channels have a first in first out system. I have implemented channels almost exactly the same as how they are implemented in Go.

2.2 Differences From Go

Although the language I have created is based on Go there are quite a few noticeable differences. The biggest difference is that Go is statically typed and Goal I have made dynamically typed, although there is a very basic type checker in place at compile time. This mainly means that if you convert your Go code to Goal you need to take care that you don't overwrite any variables.

Other noticeable differences come from the output functions being part of Goals standard library where as in Go you eed to import a package to output to a console.

Obviously the scope of Goal is much smaller as it has a much more limited type system and does not allow for multiple classes or files. Though it is noticeably similar and does contain many of the key features that Go contains

2.3 Possible Uses

Towards the end of this project I began to reflect on Goal as a standalone language and not just as something for me and felt it could have some uses of it's own.

Where Go is most popular at the moment is for a

Although I don't think Goal has a future competing with Go as tool for implementing large scale servers. I do feel that given Goal's simplistic syntax and easily understandable ,and usable, features it could have some use as a tool for teaching programming. More specifically helping people to understand how concurrent programs work, and giving them the tools to create their own exciting concurrent programs.

You can see I have provided with my project some example programs and as part of that I have included some programs I feel highlight how Goal could be used as a teaching tool in the future. I will revisit this idea as part of my evaluation of the project as whole.

Chapter 3

Parsing

Parsing is a fundamental part of compiling, you can think of it as the front end of the compiler. In simplistic terms it takes ina program as raw text and then builds a data structure that represents the program the user has written. In my project the raw text is the code from a .gol file and the data structure is the intermediate representation of Goal.

Although I will be talking a lot about the intermediate representation I've created to parse my text into, I will not be going into much detail about the design of that data structure. That is covered in more detail in the next chapter. In this section a preexisting understanding of Monads in Haskell is assumed.

3.1 Introduction to Using Monadic Parser Combinators

The technique I used to parse data in my compiler was to make use of Monadic Parser Combinators. If we think of a parser in Haskell as something that takes in a string and returns a data structure, we can think of parser combinators as high order functions that take in several parsers as its input and returns a new parser as it's output.

3.2 Goal Syntax Rules and Justifications

As was discussed in the previous structure Goal gets most of its syntax rules from Go. There are however some unique syntax rules I decided to implement in Goal that you will not find in Go. There were two main reasons why syntax rules my differ from Go.

The first being it was a design choice. For example if you look at variable declaration in Go it can be done one of 2 ways.

```
var i int = 42
j := 42
```

Although I could have implemented variable declarations to look like this because I decided because variables can be declared on the fly and because of the simplistic nature of my type system I may as well keep declarations and assignments the same, and as simple as possible. Hence in Goal all you need to write to declare or assign a variable is;

```
l = 42;
```

There are several examples of these changes in syntax for design choice, such as global variable definitions and the requirement of brackets to hold conditional statements. These small changes where made merely as a way to tidy up Goal and make it have a more complete and consistent syntax.

The other reason Goal's syntax varies from Go is because of the limitations within the implementation of my parser. For example, in Go you don't need to use semi colons at the end of a command. Where as I decided that in Goal I would make it necessary to include semi colons at the end of every command, including if statements and function declarations. This is because it makes it easier to split up commands based on every time I see a semi colon. The reasons I made some of these choices was not because it was not possible for me to implement different syntax rules, but because the main focus of this project was not on parsing and if a small change to syntax meant a quicker implementation sometimes I felt it necessary.

3.3 Parser Implementation

3.3.1 Example of Parser Implementation

3.3.2 Analysis of Parser Example

3.4 Potential for Expansion

Chapter 4

Code Generation & Intermediate Representation

Code generation is where I take the parsed input and then generate code from a low level instruction set that can then be executed. I decided to implement my own low level language and create an executor alongside that. Thus meaning that I would be compiling the parsed input down to code that used an instruction set that I had defined.

A key part of this section is looking at how I designed and implemented an intermediate representation of Goal in the form of a data structure. This intermediate representation was then simpler to compile into my low level instruction set. The importance of a good intermediate representation can be seen in [Cooper and Torczan(2012), p. 221] where they state;

> Most passes in the compiler read and manipulate the IR form of code. Thus, decisions about what to represent and how to represent it play a crucial role in both the cost of compilation and it's effectiveness.

In this section I will go into detail about the design of the data structure that you are required to parse down to, and which represents all of the features I implemented in my language. I will also be discussing some of the more interesting features that I implemented and how they were dealt with by the code generator.

Although in this section I will be talking a lot about code generated using the instruction set defined in my executer, I will not be going into detail about the design of the executor or the low level language I created. For more information on this you can go to the next section that does focus more on the design and execution of the low level instruction set.

4.1 Intermediate Representation

Having a good intermediate representation of the program being compiled is very important part of creating a compiler. It is important to make sure that you do not misinterpret the program you are compiling and a good way to do this is to create a data structure that clearly represents what your program is doing, whilst ensuring this data structure is easy enough to compile into your target language.

4.1.1 Introduction to Intermediate Representations

An intermediate representation of a program is where you create some data structure of your program that can be more easily interpreted and handled by your compiler. It is also important to ensure your IR[1] is designed in such a way that you capture the meaning of the program you wish to represent, and do not end up misrepresenting your program.

There are several approaches to generating an IR. These can been seen in [Cooper and Torczan(2012), p. 223] where they state there are three main approaches to generating an IR;

- Graphical IRs, This uses tree or graph data structures.
- Linear IRS, This will closely resemble pseudo-code.
- Hybrid IRS, A combination of both Linear and Graphical approaches.

I chose to go with a Hybrid IR. A most suitable approach for me because, as stated in [Chattapoadhyay(2005), p. 113], Linear IR are often used when compiling for stack base virtual machines, which is the architecture of the virtual machine that I created. But also this approached allowed me to take advantage of how easily Haskell allows you to create recursive data structures, that can be used when generating trees you would find in Graphical IRs.

This approach provided me with a relatively high level representation of programs, meaning it gave a representation that would closely resemble the pseudo-code of the program you are trying to compile. This makes it nice format to work with as it makes it easier to visualize what is actually occurring during the later stages of code generation.

It is important to note there are several different approaches I could have used when creating an intermediate representation and the choice to use a high level approach was both a design choice and also to do with ease of implementation.

To best understand the approach I used to create the data structure for my intermediate representation it is good to follow a simple example of how you take a specific command then translate that into my intermediate representation.

[1] IR is shorthand for Intermediate Representation

4.1.2 Example Creating an Intermediate Representation

I will now show an example of the process I went through when generating my IR. The example I will use is creating a data structure that best represents an if statement. This is what an if statement looks like in Goal;

```
if (x < 1) {
        return x;
};
```

You can see that there are two main parts to this statement. The condition directly after the if command, $(x > 1)$ and then the code ,in this case *return* x;, inside of the curly brackets.

What is important to notice is that you could replace the return statement with any other acceptable code, even another If statement. Whereas the expression after the if is limited only to be certain things. In this case we could rewrite a general definition for an if statement to look like this;

```
if EXPRESSION {
        PROGRAM
}
```

This can be described by saying; If some expression is true, then execute the program inside the curly brackets.

Now all we need to do is define what EXPRESSION and PROGRAM can be. Then we can create a data type in Haskell code to represent If statements that looks like this;

```
data IfStatement = If Expr Prog
```

A program is quite easy to define, it's just a one or more instructions written to perform a specific task[2]. Therefore we must define all of the instructions that can be used to make a program in our data type Prog. Looking at our example that needs to include return statements, but I also said we were allowed to have nested if statements. Therefore we can create a recursive data structure that will allow this;

```
data Prog = If Expr Prog | Return Expr
```

An expression is slightly harder to define. In the case of If statements we know our expression needs to do something. It needs to give us some condition which we can then decide is true or false.

We can now create a data type for our expressions. We will need to allow for nested expressions but also allow for different comparison operators such as equals or less than. This can be shown here;

[2] $wikipedia.org/wiki/Computer_program$

```
data Expr = ExprComp Op Expr Expr
          | Val Number
          | Var Name

data Op   = GET | LET | NEQ | EQU
```

Although this example gives us quite a strict definition of what an expression needs to be for an If statement later on we will need to come up with a more general definition that is more applicable to other instances of using expressions (such as in arithmetic operations). The language specification for Go describes an expression by saying; "An expression specifies the computation of a value by applying operators and functions to operators". Another way to think of an expression is it is any valid unit of code that resolves to a single value [3].

[3]

Now that we have our definitions for what Expressions and Programs can be, we can begin to group what instructions belong in which section. With all our expressions being defined in the data type Expr and all our instructions used to make programs in the data type Prog.

Therefore if we are only considering our If statement example from earlier (ignoring how we defined the variable x) we end up with the following data structure;

```
data Prog   = If Expr Prog | Return Expr

data Expr   = ExprComp Op Expr Expr
             |Val Number
             |Var Name

data Op     = GET | LET | NEQ | EQU

type Name   = Char

type Number = Int
```

The above data structure now enables you to create high level intermediate representation of the original If statement we looked at, through the use of Haskell data types.

You can see here how the data structure has a broad scope actually allowing expressions such as $(4 < 6)! = (4 < (5 > 6))$, which don't really make sense. But I felt it was better to allow for these bizarre expressions, and handle any stricter rules in parsing stages. Rather than create an IR that was too restrictive.

4.1.3 Analysis and Expansion of Creating an IR Example

The key point of the above example is to understand that each part of a languages features can be, and needs to be, strictly defined. I found it important to

[3] url for quote

remember the data structure I am creating has a purpose other than just being a new representation of Goal. It needs to actually extract the important pieces of information from the code, such that you are left with a series of concise statements that hold all the information necessary for you to start to rebuild the program using a new instruction set.

The example also highlights the process of abstracting out each part of a languages functionality, then choosing the appropriate structure to use to represent them. An important part of this example is showing the use of recursive data structures to generate tree like data types.

For example look at the code;

```
if (x > 1){
        if (x > 5){
                if (x == 6) {
                        return 1;
                };
        };
};
```

It will be represented using data structure from our example by the following Haskell expression;

```
If (ExprComp GET (Var 'x') (Val 1))
        (If (ExprComp GET (Var 'x') (Val 5))
                (If (ExprComp EQU (Var 'x') (Val 6))
                        (Return (Val 1))))
```

This can also be visualized in a tree structure with branches to the right being the outcome if the condition is true and branches to the left if the condition is false;

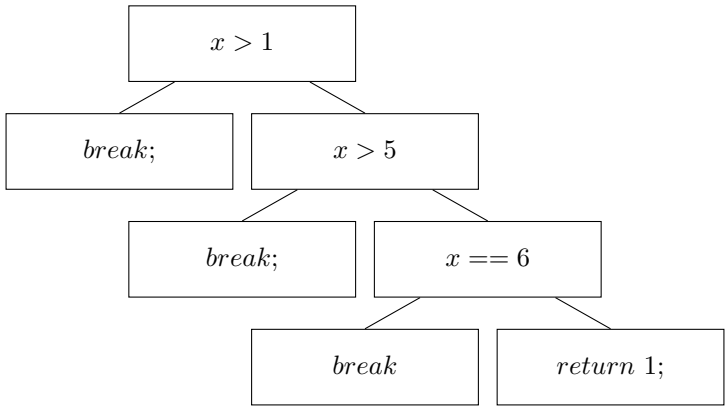

Figure 4.1: Visual representation of tree structure of nested if statements

What figure 4.1 shows you is the importance of using recursive data structures when allowing an arbitrary number of nested commands. You will see that

as you start to expand your data structure to include multiple function declarations, or even large conditional expressions, building a tree like data structure is a good way to represent any program.

Overall this example helps to introduce all the key concepts behind the creation of the IR I used in my compiler.

4.1.4 My Intermediate Representation of Goal

4.1.5 Handling More Complex Features in my IR

4.2 Code Generation

4.2.1 Introduction to Code Generation

4.2.2 Brief Introduction to Target Language Instruction Set

4.2.3 Example Code Generation

4.2.4 Analysis and Expansion of Code Generation Example

4.2.5 Examples of Generating Code For More Complex Features

Chapter 5

Execution using a Stack Based Virtual Machine

This chapter goes into detail about the method I used to execute the code generated by the code generator. The previous sections have mainly focused on compiling programs down to an instruction set. In this section I will go into detail about the design of that instruction set and the method I used to execute it.

When creating a compiler you must take in a source language, the code you wish to compile, and then output a target language. A target language is the language you wish to compile into. In simpler terms, your compiler is taking in programs in your source language and then outputting the same programs but they are now represented using your target language.

For a project like this there are two approaches you can take when choosing what your target language should be. You can either find a preexisting low level language like ARM code or Java Byte Code. Or you can create your own instruction set and a virtual machine that is capable of executing those instructions. I decided to make my own instruction set and a virtual machine to execute it.

5.1 Introduction to Stack Based Virtual Machines

There are two things too define here before we can start talking about stacked based virtual machines . We need to first define what is a virtual machine, then secondly what is a stack machine.

A virtual machine can be described as a self contained operating environment that behaves as if it is a separate computer[1] . This can be useful because it means no matter what machine you are running your VM[2] on any application you run on your VM will run the same regardless of the physical machine you

are using.

A good example of a virtual machine is if you have ever played a retro games console emulators on your phone or computers. Those are examples of fully functional virtual machines running old programs independently of the machine they are running on.

A stack machine is a machine or (virtual machine) that uses a pushdown stack instead of set registers to evaluate different expressions[3]. A pushdown stack is a data structure with two main operations;

- Push, which puts something onto the stack.
- Pop, which removes the last element put on to the stack.

You can think of a stack like a PEZ sweet dispenser. The first bit of candy you put in goes straight to the bottom of your dispenser and if you put more in they get pilled in on top. This is the same way you push objects onto a stack. Then when you wish to eat your candy, the last piece you put in comes back out first. This is the equivalent of popping the stack where the last object you pushed onto the stack comes out first. This is called a last in first out system.

Therefore a stack machine is quite simply a machine that uses a stack instead of allocated registers to handle expressions. If you look at it shows you how the expression $2 - 1$ would be evaluated using a stack in a stack machine.

Now we understand what a stack machine is and what a virtual machine is it becomes quite easy to understand what a stack based virtual machine is. Quite simply it is a virtual machine that uses a stack architecture, effectively you can think of the virtual machine I created as a simple stack machine emulator.

5.2 Implementing a Stack Based Virtual Machine in Haskell

The way the executor I crated works is it takes in code written from a simple instruction set then the virtual machine will execute the code one instruction at a time. For a quick overview of how my stack based virtual machine works you can look at the type of the main functions that handle executing code.

```
exec    :: Code -> String
```

Which simply starts a call to the recursive function;

```
exec'   ::exec' type goes here
```

[1] `http://www.webopedia.com/TERM/V/virtual_machine.html`
[2] VM is short hand for virtual machine
[3] `http://en.wikipedia.org/wiki/Stack_machine`

What looking at *exec* shows is all that is needed to start the executor is the code that has been generated by the code generator. Then looking at $exec'$ you can see the different data structures that are used in implementing my virtual machine in my executor. The main components are;

- The Code
- The Program Counter
- The Stack
- Memory
- Channels
- List of Subroutines

The code is simply referring to the code you are currently executing and the program counter tells you where in that code you are. The stack is used like registers would be, where you will hold values you are currently interested in handling. Memory is where you store any variables that you will need to refer back to. Subroutines and channels are used for concurrency and I will go into more detail about them later.

The key concept to take way is that memory is for storing objects for the long term where as the stack is for passing around values and holding them temporarily.

5.2.1 Explanation of Instruction Set

Chapter 4 shows how the code is generated from an instruction set, and in figure 5.1 you can see the data structure I created in Haskell that represents my instruction set.

To understand how this instruction set works it is good too look at a couple of examples of generated code from the instruction set and what exactly different instructions mean. Then we can move on to looking at more detailed example of how you go about building an executor for this instruction set in Haskell.

If we first look at a variable assignment. If we wanted to compile the code: $x = 7$; it becomes;

```
PUSH (Integer 7)
POP "x"
```

Or more accurately in Haskell, since Code has the type $[Inst]$, it would look like this;

```
[PUSH (Integer 7), POP "x"]
```

This uses two of the most important instructions. $PUSH$ Takes a number as an argument and pushes it onto the stack, POP takes a name as an argument and removes the head of the stack saving it in memory, overwriting any existing

```
data Inst  = PUSH Number
           | PUSHV Name
           | POP Name
           | SHOW
           | PRINT String
           | DO ArthOp
           | COMP CompOp
           | JUMP Label
           | JUMPZ Label
           | LABEL Label
           | FUNC FName
           | FEND
           | VCALL Name
           | CALL Name
           | STOP
           | RSTOP
           | MAIN
           | PUSHC Name
           | POPC Name
           | CHANNEL Name
           | WAIT
           | KILL
           | GO Name

type Label =  Int
```

Figure 5.1: Haskell data structure used to represent all the instructions in my instruction set.

variables with the same name in the same scope in memory. You can use $PUSHV$ which takes a name as argument to push a variable from memory onto the stack.

In code generation one of the few features of the instruction set I went into in any detail was the idea behind the $LABEL$ and $JUMP$ instructions. So you may remember that compiled code for the expression;

```
for (x < 5) {
        x++;
};
```

Will look like this;

```
[LABEL 0, PUSHV "x", PUSH (Integer 5), COMP LET,
JUMPZ 1, PUSHV "x", PUSH (Integer 1), DO ADD,
JUMP 0, LABEL 1]
```

To understand this better we can annotate the code to show what is going on at each instruction, this is shown in figure 5.2. It is a good way to explain several instructions that I will be using frequently throughout this chapter.

```
LABEL 0           -- places a label with name 0

PUSHV "x"         -- pushes variable x onto the stack

PUSH (Integer 5) -- pushes the number 5 onto the stack

                     performs "<" comparison between the
COMP LET          -- top 2 elements of the stack, places 0
                     ontop of the stack if it's false and
                     1 if true

JUMPZ 1           -- jumps to label 1 if the head of the
                     stack is 0

PUSHV "x"         -- pushes varaible x onto the stack

PUSH (Integer 1) -- pushes number 1 onto the stack

DO ADD            -- adds the top two items on top of the
                     stack, places result on top of stack

POP "x"           -- pops the item from top of stack and
                     saves it memory with name "x"

JUMP 0            -- Jumps to the label 0

LABEL 1           -- places label 1
```

Figure 5.2: Annotated example of what each instruction is doing in a simple while loop

This gives a more detailed insight into how $JUMP$, $JUMPZ$ and $LABEL$ are used to represent conditional expressions on a lower level. It also introduces the idea of how you perform operations involving two values, where you have instructions that pop the top two elements of the stack then push the result of the operation back onto the stack.

5.2.2 Example Code Execution

I will now run through an example of how I created an executer for my instruction set. To do this i will start with a data structure representing an much smaller set of instrucions and then show how I build a virtual machine capable of executing code built from these instructions.

The instructions I will use in my example can be shown in the following data structure;

```
data Inst  = PUSH Number
           | PUSHV Name
           | POP Name
           | DO ArthOp
           | JUMP Label
           | JUMPZ Label
           | LABEL Label
```

This is a relatively small set of instructions in comparison to what I used in my final project, but you should be familiar with most of the commands, as they were all introduced in the previous section.

Now we have our instruction set we can move on to defining everything we will need to make it work.

```
type Name   = String

type Number = Int

type Label  = Int

data ArthOp = ADD | SUB
```

Now we have some raw data types to use, we now need to think about what we will need in our virtual machine to make it work. For this example we will definitely need a memory and a stack. So we add;

```
type Code    = [Inst]

type Stack   = [Number]

type Memory  = [(Name, Number)]
```

It make sense that memory should be a list of tuples because memory is a series of values with an allocated name used to reference them. It also makes sense that our Stack just needs to be a list of numbers. We now need to decide what our executors type should be. I think its good if it takes in some code and returns us the stack. But the problem with that is how do we pass about all the components of our virtual machine. well we need too create a helper funstion thats going to do all the hard work. Seeing as we just need the code to start our executor we can write;

```
exec     :: Code Stack
exec c   = exec' c 0 [] []

exec'    :: Code -> Int -> Stack -> Memory -> Stack
```

So now we need to decide how we will handle our code. If our code is just a list of instructions then we can use our program counter to find the element in the list we should be dealing with then use a case analysis to say what to do depending on our instruction.

Lets pretends we have already defined some functions to speed things up.

- *pop* will take in the stack and return the stack after popping it
- *push* will take in the stack and a value, pushing that value to the top of the stack, then returning the updated stack
- *pushv* will do the same as *push* but takes in a name and memory and pushes the variable from memory to the stack
- *save* will take the head of the stack and a name, then save that value to memory
- *jump* will take in the code and a label number and return a program counter that is set to the location of that label.
- *jumpz* does the same as *jump* but takes in the stack as well only performing a jump if the head of the stack is 0, otherwise it just increments the program counter as normal.

We will also need to define a function *do* that will take take in the stack and an *ArthOp*. Then it will return the stack after performing the required operation on the top two elements of the stack and pushing the result on top. I will define this function below as an example, so as to give you an idea as to how some of the other functions could be implemented.

```
do       :: ArthOp -> Stack -> Stack
do o s  = case o of
             ADD -> push (v2 + v1) ns
             SUB -> push (v2 - v1) ns
          where
             v1   = head s
             v2   = head (tail s)
             ns = (pop (pop s))
```

Now that we have these functions we can easily create a recursive function to handle a long list of instructions.

```
exec'     :: Code -> Int -> Stack -> Memory -> Stack
exec'c pc s m
          = if pc >= (length s) then s
            else
              case c !! pc of
                POP n    -> exec' c (pc+1) (pop s) (save s m)
                PUSH v   -> exec' c (pc+1) (push v s) m
                PUSHV n  -> exec' c (pc+1) (pushv n m s) m
                LABEL l  -> exec' c (pc+1) s m
                JUMP l   -> exec' c (jump l c) s m
                JUMPZ l  -> exec' c (jumpz l c s) (pop s) m
                DO o     -> exec' c (pc+1) (do o s) m
```

The above function is a very basic example of how the main function in my executor works. Using recursion and pattern matching I am able to iterate over a list of instructions and create multiple functions to assist me in creating an efficient implementation of a virtual machine.

5.2.3 Analysis and Expansion of Code Execution Example

5.2.4 Memory Design and Implementation

Memory is handled very simply in this project it has the same type it was given in the example in section 5.2.2 and variables are referenced in almost exactly the same manner. The only major difference is that I included variable scope in my project and as such had to handle a number of different possible outcomes.

Handling Variable Scope

The main difference from the example is that memory in my executor is partitioned when it is initialized. My parser does not allow for variables the empty string as the name. Therefore to partition memory I initiate memory by creating a list of ten empty values , $(\text{""}, 0)$. Then all global memory is stored at the head of the list and local memory is stored at the tail, meaning memory is split in two. This makes it easy to handle scope because every time there is a change of scope, e.g. you begin to execute a function, you can simply drop everything after the list of empty variables as they will no longer be in scope.

This is a very simple approach to partitioning memory and one of the weaker aspects of this project. The current implementation does not facilitate the use of pointers, due to the complete dropping of the previous local memory on a change of scope, but this would not be too difficult to update my executor to handle. Though it may require restructuring the way I currently handle memory.

5.2.5 Stack Management

Manging a stack is a very important part of implementing an efficient and, more importantly, working stack based virtual machine.

Handling Function Calls

Handling function calls is one of the more interesting things to implement in a stack. If we first look at what happens in our executor when we hit the two commands that deal with function calls; either $VCALL$ (for void function calls) or $CALL$, a function called *handleCall* is called. The purpose of this function is to execute the function call before moving onto the next command. This seems simple enough but there are two complications; functions that take in arguments and functions that return values.

If we first look at handling arguments, one approach could be to change the instruction of $CALL$ to now incorporate arguments is as well like this;

```
data Inst     = ...
              | CALL Name [Argument]
              | VCALL Name [Argument]
              ...

type Argument = Number
```

This looks like a good solution at first, but what if we want to call a function using an expression or a variable as arguments. Calls like $fun(a+5,(4*b)-c)$; would start to get very messy to deal with.

A better approach would be to use stack frames. A stack frame is a frame of data that is put on top of the stack. In the case of function calls it means putting all the arguments onto a the stack, then ensuring they are popped off in the correct order.

If we look at what happens if we were to call a function that takes in 2 arguments we can see that both arguments get pushed onto the stack before calling the function, this is shown below.

```
fun(12, 30);     -->  PUSH (Integer 30)
                      PUSH (Integer 12)
                      CALL "fun"
```

Now to understand what happens in *handleCall* we must look at what the function does. In simple terms it calls $funExec'$, a function which behaves almost exactly as our main executor except with some limits to what can be done inside a function (mainly to do with concurrent processes), and then moves onto the next instruction by updating the program counter and calling $exec'$ after updating the stack and memory.

Lets look at the code to be executed now in $funExec$ if we see that function fun looks like this;

```
func fun(a int, b int){
```

```
            ...
};
```

So the way in which the arguments would be passed is to pass the stack into $funExec'$ then pop of the items in the correct order. This can be show here;

```
-main executor;            -function executor;
...
PUSH (Integer 30)
PUSH (Integer 12)
CALL "fun"
          -- pass stack -->
                           FUNC "fun
                           POP "a"
                           POP "b"
                           ...
```

Now all your arguments are set up in memory to use within the function. It is important to note that when calling functions rather than just insert the function code into the current executing code I actually cause a new instance of an executor to run. This was a design choice as I felt that this approach gave me more control as to how I wanted to implement recursion and how I wanted functions to behave in general. I also felt it was a cleaner approach that more closely mirrored how I wanted functions to be treated.

Now to deal with functions that return values this is very similar to how we handled arguments just the other way round. If we look at an assignment;

```
j = 13;  --> PUSH (Integer 13)
             POP "j"
```

You can see that when dealing with an assignment you are popping the top of the stack. Therefore if our assignment looked like this;

```
j = gun();
```

We must ensure that the value our function $gun()$ returns is left on top of the stack to be popped of our assignment. The instruction $RSTOP$ is used to signify a return value and will break out of $funExec'$. Therefore if we combine both our previous examples we can show how a function which takes in arguments will be called and how the stack is updated and passed around.

Lets look at the function;

```
func fun(a int, b int){
        Return a + b;
};
```

Then how it would behave if we use it as part of assignment;

```
j = fun(5, 8);
```

Looking at figure 5.3 we can see how this simple command is dealt with by the executor and the function executor functions.

```
                              ]
  -main executor;              -function executor;
  ...
  PUSH (Integer 30)
  PUSH (Integer 12)
  CALL "fun"
              -- pass stack -->
                               FUNC "fun
                               POP "a"
                               POP "b"
                               PUSHV "a"
                               PUSHCV "b"
                               DO ADD
                               RSTOP
              <-- pass stack --
 POP "j"
```

Figure 5.3: Example of how functions are called, including passing arguments and returning a value

Handling Recursion

5.2.6 Implementing Concurrency

How Goal Handles Concurrency

How my Virtual Machine Handles Concurrency

Chapter 6

Testing

Bibliography

[Aho et al.(2007)Aho, Lam, Sethi, and Ullman] AV. Aho, MS. Lam, R. Sethi, and JD. Ullman. *Compilers; Principles, Techniques & Tools*. Pearson, 2nd edition, 2007.

[Chattapoadhyay(2005)] Santanu Chattapoadhyay. *Compiler Design*. Prentice-Hall of India, 1st edition, 2005.

[Cooper and Torczan(2012)] Kieth D. Cooper and Linda Torczan. *Engineering a Compiler*. Morgan Kaufmann, 2nd edition, 2012.

[Jackson(2015)] Luke Jackson. Goal programming language documentation, 2015.

[Terry(1997)] P. D. Terry. *Compilers and Compiler Generators; An Introduction With C++*. Thompson Computer Press, 1st edition, 1997.