

## 1 General

The VM features a very basic assembler capable of little more than adress resolution. But yet it gives us an ability to create some neat little programs.

## 2 Syntax

The syntax for the assembly code is pretty straight forward. Each declaration is written on a single line. There are a few reserved identifiers:

Identifier	Name	Description
<code>%&lt;text&gt;</code>	Comment	Will be ignored by the assembler
<code>#&lt;name&gt;</code>	Label	Declares a label called <code>&lt;name&gt;</code>
<code>@&lt;name&gt;</code>	Value	Declares a value called <code>&lt;name&gt;</code>
<code>:&lt;data&gt;</code>	Raw input	Returns <code>&lt;data&gt;</code> as is
<code>\$&lt;a&gt;</code>	Address	Dereferences <code>a</code>
<code>x</code>	<code>x</code>	The <b>X</b> register
<code>y</code>	<code>y</code>	The <b>Y</b> register
<code>s</code>	<code>s</code>	The stack <b>S</b>
<code>q1</code>	IRQ1	The <b>Q<sub>1</sub></b> register
<code>q2</code>	IRQ2	The <b>Q<sub>2</sub></b> register

The names given to *labels* and *values* can contain any characters except for whitespace ones.

Operations are declared in a straightforward approach as:

`<opcode> <arg1> <arg2>`

Wich arguments are allowed are dependent upon the opcode.

## 3 Inner Workings

The assembler works in a fairly straight forward way. The first step in the process of assembling is the lexiographical analysis in wich the lines in the text file gets tokenized. In this tage an “intermediate structure”<sup>1</sup> gets constructed. This is an object wich contains all of the labels, values and alist of the tokenized lines. The list of the tokens contains tupels of (`label`,`offsett`,`token`) wherer the lable is the last initialized label and the offset is how many addresses away from that line the current token is. Allof the labels and values will not be assigned an adress in this phase. It is in theis phase where the opcodes and their arguments gets converted in to there corresponding numerical instruction code. It is also during this phase in wich the syntax gets checked. If a syntax error is encountered the assembler will stop emideatly. When the lexiographical analysis has been completed a check for duplicate pointer declaration is performed.

---

<sup>1</sup>The use of the word structure here is a bit ambigious since it actually is a structure in sml. But in this text it will reffere to an abstract structure of data.

The next phase in the assembly is the address resolution phase. This is done in two phases, in the first one the labels gets resolved and in the second the values gets resolved. It begins by first resolving the labels. This is done by first giving the assembler a base address wich is the adress of the first label. As of now the first non comment line in the input file has to be a label since every line has to have a label assigned to it. Then the address resolution function continues down the intermediate sturcture and remebering wich line it is at and what its last read label was. When it runns into a new label token it will sett the new label to its current address and then continue on untill it has gone trough the entire intermediate structure. After the labels have been resolved the assembler starts to resolve the values. This is done in a very straightforward way. The assembler just looks at the last address of the last entry in the output of the frstst pass and looks at the last address, adds one to it and the just places the all the values in after that address in the same order as they appeared in the file. After all the address has been resolved the assembler runns trough the list of tokens and replaces every pointer token with its correct adress.

After this is completed the assembler finalizes the code by converting everything into a list of integers wich then gets outputed to a file.

## 4 Usage

### 4.1 General

One noteworthy thing to point out is the limitations on the arguments. Due to limitations in the VM only one “none registry” argument can be used for any operation. A “non registry” argument is one wich is either a number or a pointer. Derefferencing a pointer is a registry operations so they are valid. Below follows some examples:

```
#label
@value
MOV label value      This is not accepted
MOV $label $value    This is perfectly fine
MOV 10 value         This is invalid
MOV 10 $value        But this is
ADD 1 10             This is invalid
ADD $1 $10           This is valid
ADD 1 $1             So is this
```

### 4.2 Registers

The usage of the registers is pretty straight forward. One has to remeber that q1 and q2 are write only registers and that s cant be used for addressing so \$s is not allowed and will generate a **syntax** error. It is also good to keep in mind that all operations reading from the stack will consume what is on top of the stack.

### 4.3 Pointers

Using pointers is fairly straight forward. Allthough one has to keep in mind how the addresses are resolved. All pointers will be resolved after the tokenazation fo the code. First the *labels* will be resolved and then the *values*. This means that the first *value* will lie after the last line of code. Since the address of a *label* depends on where in the code their addresses are easy to reason about. However for *values* things are bit different. Since vlues will be given addresses wich are “independent” of where in the code they appear it is hard to reason about the address of a *value*. Although the *value* pointers are resolved in order the first *value* declared whill lie intermediatley after the last line of code and the last *value* declared will lie “at the end” of the memmory used by the program. This can be exploited to use rellative addressing. Allthoug great care has to be taken.

Its important to remember that all pointers are reffered to troughout the entire program therefor it’s not allowed to define two pointers with the same name. If this where to be allowed it would generate unpredictable behaviour so instead the assembler will return a **assembler** error.

*labels* and *values* are interchangeable. Since opcodes takes pointers as arguments and has no idea weather or not they are *labels* or *values*. From this the need for caution arises. Since one can use *value* pointers as arguments to jump operation like this:

```
@bad_idea
ADD x y
MOV s x
MUL x y
JMP bad_idea
```

Since it is not known what where **bad\_idea** points jumping to it is suicidal.

Since pointers are just numbers under the hood one needs to take into account weather or not one uses them for their adress orr for their *values*. Here is some examples

```
@pointer
% This stores x in pointer
MOV x value
% This stroes x in the address wich is
% stored at pointer
MOV x $value
% This adds one to the value stored at
% pointer
ADD $pointer 1
% This adds one to the address of pointer
ADD pointer 1
```

Pointers are imutable and once they has been declared they can not be changed. One has to do some tricking to achive relative addressing using *labels* or *values*.

#### 4.3.1 Labels

*Labels* are declared using the `#` identifier. *Labels* are resolved first and their addresses correspond to location in the code where they are written. For example:

```
MOV x y
#loop
INC x
MOV x s
JMP loop
```

In this code `loop` points to the address where `INC x` is stored. In the tokenization of the assembly code the lines where a pointer is defined will be ignored and the address where the next instruction or raw entry occurs. This can lead to that poorly written code becomes ambiguous. For example:

```
MOV x y
#loop
#silly
INC y
```

Here `loop` and `silly` will both point to the same address which is `silly`.

Because tokenization of the code happens before the address resolving a *label* will be “in scope” throughout the entire code. So this code is perfectly valid:

```
MOV x y
JMP ahead
INC x
ADD x y
#ahead
ADD s x
```

The `JMP ahead` will jump to `ADD s x` even though the `ahead` flag is defined after the jump. This was not a conscious design choice but it is actually quite useful since one can define subroutines anywhere in the code which can be accessed from anywhere in the code.

One possible pitfall arises due to the fact that the assembler does not know the difference between a *label* and a *value* after their addresses have been resolved. So this code is valid assembly code:

```
#loop
ADD s x
MOV s $loop
JMP loop
```

Although what this will do is that it will change what is at the address of `loop`. But there `ADD s x` lies! This is what is known as self-modifying code and it's the spawn of Satan and should be avoided like one avoids Miami Beach during

spring break. Although in some cases the interchangeability of *value* and *label* can be very useful if one wants to have “arrays” in one’s code. This is easily achieved like this:

```
#array
:0
:1
:2
:3
```

Here `array` can be used as a pointer to the array. One can then manipulate the array through using relative addressing of `array` like this:

```
ADD 2 array
MOV s y
MOV 5 $y
#array
:0
:1
:2
:3
```

This code would change the 2 into a 5. But great care needs to be taken since one could easily end up outside of the “array” and corrupt the program.

#### 4.3.2 Values

*Values* are far more straightforward than *label*. One only has to take into account that what address a *value* is given is somewhat independent of where in the code it gets defined.