University of Arizona, Department of Computer Science

CSc 553 — Assignment 1 — Due noon, Tue Feb 8 — 10%

Christian Collberg
January 18, 2011

# 1 Introduction

This assignment consists of three tasks:

1. Design a virtual machine code suitable for executing the language LUCA.
2. Given a `Luca` compiler front-ent, wite a back-end that generates virtual machine code.
3. Write an efficent interpreter that reads and executes this virtual machine code.

Note the following:

1. The LUCA language is defined in Section A. Of particular interest to you is Section A.5 which describes the runtime errors that your interpreter must check for.
2. You should work in teams of two students.
3. Download the compiler front-end and test-cases from `http://www.cs.arizona.edu/~collberg/Teaching/553/2011/Assignments/lucadist.zip`
4. The interpreter should be implemented using *indirect threaded code*.
5. You should write your interpreter in C or C++ using `gcc`'s *labels-as-values*.
6. The compiler should be named `lucac` and the interpreter should be named `lucax`. They should be called like this:

   ```
   > lucac x.luc -o x.vm
   > lucax x.vm
   ```

   The script `lucac` has been given to you. `lucac -h` describes the input arguments to the compiler. The main compiler class is called `LucaCompiler.java` — modify this to call your back-end.
7. The front-end is written in Java and generates a *sequence-of-expression-tree* intermediate representation. Tree node types are defined in `tree/*.java` and described in Section B.2.
8. To make this a realistic exercise, the bytecode in the `.vm`-file should be written as *integers*, not strings. I.e. the file shouldn't look like this:

   ```
   load x
   load y
   iadd
   store z
   ```

but rather something like this:

```
10 1 10 2 5 11
```

Whether your `.vm`-file is in a binary (like Java class-files) or a text format is up to you.

9. You should test the interpreter on `lectura`.

# 2 Submission and Assessment

The deadline for this assignment is noon, Tue Feb 8. It is worth 10% of your final grade.

You should submit the assignment to `d2l.arizona.edu`.

You should submit *one* file, ass1.zip, containing all the files necessary to build the compiler and interpreter. Modify the `makefile` so that the grader can build the project by simply typing `make`, and nothing else.

**Don't show your code to anyone, don't read anyone else's code, don't discuss the details of your code with anyone. If you need help with the assignment see the instructor or the TA.**

# A The Luca Language

## A.1 Luca Lexical Rules

- Luca line comments start with a `--`-sign and extend to the end of the line.
- Luca structured comments start with a `(*` and must end with `*)`. They are not allowed to be nested.
- Luca is case-sensitive.
- Strings start and end with a `"`-character and cannot contain a `"`-character. They cannot extend past the end of a line.
- Character literals start and end with a `'`-character and must contain exactly one character (not a `'`).
- Identifiers consist of letters and digits, and must start with a letter.
- Integer literals consist of a sequence of digits.
- Real literals have the syntax

$$((\texttt{digit}*\texttt{.digit}+)|(\texttt{digit}+\texttt{.digit}*))(\texttt{E}(+|-)?\texttt{digit}+)?$$

  Examples of valid floating-point numbers:

```
0.5    .5    5.    5.0    5.E-6    100.587E99
```

- Control characters other than tabs and newlines are not allowed in Luca source files.

## A.2 LUCA Syntax

program ::= '**PROGRAM**' ident ';' decl_list block '.'
block ::= '**BEGIN**' stat_seq '**END**'
decl_list ::= { declaration ';' }
declaration ::= '**CONST**' ident ':' ident '=' expression |
'**VAR**' ident ':' ident |
'**TYPE**' ident '=' '**ARRAY**' expression '**OF**' ident |
'**TYPE**' ident '=' '**RECORD**' '[' [ field_list ] ']' |
'**PROCEDURE**' ident '(' [formal_list] ')' ';' decl_list block
formal_list ::= formal_param { ';' formal_param }
field_list ::= field { ';' field }
formal_param ::= ['**VAR**'] ident ':' ident
field ::= ident ':' ident
stat_seq ::= { statement ';' }
statement ::= designator ':=' expression |
designator '(' [ actual_list ] ')' |
'**IF**' expression '**THEN**' stat_seq '**ENDIF**' |
'**IF**' expression '**THEN**' stat_seq '**ELSE**' stat_seq '**ENDIF**' |
'**WHILE**' expression '**DO**' stat_seq '**ENDDO**' |
'**REPEAT**' stat_seq '**UNTIL**' expression|
'**LOOP**' stat_seq '**ENDLOOP**' |
'**EXIT**' |
'**WRITE**' expression | '**WRITELN**' |
'**READ**' designator
actual_list ::= expression { ',' expression }
expression ::= expression bin_operator expression |
unary_operator expression |
'(' expression ')' |
integer_literal | char_literal | real_literal | string_literal | designator
designator ::= ident { designator' }
designator' ::= '[' expression ']' | '.' ident
bin_operator ::= '+' | '−' | '∗' | '/' | '%' | '**AND**' | '**OR**' | '<' | '<=' | '=' |'#' | '>=' |'>'
unary_operator ::= '−' | '**NOT**' | '**TRUNC**' | '**FLOAT**'

This grammar is highly ambiguous. Here are the relevant operator precedence rules:

| **precedence** | **operator** | **arity** | **associativity** |
|---|---|---|---|
| low | `+, -` | binary | left associative |
| | `*, /, %` | binary | left associative |
| | `AND,OR` | binary | left associative |
| | `<, <=, #, >, >=, =` | binary | left associative |
| high | `NOT, TRUNC, FLOAT,` $-_{\texttt{unary}}$ | unary | right associative |

## A.3 Static Semantics

- The LUCA language is case sensitive.

- Luca has four (incompatible) built-in types: `INTEGER`, `CHAR`, `BOOLEAN` and `REAL`. All basic types are 32 bits wide.
- The '`#`' symbol means "not equal to". `AND` and `OR` have lower precedence than the comparison operators, which in turn have lower precedence than the arithmetic operators.
- LUCA does not allow *mixed arithmetic*, i.e. there is no *implicit conversion* of integers to reals in an expression. For example, if `I` is an integer and `R` is real, then `R:=I+R` is illegal. LUCA instead supports two explicit conversion operators, `TRUNC` and `FLOAT`. `TRUNC R` returns the integer part of `R`, and `FLOAT I` returns a real number representation of `I`. Note also that `%` (remainder) is not defined on real numbers.
- These are the type rules for Luca:

| Left | Operators | Right | | Result |
|---|---|---|---|---|
| Int | '+', '−', '∗', '/', '%' | Int | ⇒ | Int |
| Real | '+', '−', '∗', '/' | Real | ⇒ | Real |
| Int | '<','<=', '=', '#', '>=', '>' | Int | ⇒ | Bool |
| Real | '<','<=', '=', '#', '>=', '>' | Real | ⇒ | Bool |
| Char | '<','<=', '=', '#', '>=', '>' | Char | ⇒ | Bool |
| Bool | '**AND**', '**OR**' | Bool | ⇒ | Bool |
| | '**NOT**' | Bool | ⇒ | Bool |
| | '−' | Int | ⇒ | Int |
| | '−' | Real | ⇒ | Real |
| | '**TRUNC**' | Real | ⇒ | Int |
| | '**FLOAT**' | Int | ⇒ | Real |

- The identifiers `TRUE` and `FALSE` are predeclared in the language.
- The **FOR**-loop **BY**-expression must be a compile-time constant.
- Assignment is defined for scalars only, not for variables of structured type. In other words, the assignment `A:=B` is illegal if `A` or `B` are records or arrays.
- **READ** is only defined for scalar values (integers, reals, and characters).
- **WRITE** is defined for scalar values (integers, reals, and characters). and literal strings.
- A procedure's formal parameters and local declarations form one scope, which means that it is illegal for a procedure to have a formal parameter and a local variable of the same name.
- Parameters are passed by value unless the formal parameter has been declared **VAR**. Only L-valued esxpressions (such as '`A`' and '`A[5]`') can be passed to a **VAR** formal.
- Procedures cannot be nested.
- Identifiers have to be declared before they are used.

## A.4 Context Conditions

Below are the error conditions the compiler needs to check, organized by AST node.

**DECL:**

An identifier can only be declared once in each scope. A procedure's formal parameters and local declarations form one scope. If `ID` is declared more than once, the compiler should issue this error message:

```
<SEMANTIC_ERROR pos="..." message="Multiple declaration" argument="ID"/>
```

**VARDECL, FIELDDECL, FORMALDECL:**

1. The type name must be declared:
   ```
   <SEMANTIC_ERROR pos="..." message="Identifier not declared" argument="TypeName"/>
   ```
2. And, if the type name is declared, it has to be declared to be a type:
   ```
   <SEMANTIC_ERROR pos="..." message="Type identifier expected" argument="TypeName"/>
   ```

**CONSTDECL:**

1. The type name must be declared:
   ```
   <SEMANTIC_ERROR pos="..." message="Identifier not declared" argument="TypeName"/>
   ```
2. And, if the type name is declared, it has to be declared to be a type:
   ```
   <SEMANTIC_ERROR pos="..." message="Type identifier expected" argument="TypeName"/>
   ```
3. And, if it's declared a type, it has to be declared a *scalar* type (integer, character, real, boolean):
   ```
   <SEMANTIC_ERROR pos="..." message="Scalar type expected"/>
   ```
4. If the declared type is OK, you need to check that the expression is of the same type:
   ```
   <SEMANTIC_ERROR pos="..." message="Wrong expression type"/>
   ```
5. Regardless of the type checks above, the expression has to be constant-valued:
   ```
   <SEMANTIC_ERROR pos="..." message="Constant expression expected"/>
   ```

**ARRAYDECL:**

1. The type name must be declared:
   ```
   <SEMANTIC_ERROR pos="..." message="Identifier not declared" argument="TypeName"/>
   ```
2. And, if the type name is declared, it has to be declared to be a type:
   ```
   <SEMANTIC_ERROR pos="..." message="Type identifier expected" argument="TypeName"/>
   ```
3. The array size must be of type integer:
   ```
   <SEMANTIC_ERROR pos="..." message="Integer expression expected"/>
   ```
4. Regardless, of its type, the array size must be a constant expression:
   ```
   <SEMANTIC_ERROR pos="..." message="Constant expression expected"/>
   ```

**ASSIGN:**

1. The left hand and the right hand side must be of scalar (integer, real, char, boolean) type.
   ```
   <SEMANTIC_ERROR pos="..." message="Scalar type expected"/>
   ```
2. The left hand and the right hand side must be the same type.
   ```
   <SEMANTIC_ERROR pos="..." message="Type missmatch in assignment statement"/>
   ```

3. The left hand side must be a L-value, i.e. something you can assign to.

   ```
   <SEMANTIC_ERROR pos="..." message="Can't assign to a constant"/>
   ```

**PROCCALL:**

1. The designator must be a single declared identifier:

   ```
   <SEMANTIC_ERROR pos="..." message="Identifier not declared"/>
   ```

2. If the identifier is declared, it must be declared to be a procedure:

   ```
   <SEMANTIC_ERROR pos="..." message="Procedure identifier expected"/>
   ```

**WRITE:**
The expression must evaluate to an integer, real, character, or string.

```
<SEMANTIC_ERROR pos="..." message="INTEGER, REAL, CHAR, STRING type expected"/>
```

**READ:**

1. The designator must evaluate to an integer, real, or character:

   ```
   <SEMANTIC_ERROR pos="5" message="INTEGER, REAL, CHAR type expected"/>
   ```

2. The designator has to be an L-value (i.e. something you can assign to):

   ```
   <SEMANTIC_ERROR pos="..." message="Can't read to a constant"/>
   ```

**WHILE, REPEAT, IF1, IF2:**
The expression must be a boolean:

```
<SEMANTIC_ERROR pos="..." message="Boolean type expected"/>
```

**EXIT:**
`EXIT` must not occur outside of a loop:

```
<SEMANTIC_ERROR pos="..." message="EXIT only within LOOP"/>
```

**ACTUAL:**

1. There have to be the same number of actual and formal parameters:

   ```
   <SEMANTIC_ERROR pos="..." message="Too many actual parameters"/>
   <SEMANTIC_ERROR pos="..." message="Too few actual parameters"/>
   ```

2. Regardless, the actual parameter has to be assignable to the corresponding formal parameter.

   ```
   <SEMANTIC_ERROR pos="..." message="Actual/formal parameter type missmatch"/>
   ```

3. Regardless, if a formal parameter is declared to be a **VAR** parameter, then the corresponding actual has to be an L-value (cannot be a constant):

   ```
   <SEMANTIC_ERROR pos="..." message="VAR formal parameter requires variable actual"/>
   ```

**VARREF:**

1. The identifier has to be declared:

   ```
   <SEMANTIC_ERROR pos="..." message="Identifier not declared" argument="ID"/>
   ```

2. The identifier must be a formal parameter, a variable (global or local), a constant identifier, or a procedure:

   ```
   <SEMANTIC_ERROR pos="..." message="Variable expected"/>
   ```

**INDEX:**

1. The index expression must evaluate to an integer type:

   ```
   <SEMANTIC_ERROR pos="..." message="Integer type expected"/>
   ```

2. The designator must be of array type:

   ```
   <SEMANTIC_ERROR pos="..." message="Array variable expected"/>
   ```

**FIELDREF:**

1. The designator must be of record type:

   ```
   <SEMANTIC_ERROR pos="..." message="Record variable expected"/>
   ```

2. If the designator is or record type, then the field must be declared in the record:

   ```
   <SEMANTIC_ERROR pos="..." message="Field identifier not declared" argument="ID"/>
   ```

**BINARY, UNARY:**

The table in the previous section gives the semantic rules for expressions. For constant expressions, division by zero isn't allowed.

1. In arithmetic expressions, when a real or integer is expected, issue:

   ```
   <SEMANTIC_ERROR pos="..." message="Numeric type expected"/>
   ```

2. For $a\%b$, if $a$ and $b$ aren't integer types, issue

   ```
   <SEMANTIC_ERROR pos="..." message="Integer type expected"/>
   ```

3. For `AND`, `OR`, `NOT`, if the arguments aren't boolean types, issue

   ```
   <SEMANTIC_ERROR pos="..." message="Boolean type expected"/>
   ```

4. When the arguments to comparison operators (`#, <, >, <=, >=, =`) aren't integers, reals, booleans, or chars, issue

   ```
   <SEMANTIC_ERROR pos="..." message="Scalar or reference type expected"/>
   ```

   (Reference type refers to another version of Luca that also has pointer types.)

5. For `TRUNC` and `FLOAT`, respectively, when the arguments are of the wrong type, issue

   ```
   <SEMANTIC_ERROR pos="..." message="Real type expected"/>
   <SEMANTIC_ERROR pos="..." message="Integer type expected"/>
   ```

6. Otherwise, if the left and right hand sides are of different types, issue:

   ```
   <SEMANTIC_ERROR pos="..." message="Type missmatch"/>
   ```

### A.5 Checked Runtime Errors

- Arrays are indexed from 0; that is, an array declared as `ARRAY 100 OF INTEGER` has the index range `[0..99]`. It is a checked run-time error to go outside these index bounds. You should generate the following error message:

  ```
  <RUNTIME_ERROR pos="3" message="Array index out of range"/>
  ```

  Note that the source code line number is part of the error message.

- Division by zero should generate this error message:

  ```
  <RUNTIME_ERROR pos="3" message="Division by zero"/>
  ```

## B The Luca Translator

### B.1 The Luca Virtual Machine

- The Luca virtual machine is a word-addressed machine. Words are 32 bits wide. The size of all basic types (integers, reals, booleans, and chars) is one word.
- The Luca virtual machine is a stack machine. Conceptually, there is just one stack and it is used both for parameter passing and for expression evaluation. An implementation may – for efficiency or convenience – use several stacks. For example, in a three stack implementation one stack can be used to store activation records, one can be used for integer arithmetic and one can be used for real arithmetic.
- Execution begins at the (parameterless) procedure named `$MAIN`.
- Large value parameters are passed by reference. It is the responsibility of the called procedure to make a local copy of the parameter. For example, if procedure `P` passes an array `A` by value to procedure `Q`, `P` actually pushes the *address* of `A` on the stack. Before execution continues at the body of `Q`, a local copy of `A` is stored in `Q`'s activation record. The body of `Q` accesses this copy. A special instruction `Copy` is inserted by the front end to deal with this case.
- When a `ProcCall` instruction is encountered the arguments to the procedure are on the stack, with the first argument on the top. In other words, arguments are pushed in the *reverse* order.
- Variables whose names start with "$" are temporaries inserted by the front end. They are currently only used in the implementation of `FOR`-loops.

### B.2 The tree intermediate code

The front-end generates an intermediate representation that is a sequence of expression trees. Below are listed the tree-code node types the frontend generates.

## Declarations

`Version(Major,Minor,Pos)` The version of the intermediate code language.

`VarDecl(Symbol,Pos)` `VarDecl` declares a global or local variable. `Symbol` is the symbol table entry for the variable, from which we can retrieve information such as size(`S.GetSize()`), level of declaration (`S.GetLevel()`), type(`S.GetType()`), and address(`S.GetOffset()`).

`FormalDecl(Symbol,Pos)` Declares the formal parameter of a procedure. `Symbol` is the symbol table entry, from which we can retrieve information such as mode(`S.GetFormalMode()`), size(`S.GetSize()`), level of declaration (`S.GetLevel()`), type(`S.GetType()`), and offset(`S.GetOffset()`). Note that the offset returned by `S.GetOffset()` is a suggestion only; you may find that a different activation record layout suits your back-end better.

`TypeDecl(Symbol,Pos)` Declares a record or array type. `Symbol` is the symbol table entry.

## Loads and Stores

`Store(Type,Left,Right,Pos)` `Left` is an expresion tree computing an address. `Right` is an expression tree computing a value (it's type is given by `Type`) to be stored at that address.

`Load(Type,Des,Pos)` `Des` is an expression tree computing an address. `Load` should load the value (whose type is given by `Type`) stored at that address.

## Expressions

`BinExpr(Op,Type,Left,Right,Pos)` A node in an expression tree that computes `Left Op Right`. `Op` is a string.

`UnaryExpr(Op,Type,Left,Pos)` A node in an expression tree that computes `Op Left`. `Op` is a string. Type is a symbol table reference.

`LoadLit(Type,Value,Pos)` Load the literal value `Value`. In case of *strings*, the *address* should be loaded, not the value.

## Designators

`AddressOf(Symbol,Type,Pos)` Load the address of `Symbol`(which could be a global variable, local variable, or formal parameter). `Type` is the type of the symbol.

`IndexOf(Type,Base,Index,Pos)` Compute the *address* of an array element, i.e. `Base` + `S.GetSize(S.GetArrayElementType(Type))` $*$ `Index`. `Base` is an expression tree computing the base address of the array. `Index` is an expression tree computing the index value. `Type` is a symbol table reference to the array from which we can retrieve information such as `S.GetArrayCount()` and `S.GetArrayElementType()`. It's a *checked, fatal, run-time error* for `Index` to be `<0` or `>S.GetArrayCount()-1`.

`FieldOf(Type,Field,Base,Pos)` Compute the *address* of a record field, i.e. `Base` + `S.GetOffset(Field)`. `Base` is an expression tree computing the base address of the record. `Field` is a symbol table entry for the field from which we can retrieve information such as offset(`S.GetOffset()`). `Type` is a symbol table reference to the record type.

## Control

`Branch(Op,Type,Left,Right,Label,Pos)` Equivalent to `if Left Op Right then goto Label`. `Left` and `Right` are expression trees, `Op` is a string, and `Label` is the number of the label to which we should jump.

`Goto(Label,Pos)` Jump to `Label`.

`Label(Label,Pos)` Declare a `Label`.

## Input and Output

`Write(Type,Expr,Pos)` Write the value of `Expr`(an expression tree) to the standard output. If `Expr` is a(constant) string, `Expr` will compute the string's *address*, not its value.

`Read(Type,Des,Pos)` Read a value into the address held by `Des`, an expression tree. The type of the data to be read is given by `Type`, a symbol table reference.

`WriteLn(Pos)` Write a newline character to the standard output.

## Procedure call

`ProcCall(Symbol,Actuals,Pos)` Call procedure `Symbol`. `Symbol` is a symbol table entry from which we can retrieve information such as formal parameters(`S.GetProcFormals()`), local data size(`S.GetLocalSize()`), level of declaration (`S.GetLevel()`), and size of formal parameters (`S.GetFormalSize()`). `Actual` is an expression tree.

`Actual(Type,Formal,Expr,Next,Pos)` `Actual` nodes are linked together on `Next` to make a list of actual parameters. `Expr`(an expression tree) computes the value/address of the actual. `Formal` is a reference to the symbol table entry for the corresponding formal parameter, from which we can retrieve information such as size(`GetSize()`, offset within the activation record (`GetOffset()`), number(`GetFormalNumber()`), and mode(`GetFormalMode()`).

`Null(Pos)` `Null` terminates a sequence of `Actual` nodes.

## B.3 Symbols

These are the procedures available in `sym/*.java`, to extract data on symbols:

| Procedure | Description |
|---|---|
| `S.GetName()` | Get the name of symbol `S`. |
| `S.GetNumber()` | Get the unique identifying number of symbol `S`. |
| `S.GetLevel()` | Get the declaration level of symbol `S`. |
| `S.SetLevel(Level)` | Set the declaration level of symbol `S`. |
| `S.SetSize(Size)` | Set size of symbol `S`, a type, formal, field, variable, or constant. |
| `S.GetSize()` | Set size of symbol `S`, a type, formal, field, variable, or constant symbol. |
| `S.GetType()` | Get the type of a variable, field, constant, or formal. |
| `S.SetType(Type)` | Set the type of a variable, field, constant, or formal. |
| `S.GetOffset()` | Get the offset/address of a variable, field, or formal. |
| `S.SetOffset(Offset)` | Set the offset/address of a variable, field, or formal. |
| `S.GetArrayCount()` | Return the number of elements in the array `S`. |
| `S.SetArrayCount(Count)` | Set the number of elements in the array `S`. |
| `S.GetArrayElementType()` | Return the element type(a symbol) of array `S`. |
| `S.SetArrayElementType(ET)` | Set the element type `ET` (a symbol) of array `S`. |
| `S.GetFields()` | Get the fields (a `sym.SyTab`) of record type `S`. |
| `S.SetFields(Fields)` | Set the fields (a `sym.SyTab`) of record type `S`. |
| `S.GetConstantValue()` | Get the value of a constant. |
| `S.SetConstantValue(Value)` | Set the value of a constant. |
| `S.GetProcLocals()` | Get the local variables(a `sym.SyTab`) of procedure `S`. |
| `S.SetProcLocals(Locals)` | Set the local variables(a `sym.SyTab`) of procedure `S`. |
| `S.GetProcFormals()` | Get the formal parameters(a `sym.SyTab`) of procedure `S`. |
| `S.SetProcFormals(Formals)` | Set the formal parameters(a `sym.SyTab`) of procedure `S`. |
| `S.GetFormalParam(Formals, N)` | Get formal parameter number `N` of procedure `S`. |
| `S.GetLocalSize()` | Get the size of local variables of procedure `S`. |
| `S.SetLocalSize(Size)` | Set the size of local variables of procedure `S`. |
| `S.GetFormalSize()` | Get the size of formal parameters of procedure `S`. |
| `S.SetFormalSize(Size)` | Set the size of formal parameters of procedure `S`. |
| `S.GetFormalNumber()` | Get formal number of formal parameter `S`. |
| `S.SetFormalNumber(Number)` | Set formal number of formal parameter `S`. |
| `S.GetFormalMode()` | Get formal mode(string `"VAR"` or `"VAL"`) of formal parameter `S`. |
| `S.SetFormalMode(Mode)` | Set formal mode(string `"VAR"` or `"VAL"`) of formal parameter `S`. |
| `S.GetEnumValue()` | Get the value (an int) of enumeration identifier `S`. |
| `S.SetEnumValue(Value)` | Set the value (an int) of enumeration identifier `S`. |