FACULTEIT INGENIEURSWETENSCHAPPEN
EN ARCHITECTUUR

Compilers Labo

Bert De Saffel

Master in de Industriële Wetenschappen: Informatica Academiejaar 2018–2019

Gecompileerd op 20 mei 2019

UNIVERSITEIT
GENT

Inhoudsopgave

Hoofdstuk 1

Prologue

1.1 Using Docker

Each lab has a `files` directory which you should change your working directory to. Run the following command to start an interactive Docker container.

```
$ docker run --rm -it -v "$(pwd)":/files tbesard/compilers:practx
```

After the container has launched, you can navigate to the correct directory with

`cd files`

You can edit these files on your local host machine, but it is recommended to edit in the container (see section 1.2).

1.2 Installing the nano editor

Editing files on a Windows computer brings incompatibility with linux tools. It is recommended to install a text editor on the container.

```
$ apt-get update
$ apt-get install nano
```

Hoofdstuk 2

Lexing

2.1 Setup

The goal of this lab is to process source files in a C-like language, and generate a stream of tokens along with location information, or an error message if the syntax of the file is invalid. This will be realised with the Flex lexical analyzer generator.

The project can be executed using the docker image `tbesard/compilers:pract1` (see Using Docker 1.1 for more information). Run `make` and execute the resulting `main` executable on a source file:

```
make && ./main test/dummy.c
```

2.2 Basic lexing

The first part is to add some new definitions and rules in `lexer.l`. The definitions consist of a label and a regex. The purpose of these definitions is to be used in rules. The rules match a regular expression and tells the lexer what to return.

```
/*
  Definitions
*/
DIGIT  [0-9]
WHITESPACE     [ \t\n\r]
ID       [a-z_][a-z0-9_]*
COMMENT     "//"[^\n\r]*
FLOAT      [0-9]+\.[0-9]*
STRING  \"[^\n]+\"


/*
  Rules
*/
"="           return EQUAL;
"=="          return CEQ;
"!="          return CNE;
"<"           return CLT;
">"           return CGT;
```

```
"<="          return CLE;
">="          return CGE;
"+"           return PLUS;
"-"           return MINUS;
"*"           return MUL;
"/"           return DIV;
"^"           return EXP;
"%"           return MOD;
","           return ',';
";"           return ';';
"("           return '(';
")"           return ')';
"{"           return '{';
"}"           return '}';
"["           return '[';
"]"           return ']';
"return"    return RETURN;
"if"        return IF;
"else"      return ELSE;
"while"     return WHILE;
"for"       return FOR;
{COMMENT}        /*must be ignored*/
{FLOAT}          return FLOAT;
{DIGIT}+         return INTEGER;
{WHITESPACE}+
{ID}             return IDENTIFIER;
{STRING}         return STRING;
.                error("unknown symbol");
```

2.3 Location information

To add location information, the method `update_location()` in `lexer.l` must be updated. This method will be called everytime the lexer returns a token from the rules. The struct `Location` in `lexer.hpp` contains two integers for the line number and column number. The `Lexer` class contains two `Location` structs which respectively represent the beginning line and column and the ending line and column of the token. The attribute `yytext` contains a string representation of the current token.

```
void Lexer::update_location() {
  if(yytext[0] == '\n' || yytext[0] == '\r'){
    begin.column = 1;
    end.column = 1;
    begin.line++;
    end.line++;
  } else {
    begin.column = end.column;
    end.column += strlen(yytext);
  }
}
```

2.4 Error reporting

To add error messages, the method `update_location()` in `lexer.l` must first be updated so that it keeps string information of the current line.

```
char *buffer = (char*)malloc(YY_BUF_SIZE);
void Lexer::update_location() {
  if(yytext[0] == '\n' || yytext[0] == '\r'){
    begin.column = 1;
    end.column = 1;
    begin.line++;
    end.line++;
    free(buffer);
    buffer = (char*)malloc(YY_BUF_SIZE);
  } else {
    begin.column = end.column;
    end.column += strlen(yytext);
    strcat(buffer, yytext);
  }
}
```

Additionaly, the method `error()` in `lexer.l` must be implemented. This method gets called when the lexer encounters an error.

```
void Lexer::error(const std::string &message) {
  std::cerr << "Syntax error: " << message
            << " at line " << begin.line
            << " at column " << begin.column
            << "\n";
  std::cerr << buffer << "\n";
  printf("%*c^\n", (int)begin.column - 1, ' ');
}
```

2.5 Context-sensitive rules

A context-sensitive rule are useful for patterns which cannot be matched with simple regular expressions. Here we add the neccesary definitions and rules to parse block comments. We add a definition `BLOCK_COMMENT`, which is preceded by `%x`. This symbol means that `BLOCK_COMMENT` is an exclusive state, which means that the lexer will only match rules which are tagged `BLOCK_COMMENT` once it enters the state. Next a new set of rules is implemented. The `INITIAL` state is a predefined state which marks the entry point for any other state. We specify that a block comment starts with /*, and if that pattern occurs, the state `BLOCK_COMMENT` is activated. Only patterns in the `BLOCK_COMMENT` state block will be matched. Now four possible rules can be matched:

1. The end of a comment is specified with */. If this occurs the comment has succesfully ended and we go back to the initial state.
2. If it's not the end of the coment, there could be an infinite number of characters we have to eat ([^*\n]+). We do not include the newline character because we want it seperately to increase the line number.
3. When we encounter a newline, we want to increase the line number of the lexer.

4. When the end of a file is reached, the block comment was not terminated, resulting in an error message.

```
/*
  Definitions
*/
%x BLOCK_COMMENT

/*
  Rules
*/
<INITIAL>{
  "/*" BEGIN(BLOCK_COMMENT);
}
<BLOCK_COMMENT>{
  "*/"   BEGIN(INITIAL);
  [^*\n]+
  \n     yylineno++;
  <<EOF>> {
    error("unterminated block comment");
    BEGIN(INITIAL)
  }
}
```

Hoofdstuk 3

Parsing

3.1 Setup

The goal of this lab is to complete the implementation of the parser for the same C-like language we have worked with in the first lab. This will be realised with the Bison parser-generator tool in order to generate the BNF grammar that implements the rule of this language.

The grammar is to be processed by the Bison LALR(1) parser generator. A LALR parser, or Look-Ahead LR parser, parses a text according to a set of production rules specified by a formal grammar. The **LR** stands for **L**eft-to-right, **R**ightmost derivation. A left-to-right parser reads input text from left to right. Rightmost derivation always choses the rightmost nonterminal to rewrite. The (1) denotes one-token lookahead, which allows the parser to peek ahead at 1 input symbol before deciding how to parse earlier symbols.

After mounting the image (see section 1.1), configure the project using

```
cmake .
```

This only need to be done once. To actaully compile the files, use

```
make
```

After compiling, a binary named `cheetah` will be created. This binary will visualize the AST of a source file in GraphViz DOT format. This DOT format can be rendered to a PNG.

```
./cheetah ../test/dummy.c > dummy.dot && dot -Tpng dummy.dot > dummy.png
```

3.2 Function call

We will first implement the function call expression. The rule needs to produce an expression, and as such needs to be part of the `expr` production. A function call is represented by a `AST::CallExpr` object and the argument list is of type `AST::ExprList`. In `parser.y`, we first add two new semantic values to the `%union` clause.

```
%union {
   ...
   AST::CallExpr *call_expr_t;
   AST::ExprList *exprlist_t;
```

```
  ...
}
```

These symbols also need to be classified as a nonterminal symbol.

```
%type <call_expr_t> call
%type <exprlist_t> call_args
```

Now we can create a rule for a function call. The first step is to add `call` to the `expr` rule.

```
expr:
    ... {
      ...
    }
  | call
  | ...
```

Now a `call` rule must be made. A function call consists of two parts: the function identifier and the optional argument list. The semantic value of this rule is a `AST::CallExpr` object.

```
call:
  ident '(' call_args ')' {
    $$ = new AST::CallExpr($1, *$3);
    delete $3;
  };
```

Consequently, a rule for `call_args` must be made. There are three cases:

1. The argument list is empty.
2. The argument list contains only one expression.
3. The argument list contains more than one expressions.

In the first case, we can use the pseudo rule `%empty` to indicate there is nothing. In the last case, we have to define a recursive rule that creates an expression list. The function `Sema::ParseExprList()` is used which returns pointers to `AST::ExprList` objects.

```
call_args:
    %empty {
      $$ = sema.ParseExprList();
    }
  | expr {
      $$ = sema.ParseExprList($1);
    }
  |   call_args ',' expr {
      $$ = sema.ParseExprList($3, $1);
    };
```

3.3 Literals

The next step will take care of the rules for literals (integer, floating-point and string). These rules return an expression but require a semantic action to parse the token value to a semantic value.

The `Sema` class does not contain such functions and have to be defined for the `IntLiteral` and `FloatLiteral` types in `semaexpr.cpp`.

```
AST::IntLiteral *Sema::ParseIntLiteral(const Location &Loc,
                                       std::string IntToken){
  return new AST::IntLiteral(atoi(IntToken.c_str()));
}
AST::FloatLiteral *Sema::ParseFloatLiteral(const Location &Loc,
                                           std::string FloatToken){
  return new AST::FloatLiteral(atof(FloatToken.c_str()));
}
AST::StringLiteral *Sema::ParseStringLiteral(const Location &Loc,
                                             std::string StringToken){
  // remove quotes " " around string
  for(int i = 1; i < StringToken.size(); i++){
    StringToken[i - 1] = StringToken[i];
  }
  StringToken.resize(StringToken.size() - 2)
  return new AST::StringLiteral(StringToken);
}
```

Before rules will be created, we first add a new `literal` nonterminal symbol of type `expr_t`.

```
%type <expr_t> ... literal ...
```

This nonterminal is also added to the `expr` rule.

```
expr:
    ... {
      ...
    }
  | call
  | literal
  | ...
```

Now the `literal` rule can be defined.

```
literal:
    INTEGER {
      $$ = sema.ParseIntLiteral(@$, yytext(lexer));
    }
  |   FLOAT {
      $$ = sema.ParseFloatLiteral(@$, yytext(lexer));
    }
  | STRING {
      $$ = sema.ParseStringLiteral(@$, yytext(lexer));
    };
```

3.4 Operators

To allow operators to be parsed, it is neccesary to define the precedence rules first. The precedence list is defined in reverse order such that the lowest item in the list has the highest associativity. In the following example, the `EXP` is the most tightly bound, right-associative operator.

```
%precedence EQUAL
%nonassoc CEQ CNE CLT CLE CGT CGE
%left PLUS MINUS
%left MUL DIV MOD
%precedence UNARY
%right EXP
```

These rules do not need semantic actions, and as such can directly construct the relevent AST Nodes. All these rules are put in the `expr` rule.

```
expr:
    ... {
      ...
    }
  | call
  | literal
  | expr CEQ expr           { $$ = new AST::BinaryOp($1, AST::Operator::CEQ,   $3); }
  | expr CNE expr           { $$ = new AST::BinaryOp($1, AST::Operator::CNE,   $3); }
  | expr CLT expr           { $$ = new AST::BinaryOp($1, AST::Operator::CLT,   $3); }
  | expr CLE expr           { $$ = new AST::BinaryOp($1, AST::Operator::CLE,   $3); }
  | expr CGT expr           { $$ = new AST::BinaryOp($1, AST::Operator::CGT,   $3); }
  | expr CGE expr           { $$ = new AST::BinaryOp($1, AST::Operator::CGE,   $3); }
  | expr PLUS expr          { $$ = new AST::BinaryOp($1, AST::Operator::PLUS,  $3); }
  | expr MINUS expr         { $$ = new AST::BinaryOp($1, AST::Operator::MINUS, $3); }
  | expr MUL expr           { $$ = new AST::BinaryOp($1, AST::Operator::MUL,   $3); }
  | expr DIV expr           { $$ = new AST::BinaryOp($1, AST::Operator::DIV,   $3); }
  | expr MOD expr           { $$ = new AST::BinaryOp($1, AST::Operator::MOD,   $3); }
  | MINUS expr %prec UNARY  { $$ = new AST::UnaryOp(    AST::Operator::MINUS, $2); }
  | PLUS expr %prec UNARY   { $$ = new AST::UnaryOp(    AST::Operator::PLUS,  $2); }
  | expr EXP expr           { $$ = new AST::BinaryOp($1, AST::Operator::EXP,   $3); }
  ;
```

3.5 Control flow

To make implementation easier, we first add new nonterminal symbols of type `stmt_t`.

```
%type <stmt_t> ... ifstmt whilestmt forstmt
```

The `stmt` rule can now be expanded with these three new symbols. We also define the `return` inline.

```
stmt:
  ...
  | ifstmt
  | whilestmt
  | forstmt
  | RETURN expr ';' {
      $$ = new AST::ReturnStmt($2);
    }
  | RETURN ';' {
      $$ = new AST::ReturnStmt();
    }
```

Each of the other three statements is a rule on its own.

```
ifstmt:
    IF '(' expr ')' block {
```

```
        $$ = new AST::IfStmt($3, $5);
    }
  | IF '(' expr ')' block ELSE block {
        $$ = new AST::IfStmt($3, $5, $7);
    }

whilestmt:
    WHILE '(' expr ')' block {
        $$ = new AST::WhileStmt($3, $5);
    }

forstmt:
    FOR '(' forinit ';' expr ';' expr ')' block {
        $$ = new AST::ForStmt($<stmt_t>3, $5, $7, $9);
    }
forinit:
    var_decl {
        $<stmt_t>$ = new AST::DeclStmt($1);
    }
  |   expr {
        $<stmt_t>$ = $1;
    }
```

Hoofdstuk 4

Code Generation

4.1 Setup

Run a docker container and configure the project with CMake.

```
$ docker run --rm -it  --cap-add=SYS_PTRACE
                --security-opt seccomp=unconfined
                -v "$(pwd)":/files tbesard/compilers:pract3
$ cd /files
$ cmake .
```

Run `make` to compile the whole project after each change. Use `cheetah` to generate the assembly code.

```
$ ./cheetah test/dummy.c
.globl main
main:
  pushq $1
  popq  %rax
  ...
```

It is also possible to write this to a file:

```
$ ./cheetah -o test/dummy.S test/dummy.c
```

Een executable aanmaken kan met `make dummy`, of als je alle testen wilt compileren kan je `make test` gebruiken.

4.2 Debugging

Met `gdb` kan een executable geïnspecteerd worden.

```
$ gdb ./test/dummy
(gdb) run
```

4.3 Compiler infrastructure

The `codegen.hpp` header defines three important datastructures:

- `Program`: This represents the program that is being emitted, and is accesible as the argument to each `emit` function. It contains a list of `Blocks`.
- `Block`: A block is identified by a label and contains a list of `Instructions`.
- `Instruction`: An instruction has three fields:
  - `name`: the textual representation of the instruction name.
  - `arguments`: a list of arguments.
  - `comment`: an optional string that will be emitted as part of the generated code.

4.4 Emitting code

We will implement the compiler as a stack machine. This means that it should push and pop values onto the stack and only use registers when absolutely neccesary. An explanation of the most useful registers:

- `%rax`: Temporary register, mainly used as the return register.
- `%rbx`: Callee-saved register which can optionally be used as a base pointer.
- `%rbp`: Callee-saved register which can optionally be used as a frame pointer.
- `%rdi`: Used to pass the first argument to functions.
- `%rsi`: Used to pass the second argument to functions.
- `%rdx`: Used to pass the third argument to functions. Can also be used as the second return register.
- `%rcx`: Used to pass the fourth argument to functions.
- `%r8`: Used to pass the fifth argument to functions.
- `%r9`: Used to pass the sixth argument to functions.
- `%r12-r15`: Callee-saved registers.
- `%esp`: The stack pointer (32 bit).

A short summary of the special registers:

- **Stack Pointer**: The stack pointer points to the top of the stack. All locations beyond the stack pointer are considered to be garbage, and all locations before the stack pointer are considered to be allocated.
- **Frame Pointer**: The area on the stack devoted to local variables, parameters, return address and other temporaries for a function is called the function's `stack frame`. The frame pointer points at the beginning of a stack frame such that the stack pointer can be restored to the frame pointer. Equivalently, the frame pointer contains the value of the stack pointer just before a function is called.

- **Base Pointer**: The base pointer is derived from the stack pointer and is used to travel trough the stack.

Each AST object now has an `emit` function, which purpose is to generate assembly code for that AST object.

4.4.1 Function calls

In `emit.cpp`, complete the implementation of `CallExpr::emit(Program &prog) const` and implement the following features:

- **emit and store arguments**

```
int j = 0;
for (size_t i = 0; i < argc; i++){
  args[i]->emit(prog)
  prog << Instruction{"popq", {param_regs[j]}};
  j++;
}
```

  Remember that this is a method of the class `CallExpr`, so we can use the attributes `args` and `name` of this class. The attribute `args` is of type `ExprList`, which can contain pointers to various expression types such as `FloatLiteral` or `Assignment` (see `expr.hpp`). First we call the emit method for each `Expr` in this list. Afterwards the value gets put into the predefined parameter registers.

- **generate a call**

```
prog << Instruction{"call", {name->string}};
```

  Here we need to emit a `call` instruction. An instruction has a `name`, a list of `arguments` and optionally a `comment`. The name of the instruction is obviously `call`. In this case the list of arguments only contain one element: the name of the function. We opted to not include a comment here since a `call` instruction is fairly obvious.

- **return a value**

```
if(std::get<0>(decl) == T_void){
  prog << Instruction{"pushq", {"$0xABCDEF"}};
} else if(std::get<0>(decl) == T_int){
  prog << Instruction{"pushq", {"%rax"}};
}
```

  Our language only has two possible return types: `void` and `int`. Because we implement the compiler as a stack machine, we also have to put a void sentinel value on the stack, which is `$0xABCDEF`. For an integer value we just push the return value on the stack.

4.4.2 Function Declarations

In `emit.cpp`, complete the implementation of `FuncDecl::emit(Program &prog) const` and implement the following features:

- **The function prologue**:

- save callee-saved registers

```
for(const std::string& s : callee_saved_regs){
  prog << Instruction{"pushq", {s}};
}
```

- set the base pointer

```
prog << Instruction{"pushq", {"%rbp"}};
prog << Instruction{"movq", {"%rsp", "%rbp"}};
```

- align the stack pointer by 16 bytes

```
prog << Instruction{"subq", {"$16", "%rsp"}};
```

- **The function epilogue**:

  - restore the stack pointer

```
prog << Instruction{"movq", {"%rbp", "%rsp"}};
prog << Instruction{"pushq", {"%rbp"}};
```

  - restore callee-saved registers

```
for(int i = callee_saved_regs.size() - 1; i >=0 ; i--){
  prog << Instruction{"popq", {callee_saved_regs[i]}};
}
```

Hoofdstuk 5

Compiler transformation at LLVM IR level

5.1 Setup

Installing `perf` on the docker machine:

```
apt-get update
apt-get install linux-tools-generic
/usr/lib/linux-tools/4.18.0.20-genric/perf ...
```

5.2 Introduction to LLVM bitcode

- LLVM intermediate representation is a low-level Static Single Assignment representation.
- Code is grouped in *basic blocks*.
  - Control flow within the application can only change at the boundaries of basic blocks.
  - Each basic block has to be terminated with an instruction directing control flow.
  - These instructions are derived from the `TerminatorInst` super class.
- To modify the IR:
  - `BasicBlocks::Create` adds new basic blocks to a function.
  - `Value::replaceAllUsesWith` changes the value of the `Value` object.
  - For more complex operations, use the `IRBuilder` class which points at a location in the IR. The `IRBuilder::Create` method inserts code after this location.
- Each LLVM object has a `dump()` method which is very handy for debugging purposes.

5.3 Array accesses in LLVM

- Goal of this lab:
  - Prevent out of bounds crashes.
  - Display a message to the user when such an error occurs.

5.4 Protecting array accesses

The operands of `getelementptr` instruction:

- A base pointer containing the memory address of the array;
- An index stepping in terms of the base pointer;
- An index stepping in terms of pointer elements.

5.5 Implementation

5.5.1 Decode array accesses

The first step is to determine which array index is accessed. For every `getelementptr` instruction we which to analyse if the indexing is out of bounds or not. First the whole source is iterated to find such instructions:

```
std::list<GetElementPtrInst *> WorkList;
for(auto &FI : F)
  for(auto &BI : FI)
    if(auto *GEP = dyn_cast<GetElementPtrInst>(&BI))
      WorkList.push_back(GEP);
```

Now every instruction can be iterated and some basic information can be gathered, such as the number of elements of the array:

```
for(auto *GEP : WorkList){
  ArrayType* array = (ArrayType*) GEP->getSourceElement();
  uint64_t numOfElements = array->getNumElements();

  // implement checks
}
```

The further implementation of the checks are discussed in the following sections.

Constant Integer Expressions

We first start with constant integer expressions and make a test file called `constant_indices.c`:

```
int foo[10];
foo[5] = 1;
foo[0] = 1;
foo[10] = 1;
```

We create an array that can hold 10 integers. The first two indexing operations are valid while the last one is invalid. Constant expressions can be evaluated at compile time so we can abort the compilation prematurely here.

```
if(GEP->hasAllConstantIndices()){
  User::const_op_iterator indices = GEP->idx_begin() + 1;
  Value* iVal = indices->get();
  ConstantInt* integer = (ConstantInt*) iVal;
  accessIndex = integer->getValue()->getLimitedValue();
  if(accessIndex >= numOfElements){
    std::ostringstream reason;
    reason << "Index out of bounds(line " << line <<
                                    ", column " << column <<
                                    ")\n\tIndex: " << accessIndex <<
                                    "\n\tMax: " << numOfElements - 1 <<
                                    "\n";
    report_fatal_error(...);
  }
}
```

This simple case will stop the compilation for expressions such as `foo[10]`.