

# The SMite Virtual Machine

Reuben Thomas

7th April 2019

## Typographical conventions

Instructions and registers are shown in Typewriter font; interface calls are shown in **Bold** type.

Addresses are given in bytes and refer to the VM address space except where stated. Addresses are written in hexadecimal; hex numbers are prefixed with “0x”.

## 1 Introduction

SMite is a simple virtual machine for study and experiment. It is a stack machine, based on the more complex register machine [3]. This paper gives a full description of SMite.

SMite is conceptually (and usually in fact) a library, embedded in other programs; it supports a simple object module format.

## 2 Architecture

SMite’s address unit is the byte, which is eight bits. Most of the quantities on which SMite operates are fixed-size words, which are stored in memory in either big- or little-endian order. These parameters enable efficient implementation on the vast majority of machine architectures.

### 2.1 Registers

The registers are word quantities; they are listed, with their functions, in table 1. Registers without fixed values are initialised to 0.

### 2.2 Memory

SMite’s memory is a contiguous sequence of bytes with addresses starting at 0. The address of a word is that of the byte in it with the lowest address.

### 2.3 Stack

The stack is a LIFO stack of words used for passing values to instructions and routines and for holding subroutine return addresses. To **push** a word on to the stack means to add a new word to the top of the stack, increasing the stack

Register	Function
PC	The Program Counter. Points to the next word from which I may be loaded.
I	The Instruction register. Contains instructions to be executed.
BAD	The invalid stack position, or invalid or unaligned memory address, that last caused an error.
STACK_DEPTH	The number of words on the stack.
ENDISM	The endianness of SMite: 0 = Little-endian, 1 = Big-endian. Fixed for a particular instance of SMite.
WORD_BYTES	The number of bytes in a word. Must be in the range 2 to 32 inclusive, and a power of 2. Fixed for a particular instance of SMite.

Table 1: Registers

depth by 1; to **pop** a word means to reduce the stack depth by 1. Instructions that change the number of words on the stack implicitly pop their arguments and push their results.

## 2.4 Execution

Execution proceeds as follows:

```

begin
    execute the instruction in the 8 least-significant bits of I
    shift I logically 8 bits to the right
repeat

```

In the latter case, the contents of the execution cycle is executed once, and control returns to the calling program.

## 2.5 Errors and termination

When SMite encounters certain abnormal situations, such as an attempt to access an invalid address, or divide by zero, an **error** is **raised**, and execution terminates; an **error code** is returned to the caller. The instruction has no effect, and PC is not advanced to the next instruction. If the error is a stack or memory access error, BAD is set to the stack position or address that caused the error.

Execution can also be terminated explicitly by performing a HALT instruction (see section 3.10).

Error codes are unsigned numbers. 0 to 128 are reserved for SMite's own error codes; the meanings of those that may be raised by SMite are shown in table 2.

## 3 Instruction set

The instruction set is listed below, with the instructions grouped according to function. The instructions are given in the following format:

Code	Meaning
0	<b>single_step()</b> has terminated without error.
1	Invalid opcode (see section 3.12).
2	Stack overflow. <b>BAD</b> is set to the extra number of words of stack space required.
3	Invalid stack read. <b>BAD</b> is set to the invalid stack position.
4	Invalid stack write. <b>BAD</b> is set to the invalid stack position.
5	Invalid memory read. <b>BAD</b> is set to the invalid address.
6	Invalid memory write. <b>BAD</b> is set to the invalid address.
7	Address alignment error: raised when an instruction is given a valid address, but insufficiently aligned.
8	Invalid size (greater than $\log_2 \text{WORD\_BYTES}$ ).
9	Division by zero attempted (see section 3.7).
128	A <b>HALT</b> instruction was executed.

Table 2: Errors raised by SMite

NAME ( *before* - *after* )

Description.

The first line consists of the name of the instruction. On the right is the stack effect, which shows the effect of the instruction on the stack. Underneath is the description.

**Stack effects** are written

( *before* - *after* )

where *before* and *after* are stack pictures showing the items on top of a stack before and after the instruction is executed. An instruction only affects the items shown in its stack effects. The brackets and dashes serve merely to delimit the stack effect and to separate *before* from *after*. **Stack pictures** are a representation of the top-most items on the stack, and are written

$$i_1 \ i_2 \dots i_{n-1} \ i_n$$

where the  $i_k$  are stack items, each of which occupies a whole number of words,<sup>1</sup> with  $i_n$  being on top of the stack. The symbols denoting different types of stack item are shown in table 3.

Types are only used to indicate how instructions treat their arguments and results; SMite does not distinguish between stack items of different types. In stack pictures the most general argument types with which each instruction can be supplied are given; subtypes may be substituted. Using the phrase " $i \Rightarrow j$ " to denote " $i$  is a subtype of  $j$ ", table 4 shows the subtype relationships. The subtype relation is transitive.

Numbers are represented in twos complement form. *addr* consists of all valid virtual machine addresses.

Each type may be suffixed by a number in stack pictures; if the same combination of type and suffix appears more than once in a stack effect, it refers each time to the identical stack item.

Ellipsis is used for indeterminate numbers of specified types of item.

<sup>1</sup>In this specification, each stack item occupies *precisely* one word.

Symbol	Data type
<i>flag</i>	a Boolean flag, 0 for false or 1 for true; other non-zero values are also treated as true
<i>size</i>	an integer in the range 0 to $\log_2 \text{WORD\_BYTES}$ inclusive.
<i>n</i>	signed number
<i>u</i>	unsigned number
<i>n u</i>	number (signed or unsigned)
<i>x</i>	unspecified word
<i>addr</i>	address
<i>a-addr</i>	word-aligned address

Table 3: Types used in stack effects

$u \Rightarrow x$
$n \Rightarrow x$
$flag \Rightarrow u$
$size \Rightarrow u$
$a-addr \Rightarrow addr \Rightarrow u$

Table 4: The subtype relation

### 3.1 Instruction fetch

NEXT ( - )

If I is not 0, raise error 1. If PC is not word-aligned, raise error 7, setting BAD to the contents of PC. Load the word pointed to by PC into I then add WORD\_BYTES to PC.

### 3.2 Control

These instructions implement unconditional and conditional branches, and subroutine call and return (subroutine return is BRANCH):

BRANCH ( *a-addr* - )

Set PC to *a-addr*. Perform the action of NEXT.

BRANCHZ ( *flag* *a-addr* - )

If *flag* is false then set PC to *a-addr* and perform the action of NEXT.

CALL ( *a-addr*<sub>1</sub> - *a-addr*<sub>2</sub> )

Exchange PC with the top stack value. Perform the action of NEXT.

### 3.3 Stack manipulation

These instructions manage the stack:

POP ( *x* - )

Remove *x* from the stack.

DUP ( *x*<sub>u</sub> . . . *x*<sub>0</sub> *u* - *x*<sub>u</sub> . . . *x*<sub>0</sub> *x*<sub>u</sub> )

Remove *u*. Copy *x*<sub>u</sub> to the top of the stack.

SWAP  $(x_{u+1} \dots x_0 \ u - x_0 \ x_u \dots x_1 \ x_{u+1})$

Exchange the top stack word with the  $u+1$ th.

GET\_STACK\_DEPTH  $(-u)$

$u$  is the value of STACK\_DEPTH, the number of words on the stack.

SET\_STACK\_DEPTH  $(u -)$

Set STACK\_DEPTH to  $u$ .

### 3.4 Literals

LIT  $(-n)$

The word pointed to by PC is pushed on to the stack, and PC is incremented to point to the following word.

LIT\_PC\_REL  $(-n)$

Like LIT, except that the initial value of PC is added to the value pushed on to the stack.

LIT\_0  $(-0)$

Push 0 on to the stack.

LIT\_1  $(-1)$

Push 1 on to the stack.

### 3.5 Logic and shifts

Logic functions:

NOT  $(x_1 - x_2)$

Invert all bits of  $x_1$ , giving its logical inverse  $x_2$ .

AND  $(x_1 \ x_2 - x_3)$

$x_3$  is the bit-by-bit logical “and” of  $x_1$  with  $x_2$ .

OR  $(x_1 \ x_2 - x_3)$

$x_3$  is the bit-by-bit inclusive-or of  $x_1$  with  $x_2$ .

XOR  $(x_1 \ x_2 - x_3)$

$x_3$  is the bit-by-bit exclusive-or of  $x_1$  with  $x_2$ .

Shifts:

LSHIFT  $(x_1 \ u - x_2)$

Perform a logical left shift of  $u$  bit-places on  $x_1$ , giving  $x_2$ . Put zero into the least significant bits vacated by the shift. If  $u$  is greater than or equal to the number of bits in a word,  $x_2$  is zero.

RSHIFT  $(x_1 \ u - x_2)$

Perform a logical right shift of  $u$  bit-places on  $x_1$ , giving  $x_2$ . Put zero into the most significant bits vacated by the shift. If  $u$  is greater than or equal to the number of bits in a word,  $x_2$  is zero.

ARSHIFT (  $x_1$   $u$  -  $x_2$  )

Perform an arithmetic right shift of  $u$  bit-places on  $x_1$ , giving  $x_2$ . Copy the original most-significant bits into the most significant bits vacated by the shift. If  $u$  is greater than or equal to the number of bits in a word, all the bits of  $x_2$  are the same as the original most-significant bit.

SIGN\_EXTEND (  $u$  *size* -  $n$  )

Sign extend the  $2^{size}$ -byte quantity  $u$  to  $n$ .

### 3.6 Comparison

These words compare two numbers (or, for equality tests, any two words) on the stack, returning a flag:

EQ (  $x_1$   $x_2$  - *flag* )

*flag* is true if and only if  $x_1$  is bit-for-bit the same as  $x_2$ .

LT (  $n_1$   $n_2$  - *flag* )

*flag* is true if and only if  $n_1$  is less than  $n_2$ .

ULT (  $u_1$   $u_2$  - *flag* )

*flag* is true if and only if  $u_1$  is less than  $u_2$ .

### 3.7 Arithmetic

These instructions consist of monadic and dyadic operators. All calculations are made without bounds or overflow checking, except as detailed for certain instructions.

Negation and addition:

NEGATE (  $n_1$  -  $n_2$  )

Negate  $n_1$ , giving its arithmetic inverse  $n_2$ .

ADD (  $n_1 | u_1$   $n_2 | u_2$  -  $n_3 | u_3$  )

Add  $n_2 | u_2$  to  $n_1 | u_1$ , giving the sum  $n_3 | u_3$ .

Multiplication and division (note that all division instructions raise error -4 if division by zero is attempted):

MUL (  $n_1 | u_1$   $n_2 | u_2$  -  $n_3 | u_3$  )

Multiply  $n_1 | u_1$  by  $n_2 | u_2$  giving the product  $n_3 | u_3$ .

DIVMOD (  $n_1$   $n_2$  -  $n_3$   $n_4$  )

Divide  $n_1$  by  $n_2$  using symmetric division, giving the single-word quotient  $n_3$  and the single-word remainder  $n_4$ . The quotient is rounded towards zero.

UDIVMOD (  $u_1$   $u_2$  -  $u_3$   $u_4$  )

Divide  $u_1$  by  $u_2$ , giving the single-word quotient  $u_3$  and the single-word remainder  $u_4$ .

### 3.8 Memory

These instructions fetch and store quantities to and from memory. If a given address is incorrectly aligned, raise error 7, setting BAD to the address.

LOAD ( *addr size - x* )  
Load the  $2^{size}$ -byte quantity *x* stored at *addr*, which must be a multiple of  $2^{size}$ . Any unused high-order bits are set to zero.

STORE ( *x addr size -* )  
Store the  $2^{size}$  least-significant bytes of *x* at *addr*, which must be a multiple of  $2^{size}$ .

### 3.9 External access

This instruction is a general-purpose escape. In contrast to extra instructions, which are intended to be frequently used and potentially optimized, EXT's behaviour is not defined; in particular, it has an arbitrary stack effect.

EXT ( )  
Perform arbitrary actions.

### 3.10 Termination

This instruction terminates execution (see section 2.5):

HALT ( - )  
Terminate execution with error code 128.

### 3.11 Instruction encoding

Instructions are encoded as 8-bit opcodes; opcodes are packed into words, which are executed starting at the least-significant bits.

### 3.12 Instruction opcodes

Table 5 lists the instruction opcodes in numerical order. Other instruction opcodes are undefined.

## 4 External interface

- Implementations should provide an **API** to create and run virtual machine instances, and provide access to its registers, stack and memory.
- Implementations can add **extra instructions** to provide extra computational primitives.
- The EXT instruction can offer access to system facilities, previously written code, native libraries and so on.
- The **object module** format allows compiled code to be saved, reloaded and shared between systems.

Opcode	Instruction	Opcode	Instruction
0x00	NEXT	0x10	EQ
0x01	BRANCH	0x11	LT
0x02	BRANCHZ	0x12	ULT
0x03	CALL	0x13	NEGATE
0x04	POP	0x14	ADD
0x05	DUP	0x15	MUL
0x06	SWAP	0x16	DIVMOD
0x07	undefined	0x17	UDIVMOD
0x08	NOT	0x18	LOAD
0x09	AND	0x19	STORE
0x0a	OR	0x1a	LIT
0x0b	XOR	0x1b	LIT_PC_REL
0x0c	LSHIFT	0x1c	LIT_0
0x0d	RSHIFT	0x1d	LIT_1
0x0e	ARSHIFT	0x1e	EXT
0x0f	SIGN_EXTEND	0x1f	HALT

Table 5: Instruction opcodes

## 4.1 Object module format

The object module starts with the ASCII codes of the letters “smite” followed by ASCII NUL (0x00), then the one-byte values of the `ENDISM` and `WORD_BYTES` registers of the system which saved the module, then a word (of the given endianness and size) containing the number of bytes the code occupies. Then follows the code.

Object modules have a simple structure, as they are only intended for loading an initial memory image into SMite.

## Acknowledgements

Martin Richards’s demonstration of his BCPL-oriented Cintcode virtual machine [1] convinced me it was going to be fun working on virtual machines. He also supervised my BA dissertation project, Beetle [2], on which SMite is based.

## References

- [1] Martin Richards. Cintcode distribution, 2000. <https://www.cl.cam.ac.uk/~mr/BCPL.html>.
- [2] Reuben Thomas. Beetle and pForth: a Forth virtual machine and compiler. BA dissertation, University of Cambridge, 1995. <https://rrt.sc3d.org/>.
- [3] Reuben Thomas. *Mite: a basis for ubiquitous virtual machines*. PhD thesis, University of Cambridge Computer Laboratory, November 2000. <https://rrt.sc3d.org/>.