Appendix E
Definition of the CPRL Virtual Machine

E.1 Specification.

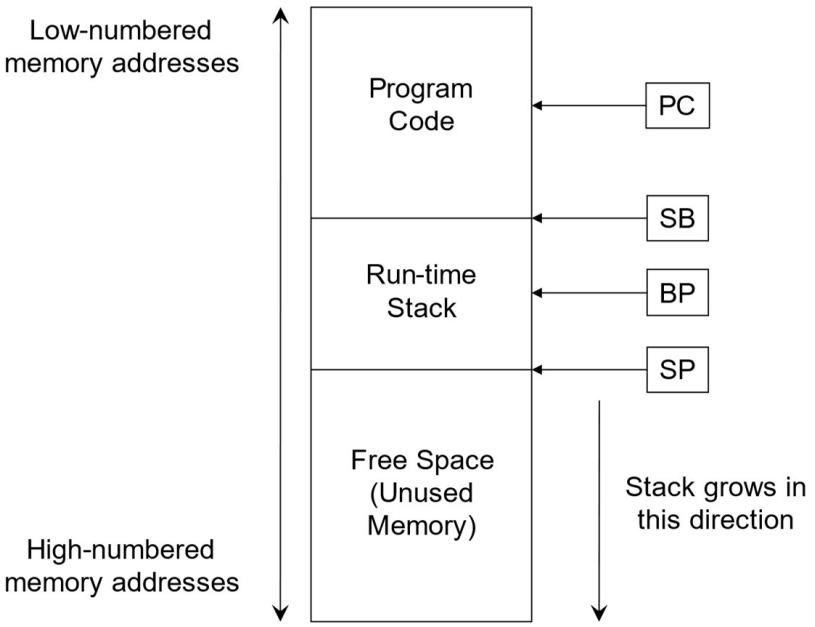
CVM (CPRL Virtual Machine) is a hypothetical computer designed to simplify the code generation phase of a compiler for CPRL (the Compiler PRoject Language. CVM has a stack architecture; i.e., most instructions either expect operands on the stack, place results on the stack, or both. Memory is organized into 8-bit bytes, and each byte is directly addressable. A word is a logical grouping of 4 consecutive bytes in memory. The address of a word is the address of its first (low) byte. `Boolean` values are represented in a single byte, character values use 2 bytes (Unicode Basic Multilingual Plane or Plane 0, code points from `U+0000` to `U+FFFF`), and integer values use a word (four bytes).

CVM has four 32-bit internal registers that are usually manipulated indirectly as a result of program execution. There are no general-purpose registers for computation. The names and functions of the internal registers are as follows.

- `PC` (program counter); a.k.a. instruction pointer: holds the address of the next instruction to be executed.
- `SP` (stack pointer): holds the address of the top of the stack. The stack grows from low-numbered memory addresses to high-numbered memory addresses. When the stack is empty, `SP` has a value of the address immediately before the first free byte in memory.
- `SB` (stack base): holds the address of the bottom of the stack. When a program is loaded, `SB` is initialized to the address of the first free byte in memory.
- `BP` (base pointer): holds the base address of the current activation record (a.k.a., frame); i.e., the base address for the subprogram currently being executed.

Each CVM instruction operation code (opcode) occupies one byte of memory. Some instructions take an immediate operand, which is always located immediately following the instruction in memory. Depending on the opcode, an immediate operand can be a single byte, two bytes (e.g., for a char), four bytes (e.g. for an integer or a memory address), or multiple bytes (e.g., for a string literal). The complete instruction set for CVM is given in the next section. Most instructions get their operands from the run-time stack. In general, the operands are removed from the stack whenever the instruction is executed, and any results are left on the top of the stack. With respect to boolean values, zero means false and any nonzero value is interpreted as true.

The following diagram illustrates a program loaded into memory.



Variable Addressing

Each time that a subprogram is called, CVM saves the current value of `BP` and sets `BP` to point to the new activation record (a.k.a., frame). When the subprogram returns, CVM restores `BP` back to the saved value.

A variable has an absolute address in memory (the address of the first byte), but variables are more commonly addressed relative to a register. A local variable is addressed relative to register `BP`, and a global variable is addressed relative to register `SB`.

Various load and store operations move data between memory and the run-time stack.

E.2 Implementation

CVM is implemented by three classes in package `edu.citadel.cvm`.

Class `Constants` defines the number of bytes for primitive types plus the number of bytes for the context part of an activation record.

```
public class Constants
  {
    public static final int BYTES_PER_OPCODE  = 1;
    public static final int BYTES_PER_INTEGER = 4;
    public static final int BYTES_PER_ADDRESS = 4;
    public static final int BYTES_PER_CHAR    = 2;
    public static final int BYTES_PER_BOOLEAN = 1;
    public static final int BYTES_PER_CONTEXT = 2*BYTES_PER_ADDRESS;
  }
```

Class `Opcode` is an enum class that defines the name and numeric values for each CVM opcode. In addition, this class defines several helper functions used by the CVM, the assembler, and disassembler, including static method `toOpcode(byte b)` that returns the opcode for the specified byte value.

```
public enum Opcode
  {
    // halt opcode
    HALT(0),

    // load opcodes (move data from memory to top of stack)
    LOAD(9),
    LOADB(10),
    LOADW(12),
    LOADSTR(13),
    LDCB(14),
    LDCCH(15),
    LDCINT(16),
    ...

    // arithmetic opcodes
    ADD(70),
    SUB(71),
    MUL(72),
    DIV(73),
    ...

    // program/procedure opcodes
    PROGRAM(90),
    PROC(91),
    CALL(92),
    RET(93),
    ...

    private final byte value;

    /**
     * Construct an opcode with its machine instruction value.
     */
    private Opcode(int value)
      {
        this.value = (byte) value;
      }

    ...
  }
```

Class `CVM` is the primary component of the implementation for the virtual machine. In addition to several helper methods, every opcode is implemented by a method. CVM method `run()` provides the basic control logic for the virtual machine using a large "`switch`" statement to dispatch opcodes to their corresponding method calls.

```
public void run()
  {
    running = true;
    pc = 0;

    while (running)
      {
        switch (Opcode.toOpcode(fetchByte()))
          {
            case ADD     -> add();
            case ALLOC   -> allocate();
            case BR      -> branch();
            case BE      -> branchEqual();
            case BNE     -> branchNotEqual();
            case BG      -> branchGreater();
            case BGE     -> branchGreaterOrEqual();
            case BL      -> branchLess();
            case BLE     -> branchLessOrEqual();
            case BZ      -> branchZero();
            case BNZ     -> branchNonZero();
            case CALL    -> call();
            case DEC     -> decrement();
            case DIV     -> divide();
            case GETCH   -> getCh();
            case GETINT  -> getInt();
            case GETSTR  -> getString();
            case HALT    -> halt();
            case INC     -> increment();
            ...
            case SHL     -> shiftLeft();
            case SHR     -> shiftRight();
            case STORE   -> store();
            case STOREB  -> storeByte();
            case STORE2B -> store2Bytes();
            case STOREW  -> storeWord();
            case SUB     -> subtract();
            default      -> error("invalid machine instruction");
          }
      }
  }
```

E.3 CVM Instruction Set Architecture

Mnemonic	Short Description	Stack before after	Definition
Arithmetic Opcodes			
ADD	Add: Pop two integers from the stack and push their sum back onto the stack.	n1 n2 n1 + n2	n2 ← popInt() n1 ← popInt() pushInt(n1 + n2)
SUB	Subtract: Pop two integers from the stack and push their difference back onto the stack.	n1 n2 n1 - n2	n2 ← popInt() n1 ← popInt() pushInt(n1 - n2)
MUL	Multiply: Pop two integers from the stack and push their product back onto the stack.	n1 n2 n1*n2	n2 ← popInt() n1 ← popInt() pushInt(n1*n2)
DIV	Divide: Pop two integers from the stack and push their quotient back onto the stack.	n1 n2 n1/n2	n2 ← popInt() n1 ← popInt() pushInt(n1/n2)
MOD	Modulo: Pop two integers from the stack, divide them, and push the remainder back onto the stack.	n1 n2 n1 % n2	n2 ← popInt() n1 ← popInt() pushInt(n1 % n2)
NEG	Negate: Pop an integer from the stack, negate it, and push the result back onto the stack.	n -n	n ← popInt() pushInt(-n)
INC	Increment: Pop an integer from the stack, add 1, and push the result back onto the stack.	n n + 1	n ← popInt() pushInt(n + 1)
DEC	Decrement: Pop an integer from the stack, subtract 1, and push the result back onto the stack.	n n - 1	n ← popInt() pushInt(n - 1)

Logical Opcodes			

`NOT`	Logical Not: Pop a byte from the stack and push its logical negation back onto the stack.	`b`
		`!b`
			`b ← popByte()`
			`if b = 0`
			`  pushByte(1)`
			`else`
			`  pushByte(0)`

Shift Opcodes			
`SHL b`	Shift Left: Pop an integer from the stack, shift the bits left by the specified amount using zero fill, and push the result back onto the stack. Note: Only the right most five bits of the argument are used for the shift.	`n` `n << b`	`n ← popInt()` `pushInt(n << b)`
`SHR b`	Shift Right: Pop an integer from the stack, shift it right by the specified amount using sign extend, and push the result back onto the stack. Note: Only the right most five bits of the argument are used for the shift.	`n` `n >> b`	`n ← popInt()` `pushInt(n >> b)`
Branch Opcodes			
`BR displ`	Branch: Branch unconditionally according to displacement argument (may be positive or negative).		`pc ← pc + displ`

`BE displ`	Branch Equal: Pop two integers from the stack and compare them. If they are equal, then branch according to displacement argument (may be positive or negative); otherwise continue with the next instruction.	`n1`
		`n2`
			`n2 ← popInt()`
			`n1 ← popInt()`
			`if n1 == n2`
			`  pc ← pc + displ`

`BNE displ`	Branch Not Equal: Pop two integers from the stack and compare them. If they are not equal, then branch according to displacement argument (may be positive or negative); otherwise continue with the next instruction.	`n1`
		`n2`
			`n2 ← popInt()`
			`n1 ← popInt()`
			`if n1 != n2`
			`  pc ← pc + displ`

`BG displ`	Branch Greater: Pop two integers from the stack and compare them. If the second integer is greater than the first, then branch according to displacement argument (may be positive or negative); otherwise continue with the next instruction.	`n1`
		`n2`
			`n2 ← popInt()`
			`n1 ← popInt()`
			`if n1 > n2`
			`  pc ← pc + displ`

`BGE displ`	Branch Greater or Equal: Pop two integers from the stack and compare them. If the second integer is greater than or equal to the first, then branch according to displacement argument (may be positive or negative); otherwise continue with the next instruction.	`n1`
		`n2`
			`n2 ← popInt()`
			`n1 ← popInt()`
			`if n1 >= n2`
			`  pc ← pc + displ`

`BL displ`	Branch Less: Pop two integers from the stack and compare them. If the second integer is less than the first, then branch according to displacement argument (may be positive or negative); otherwise continue with the next instruction.	`n1`
		`n2`
		———
			`n2 ← popInt()`
			`n1 ← popInt()`
			`if n1 < n2`
			`  pc ← pc + displ`

`BLE displ`	Branch Less or Equal: Pop two integers from the stack and compare them. If the second integer is less than or equal to the first, then branch according to displacement argument (may be positive or negative); otherwise continue with the next instruction.	`n1`
		`n2`
		———
			`n2 ← popInt()`
			`n1 ← popInt()`
			`if n1 <= n2`
			`  pc ← pc + displ`

`BZ displ`	Branch if Zero: Pop one byte from the stack. If it is zero, then branch according to displacement argument (may be positive or negative); otherwise continue with the next instruction.	`b`
		———
			`b ← popByte()`
			`if b = 0`
			`  pc ← pc + displ`

`BNZ displ`	Branch if Nonzero: Pop one byte from the stack. If it is nonzero then branch according to displacement argument (may be positive or negative); otherwise continue with the next instruction.	`b`
		———
			`b ← popByte()`
			`if b ≠ 0`
			`  pc ← pc + displ`

Load/Store Opcodes			

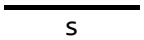
`LOAD n`	Load multiple bytes onto the stack: The number of bytes to move is part of the instruction. Pop an address from the stack and push n bytes starting at that address onto the stack.	`addr`
		———
		`b1`
		`b2`
		`...`
		`bn`
			`addr ← popInt();`
			`for i ← 0..n-1 loop`
			`  pushByte(mem[addr + i])`

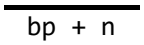
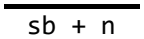
`LOADB`	Load Byte: Load (push) a single byte onto the stack. The address of the byte is obtained by popping it off the stack.	`addr`
		———
		`b`
			`addr ← popInt()`
			`b ← mem[addr]`
			`pushByte(b)`

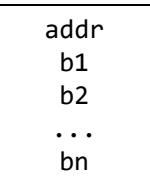
`LOAD2B`	Load Two Bytes: Load (push) two consecutive bytes onto the stack. The address of the first byte is obtained by popping it off the stack.	`addr`
		———
		`b0`
		`b1`
			`addr ← popInt()`
			`b0 ← mem[addr + 0]`
			`b1 ← mem[addr + 1]`
			`pushByte(b0)`
			`pushByte(b1)`

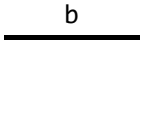
`LOADW`	Load Word: Load (push) a word (four consecutive bytes) onto the stack. The address of the word is obtained by popping it off the stack.	`addr`
		———
		`w`
			`addr ← popInt()`
			`w ← getWord(addr)`
			`pushInt(w)`

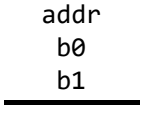
`LDCB b`	Load Constant Byte: Fetch the byte immediately following the opcode and push it onto the stack.	——— `b`	`pushByte(b)`
`LDCB0`	Load Constant Byte `0`: Optimized version of `LDCB 0`.	——— `0`	`pushByte(0)`
`LDCB1`	Load Constant Byte `1`: Optimized version of `LDCB 1`.	——— `1`	`pushByte(1)`
`LDCCH c`	Load Constant Character: Fetch the character immediately following the opcode and push it onto the stack.	——— `c`	`pushChar(c)`
`LDCINT n`	Load Constant Integer: Fetch the integer immediately following the opcode and push it onto the stack.	——— `n`	`pushInt(n)`
`LDCINT0`	Load Constant Integer `0`: Optimized version of `LDCINT 0`.	——— `0`	`pushInt(0)`
`LDCINT1`	Load Constant Integer `1`: Optimized version of `LDCINT 1`.	——— `1`	`pushInt(1)`

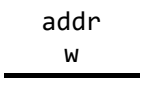
`LDCSTR s`	Load Constant String: The string (length plus characters) immediately follows the opcode. Push the string (length and characters) onto the stack.		`n ← fetchInt()`
			`pushInt(n)`
			`for i ← 0..n-1 loop`
			`  c ← fetchChar()`
			`  pushChar(c)`

`LDLADDR n`	Load Local Address: Compute the absolute address of a local variable from its relative address `n` and push the absolute address onto the stack.		`pushInt(bp + n)`
`LDGADDR n`	Load Global Address: Compute the absolute address of a global (program level) variable from its relative address `n` and push the absolute address onto the stack.		`pushInt(sb + n)`

`STORE n`	Store n bytes: Remove `n` bytes from the stack followed by an absolute address and copy the `n` bytes to the location starting at the absolute address.		`for i ← n-1..0 loop`
			`  data[i] ← popByte()`
			`addr ← popInt()`
			`for i ← 0..n-1 loop`
			`  mem[addr + i] ← data[i]`

`STOREB`	Store Byte: Store a single byte at a specified memory location. The byte to be stored and the address where it is to be stored are popped from the stack.		`b ← popByte()`
			`addr ← popInt()`
			`memory[addr] ← b`

`STORE2B`	Store Two Bytes: Store two bytes at a specified memory location. The bytes to be stored and the address where they are to be stored are popped from the stack.		`b1 ← popByte()`
			`b0 ← popByte()`
			`addr ← popInt()`
			`mem[addr + 0] ← b0`
			`mem[addr + 1] ← b1`

`STOREW`	Store Word: Store a word (4 bytes) at a specified memory location. The word to be stored and the address where it is to be stored are popped from the stack.		`w ← popWord()`
			`addr ← popInt()`
			`putWord(w, addr)`

`ALLOC n`	Allocate: Allocate space on the stack for future use.		`sp ← sp + n`
Program/Procedure Opcodes			
`PROGRAM n`	Program: Initialize base pointer and allocate space on the stack for the program's local variables.		`bp ← sb` `sp ← bp + n - 1`
`PROC n`	Procedure: Allocate space on the stack for a subprogram's local variables.		`sp ← sp + n`

`CALL disp`	Call: Call a subprogram, pushing current values for `BP` and `PC` onto the stack.	bp
		p
			`pushInt(bp)`
			`pushInt(pc)`
			`bp ← sp - 7`
			`pc ← pc + disp`

`RET n`	Return: Return from a subprogram, restoring the old value for BP plus space on stack previously allocated for the subprogram's local variables and parameters.		`bpSave ← bp`
			`sp ← bpSave - n - 1`
			`bp ← getInt(bpSave)`
			`pc ← getInt(bpSave + 4)`

`RET0`	Optimized version of `RET 0`.		`bpSave ← bp`
			`sp ← bpSave - 1`
			`bp ← getInt(bpSave)`
			`pc ← getInt(bpSave + 4)`

`RET4`	Optimized version of `RET 4`.		`bpSave ← bp`
			`sp ← bpSave - 5`
			`bp ← getInt(bpSave)`
			`pc ← getInt(bpSave + 4)`

`HALT`	Halt: Stop the virtual machine.		`halt`
I/O Opcodes			

`GETINT`	Get Integer: Read digits from standard input, convert them to an integer, and store the integer at the address on top of stack.	addr	`addr ← popInt()`
			`n ← readInt()`
			`putInt(n, addr)`

`GETCH`	Get Character: Read character from standard input and store it at the address on top of stack.	`addr`	`addr ← popInt()`
			`c ← readChar()`
			`putChar(c, addr)`

`GETSTR n`	Get String: Read string from standard input and store it at the address on top of stack.	`addr`	`addr ← popInt()`
			`s ← readStr()`
			`strLen = min(s.length, n)`
			`putInt(n, addr)`
			`for i ← 0..strLen-1 loop`
			`  putChar(s[i], addr)`
			`  addr ← addr + 2`

`PUTBYTE`	Put Byte: Pop byte from top of the stack and write its value to standard output.	`b`	`b ← popByte()` `writeByte(b)`
`PUTINT`	Put Integer: Pop integer from top of the stack and write its value to standard output.	`n`	`n ← popInt()` `writeInt(n)`
`PUTCH`	Put Character: Pop character from top of stack and write its value to standard output.	`c`	`c ← popChar()` `writeChar(c)`

`PUTSTR n`	Put String: Write a string of n characters to standard output. The string (length plus characters) was previously pushed onto the stack.	`s`	`nBytes ← 4 + 2*n`
			`addr ← sp – nBytes + 1`
			`strLen ← getInt(addr)`
			`for i ← 0..strLen-1 loop`
			`  writeChar(mem[2*i])`
			`sp ← sp - n`

`PUTEOL`	Put End-of-Line: Write a line terminator to standard output.		`write(EOL)`