LoliCCompiler

LOLI : LOLI Oriented Language Implementation

倪昊斌/RobbinNi

2015.4~5

~核融合炉 Ver~

目录

1. Introduction

LoliCCompiler 是上海交通大学 13 级 ACM 班编译原理课程的一项学生项目。作者为 13 级 ACM 学生倪昊斌。目标实现一个采用现代编译器的通用架构模式，以 C 语言的一个较为精简的子集为源语言，以 MIPS 指令集的汇编器 SPIM 为目标机，并能够进行一些较为基本的编译器优化的编译器。编程语言是 JAVA。本文就这一项目所采取的基本架构、所支持的特性和部分具体实现进行报告式的介绍。

1.1 基本架构

LoliCCompiler 整体可分为五个部分：

1、分词和语法分析(Lexing & Parsing)
- 原程序(代码)→符号流→抽象语法树(AST)
- 检查原程序中存在的语法错误

2、语义检查(Semantic Check)
- 抽象语法树(AST)→中间表示树(IRT)
- 检查原程序中存在的语义错误

3、中间表示(Intermediate Representation)
- 中间表示树(IRT)→三地址代码(IR)
- 静态单赋值形式(SSA)
- 控制流优化

4、代码生成(Code Generation)
- 三地址代码(IR)→MIPS 汇编代码
- 全局（跨过程）寄存器分配

5、优化(Optimization)
- 公共子表达式消除
- 冗余复制消除
- 死代码消除

而这一架构也是现代编译器多采用的围绕不同的中间表示之间的转换进行的模块化架构的一个微缩版本。后文将按照这一顺序逐模块进行介绍。

1.2 主要特性

LoliCCompiler 支持许多独特特性，具体可分为语言特性、功能特性和优化特性三类。

1.2.1 语言特性

1、复杂类型系统

• 支持结构/联合类型、函数类型、数组类型、指针类型的任意嵌套。

2、内嵌函数支持

• 可以在函数体内定义仅能在这一函数内调用或通过函数指针进行调用的内嵌函数。

3、未定大小数组

• 允许声明未定大小的数组类型。LoliCComplier 将会根据初始化列表大小等信息自动计算数组大小。

4、typedef 支持

• 允许通过 typedef 简化复杂类型的定义。LoliCCompiler 可以正确区分被 typedef 的类型和与之同名的变量。

5、函数指针

• 可以定义、赋值、传递、调用函数指针，以实现多态、代码隐藏等功能。

6、高阶函数

• 对于内嵌函数，支持通过函数指针将其作为参数传递。LoliCCompiler 实现了 gcc 的 trampoline 技术来支持对于调用函数的变量的正确访问。但执行运行时生成代码这一特性目前暂不被目标机 SPIM 所支持。

7、scanf 支持

除了 printf/getchar/malloc，LoliCCompiler 和 LoliCInterpreter 均额外支持 STL 中的 scanf 函数，支持%d、%s、%c 作为描述符。

1.2.2 功能特性

1、用户友好型 GUI

• 设计简洁的 GUI 控制面板集成了 LoliCCompiler 的各项特性。用户可以方便而清晰地看到 LoliCCompiler 的内部中间表示细节。

2、代码美化工具

• 可以对于一段源代码进行自动缩进、插入空格来进行美化，使之更易于阅读。

3、语法/语义错误信息

• 对于语法和语义错误，能够在 GUI 界面上返回具体的错误信息。

4、C 语言解释器(LoliCInterpreter)

• LoliCCompiler 内含一个功能完整的 C 语言解释器(LoliCInterpreter)，可以直接运行 C 语言源代码，并且实现了与用户进行 IO 交互的仿控制台。

1.2.3 优化特性

1、控制流优化

• 在代码生成中，对于分支语句判断表达式的计算进行了分析，能够大量减少生成的跳转指令。

2、静态单赋值形式(SSA)

• 将中间表示转换为静态单赋值形式(SSA)并进行优化再转回三地址代码进行代码生成。

3、全局寄存器分配

• 正确实现了循环分析、过程间数据流/控制流分析、符号寄存器汇编生成、并以此为

基础实现了溢出改进的图染色寄存器分配算法。优化效果显著。

4、公共子表达式消除

- 能够提出在程序中出现的公共子表达式，并通过替换减少计算量。

5、冗余复制消除

- 基于 SSA，能够通过等价替换消除冗余的复制指令。

6、死代码消除

- 基于 SSA，对于定义后不再会被使用的变量，其定义会被消除。

2 Lexing & Parsing

分词和语法分析(Lexing & Parsing)是 LoliCCompiler 的第一个模块。主要是将源代码通过分词转化为符号流，再通过语法分析转化为抽象语法树。我使用了正则表达式+Jflex 生成分词程序，上下文无关文法+JCup 生成语法分析程序，并通过在语法分析中插入 action code 来实现抽象语法树(AST)的构建。

2.1 分词：正则表达式+Jflex

分词完成由源程序(字符串)到符号流的转化。一个符号(Token)是语言中表达最基本含义的单元。一个符号可以用一个正则表达式来表述，字符串和正则表达式的匹配则可以由确定状态自动机高效完成。对于在程序当中某一处有多个符号可以匹配的情况采用最长优先原则。

我通过描述正则表达式、设定不同的分词状态并加入 Java 代码处理转义字符，使用 Jflex 工具生成了分词代码。

```
118  commonCharacter = [[\x20-\xff]--['\"\\]]
119  translatedCharacter = \\[bfnrt\\'\"0]
120  asciiNumberCharacter = (\\x[0-9A-Fa-f][0-9A-Fa-f])|(\\[0-3][0-7][0-7])
121  //char = {commonCharacter} | {translatedCharacter} | {asciiNumberCharacter}
122
123  %state YYSTRING
124  %state YYCOMMENT
125
126  %%
127
128  <YYINITIAL> {
129      /* Pre-Compile Command */
130      {preCompileCommand}       { /* ignore */ }
131
132      /* Comments */
133      {singleLineComments}      { /* ignore */ }
134      "/*"            { yybegin(YYCOMMENT); }
135
136      /* Keywords */
137      "void"          { return symbol(VOID); }
138      "char"          { return symbol(CHAR); }
139      "int"           { return symbol(INT); }
```

图 1：loliccompiler.flex 片段

2.2 语法分析：上下文无关文法+JCup

分词产生的符号流输入到语法分析器中。语法分析器的作用是识别语法结构，并根据所识别的语法结构建立抽象语法树。这里也可以通过递归下降法来手写分析。但是为了支持一些较为复杂的语法模式和 typedef 所要采取的 lexer-hack，我使用了上下文无关文法来描述 C 语言的语法再 JCup 自动生成语法分析器代码的方法。通过在识别出特定语法模式时执行相应的 action code，可以建出抽象语法树。

```
224  declarator ::=      direct_declarator:dect
225                  {:  RESULT = dect;  :}
226                  |   MUL declarator:dect
227                  {:
228                      if (dect.type == null) {
229                          dect.type = new PointerType(null);
230                      } else {
231                          dect.type = dect.type.encore(new PointerType(null));
232                      }
233                      RESULT = dect;  :}
234                  ;
235
236  direct_declarator ::=   IDENTIFIER:name
237                      {:  RESULT = new VariableDecl(null, new Symbol(name), null); :}
238                      |   TYPENAME:name
239                      {:  table.addEntry(name, Symbols.IDENTIFIER);
240                          RESULT = new VariableDecl(null, new Symbol(name), null); :}
241                      |   PARAL declarator:dect PARAR
242                      {:  RESULT = dect; :}
243                      |   direct_declarator:dect PARAL {: table.addScope(); table.delScope(); :} PARAR
244                      {:
245                          if (dect.type == null) {
246                              RESULT = new FunctionDecl(new FunctionType(null, new DeclList()), dect.name,
247                          } else {
248                              dect.type = dect.type.encore(new FunctionType(null, new DeclList()));
249                              RESULT = dect;
250                          }
251                          :}
252                      |   direct_declarator:dect PARAL {: table.addScope(); :} parameter_list:para {: table
```

图 2：loliccompiler.cup 片段

2.3 抽象语法树设计

抽象语法树是 LoliCCompiler 中第一个对于源代码的等价表述。其主要目的是描述程序的抽象语法框架。我的抽象语法树设计分为以下 3 个主要模块和 2 个辅助模块：

declaration：函数、变量等的声明。9 个类。

statement：循环、分支、跳转语句。10 个类。

expression：表达式树。16 个类。

type：类型系统。12 个类。

initialization：初始化列表。3 个类。

它们之间的关系可以用这样一张图来描述。Program 是整棵 AST 的根。箭头表示了类型之间的为另一类型的属性的关系。

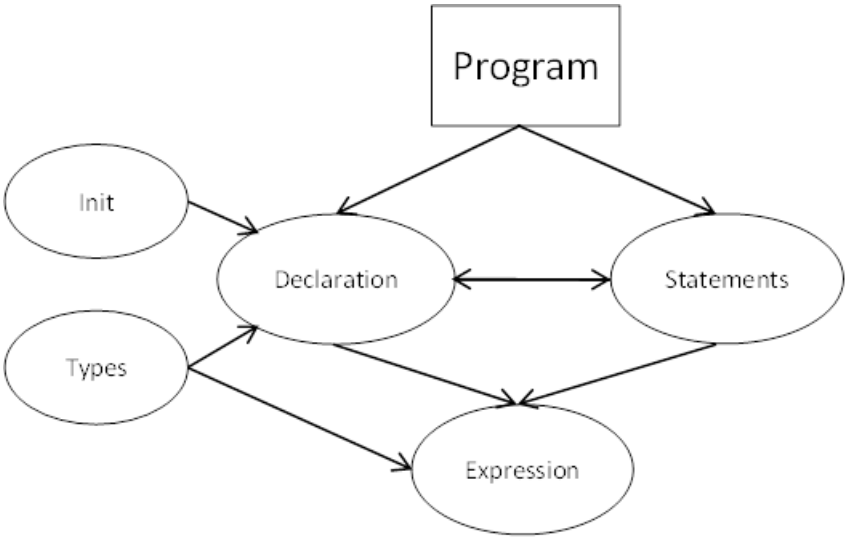

图 3：AST 类结构图

在遍历 AST 时我大量采用了访问者模式。通过不断地把控制权在具体类和访问者之前转换实现对于 AST 的遍历。这样做的好处是面向对象的 AST 的每个类可以只保留自己的属性定义，而把遍历的代码通过访问者集中起来进行管理共享环境。不过在遍历时传递参数会受到限制。

可以通过 GUI 上的 PrintAST 按钮查看生成的 AST 效果。

```
[Program]
    [FunctionDefi]
        [IntType]
        [main]
        [DeclList]
        [CompoundStat]
            [DeclList]
            [StatList]
                [ExpressionStat]
                    [FunctionCall]
                        [Variable]
                        [printf]
                        [ArgumentList]
                            [StringConst] [hello world!]
                [ReturnStat]
                    [IntConst] [0]
```

图 4：GUI 输出的 helloworld 程序的 AST

2.4 特性：复杂类型系统

LoliCCompiler 所支持的类型系统由五个部分组成：

基本类型：int、char、void。

指针类型：修饰某个具体类型。

数组类型：修饰某个具体类型，并具有固定的数组大小。

记录类型：由若干个具体类型按照固定顺序组合。

函数类型：由一个具体类型作为返回值，若干个具体类型按照顺序作为参数。（注：类型表示支持变参函数但仅用于内置 printf/scanf，语法并不能识别变参函数。）

其中除了基本类型以外的四种类型可以进行任意的嵌套组合。

这一特性的实现有两点：一是复杂类型的表示，二是复杂类型的识别。

复杂类型的表示通过面向对象的继承方法是容易设计的。而复杂类型的识别较为复杂，这是因为 C 语言的类型修饰之中指针是前缀修饰，数组和函数后缀修饰，而优先级却是以后缀修饰优先，并且函数定义的括号与改变类型修饰优先级的括号容易引发语法冲突。于是需要精妙设计相应的上下文无关文法加以解决。

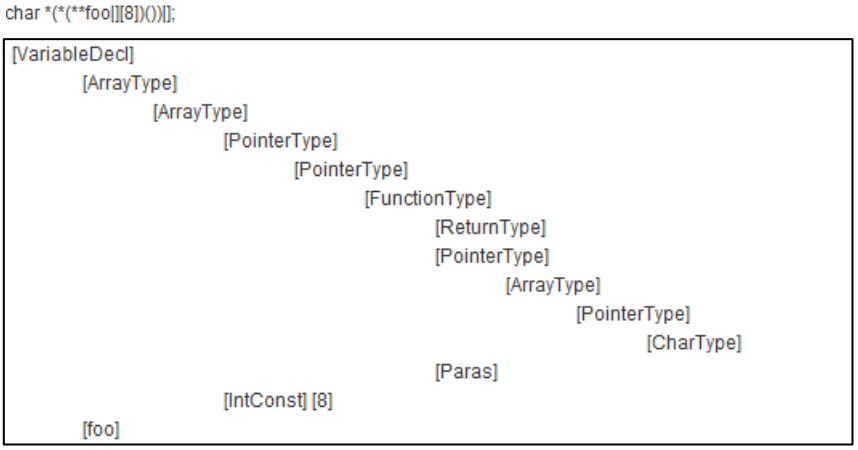

图 5：C 复杂类型定义和对应的 AST

2.5 特性：用户友好型 GUI

LoliCCompiler 相比其他甚至于企业级编译器而言的一大特色就是有着一个设计简洁且非常用户友好的 GUI。

左上是控制面板。有着包括打开/保存在内的各种功能按钮。可以输出各种中间表示。

左下是编译信息。出现语法或者语义错误时会在这里显示错误提示信息。

右侧则是输出文本框。对应的输出会显示在这大片区域中。

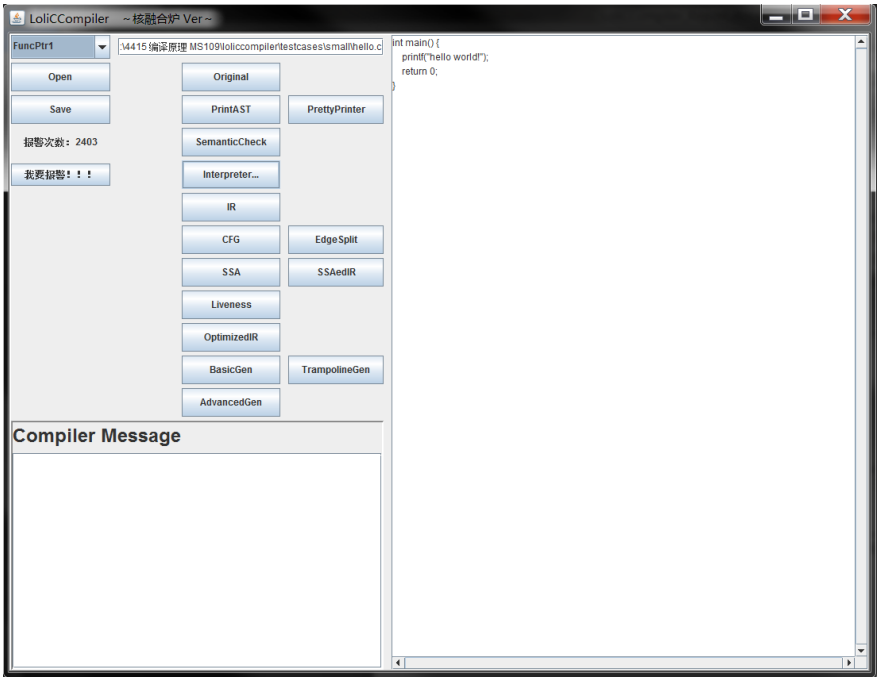

图 6：GUI 主界面

使用 java 的 GUI 库 java.swing，实现并不复杂。通过这一 GUI，可以详细地观察到 LoliCCompiler 执行的中间结果。

2.6 特性：代码美化工具

LoliCCompiler 实现了 Pretty Printer 也就是代码美化工具。其主要功能是将输入的源代码进行合理的缩进并插入/删除空格空行使其变得更为美观，便于阅读。

```
char *_ = "#include <stdio.h>%cchar* recurse=%c%s%c;%cint main(){printf(recurse,10,34,recurse,34,10,10);}%c"
int M[5000]={2},*u=M,N[5000],R=22,a[4],l[5]={0,-1,39-1,-1},m[4]={1,-39,-1,39},*b=N,
*d=N,c,e,f,g,i,j,k,s;int main(){for(M[i=39*R-1]=24;f|d>=b;){c=M[g=i];i=e;for(s=f=0;
s<4;s++)if((k=m[s]+g)>=0&&k<39*R&&l[s]!=k%39&&(!M[k]||!j&&c>=16!=M[k]>=16))a[f++
]=s;if(f){f=M[e=m[s=a[1/(1+2147483647/f)]]+g];if(j < f) j=f;f+=c&-16*!j;M[g]=
c|1<<s;M[*d++=e]=f|1<<(s+2)%4;}else if(d > b++) e=b[-1];}printf(" ");for(s=39;--s;printf("_")
)printf(" "); for(;printf("\n"),R--;printf("|"))for(e=39;e--;printf("%c", *("_ "+(*u++/8)%2)))printf("%c", *("| "+(*u/4)%2
                    ));}
```

```
char *_ = "#include <stdio.h>%cchar* recurse=%c%s%c;%cint main(){printf(recurse,10,34,recurse,34,10,10);}%c

int M[5000] = {2}, *u = M, N[5000], R = 22, a[4], l[5] = {0, -1, 39 - 1, -1}, m[4] = {1, -39, -1, 39}, *b = N, *d = N, c, e, f, g

int main() {
  for (M[i = 39 * R - 1] = 24; f | d >= b; ) {
    c = M[g = i];
    i = e;
    for (s = f = 0; s < 4; s++)
      if ((k = m[s] + g) >= 0 && k < 39 * R && l[s] != k % 39 && (!M[k] || !j && c >= 16 != M[k] >= 16))
        a[f++] = s;
    if (f) {
      f = M[e = m[s = a[1 / (1 + 2147483647 / f)]] + g];
      if (j < f)
        j = f;
      f += c & -16 * !j;
      M[g] = c | 1 <<= s;
      M[*d++ = e] = f | 1 <<= (s + 2) % 4;
    } else
      if (d > b++)
        e = b[-1];
  }
  printf(" ");
  for (s = 39; --s; printf("_"))
    printf(" ");
  for (; printf("\n"), R--; printf("|"))
    for (e = 39; e--; printf("%c", *("_ " + *u++ / 8 % 2)))
      printf("%c", *("| " + *u / 4 % 2));
}
```

图 7：代码美化前后效果对比

(使用的数据是 Phase2/Passed-new/madcalc.c)

实现上，使用了一个访问者 prettyprinter 来遍历整棵 AST，一边维护缩进信息一边输出到缓冲区中。而利用缓冲区处理类型的前后缀修饰等需要调整输出顺序等情况。

美中不足的是，由于是在 AST 进行的遍历，丢失了源代码当中的注释和括号等信息。LoliCCopmiler 是采用对于运算符优先级的运算来加上必要的括号来保证语义等价性的。

3 Semantic Check

完成抽象语法树的建立之后，需要对于语法树进行语义检查，这是 LoliCCompiler 的第二个模块。语义检查的目的主要有三点：一是检测出无法被编译的、具有语义错误的代码；二是

使得代码的表示变得更为抽象化，忽略语法结构细节，便于进一步的编译操作；三是对于变量大小等信息进行计算，为进一步的编译提供必要信息。我选择了通过将抽象语法树转化为中间表示树的方法来实现这三个目的。

3.1 中间表示树设计

中间表示树(IRT)是 LoliCCompiler 对于源程序的第二个等价描述。其主要目的是更加抽象地描述程序并聚合后面编译步骤所需要的信息。

我所设计的 IRT 仅有三个模块：声明、语句、表达式。其中声明只有一个类，即是变量声明，函数和类型的声明经过语义检查都已经不再需要，语句还是保留了程序的控制流，有循环、分支、跳转、表达式以及组合语句五个子类，而表达式也只有一个类，通过工厂模式对于运算符类进行封装，这使得在之后的中间代码生成之中，可以由各个运算符自己处理自己的翻译方法。

至于 AST 中的类型和初始化列表，经过语义检查，其中必要的信息也已经保留到了 IRT 各个类的属性之中。

IRT 的类图如下。Program 是 IRT 的根。可以看出相对 AST 已经有了很大的简化。

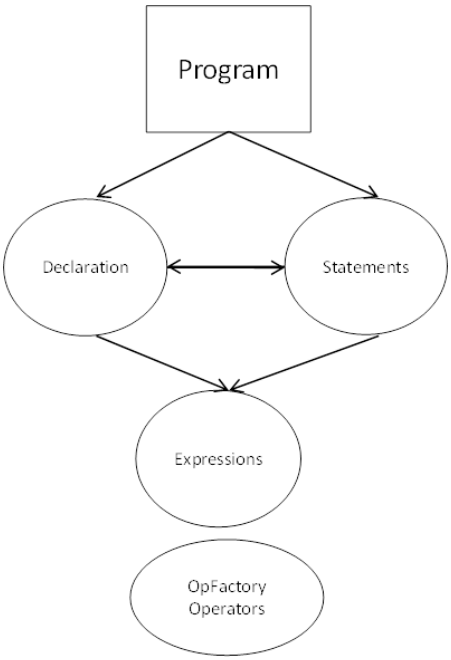

图 8：IRT 类结构图

3.2 语义检查实现

语义检查仍然是通过一个 visitor 来遍历 AST，并且在这个 visitor 当中维护变量、函数、类型等的环境信息来实现的。在语义检查的过程中利用一个栈来建立 IRT。

语义检查有很多的实现细节，这是为了支持繁多的语言特性。我觉得这并没有特别好的实现方法，还是要依靠对于各种语义的枚举亦或是对于不同情况的分类讨论。例如减法对于两个操作数分别是不是指针就有完全不同的操作，必须进行分别处理。而+=这类运算赋值运算符也不能简单地拆成+和=先运算再赋值。

我通过工厂模式对于运算符以及运算符的构造过程进行了封装，将所有的讨论都限定在了

语义检查的阶段，之后的阶段就不需要再枚举是什么运算符再来进行翻译了，这是因为这一数据信息已经被转化为了更加便于使用的结构信息，包含在了对象本身之中。

```
232     private OpFactory calPost(int op) {
233         switch (op) {
234             case Symbols.INC_OP : return Factories.POINC.getFact();
235             case Symbols.DEC_OP : return Factories.PODEC.getFact();
236             default : throw new InternalError("Unexpected Post operator.\n");
237         }
238     }
239
240     void define(int id, Type type) { define(id, type, VariableTable.VARIABLE); }
243
244     void define(int id, Type type, int isVari) {
245         if (table.checkCurId(id)) {
246             if (!isGlobal() || !typeEqual(type, table.getId(id)) || table.checkType(id) != isVari) {
247                 throw new SemanticError("Identifier " + id + " redeclared as a different kind of symbol.\n");
248             } else if (table.checkDefi(id)) {
249                 throw new SemanticError("Identifier " + id + " redefined.\n");
250             }
251         } else {
252             if (isVari == VariableTable.VARIABLE) {
253                 table.addVari(id, type);
254             } else {
255                 table.addType(id, type);
256             }
257         }
258         table.defiVari(id);
259     }
```

图 9：IRTBuilder.java 代码片段

3.3 特性：语法/语义错误信息

LoliCCompiler 有着完整的异常处理机制。其异常类 CompileError 继承 RuntimeException。有四个子类：

InternalError：编译器内部错误，属于 bug。

InterpretError：解释器运行时错误，包括除 0 等。

SemanticError：语义检查发现的语义错误。

SyntacticError：Lexer 或 Parser 发现的语法错误。

对于在程序中发现的语义或语法错误，LoliCCompilerGUI 可以捕捉到被抛出的异常，并显示在 CompilerMessage 区中。有数十种语法和语义错误提示，仿照了 GCC 的风格。

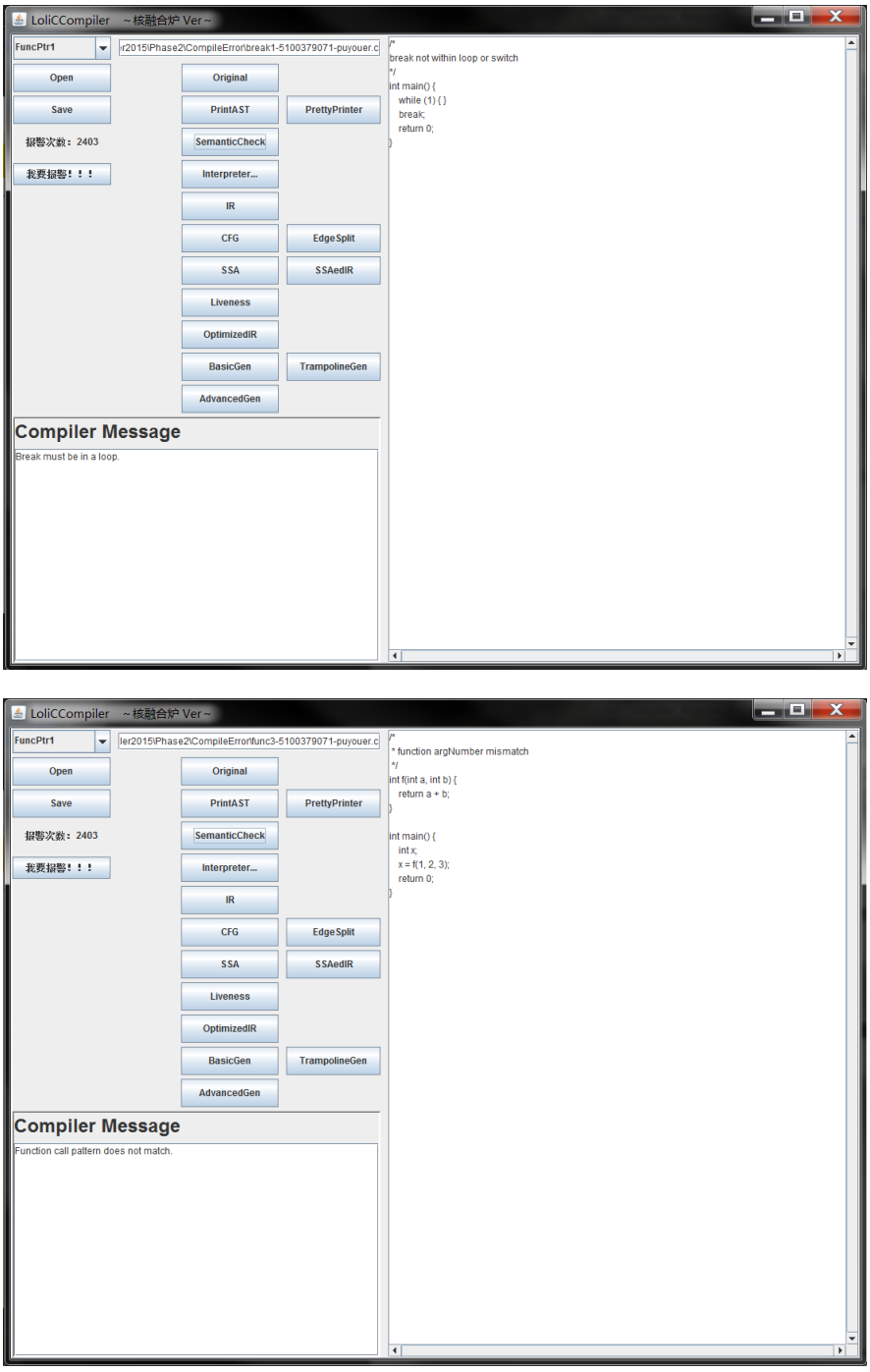

图 10、11 编译错误信息实例

3.4 特性：未定大小数组

在 C 语言中，未定大小数组共有三种：1、可以由初始化列表计算出数组大小。2、将数组作为一个函数的参数进行传递。3、对于 struct 的最后一个元素，如果是数组，可以是可变大小数组，其实际大小将由被分配到的内存大小决定。

LoliCCompiler 支持所有以上三种未定大小数组的使用方式。主要分为三个步骤：1、在语法分析阶段支持大小为空的数组。2、对于可由初始化列表计算得到的，在语义检查阶段进行计算。对于参数，将其看作一个相应类型的指针。3、在解释器中正确模拟以及在中间代码和

代码生成中正确实现对应的内存分配和地址访问。

3.5 特性：typedef 支持

typedef 是 C 语言所提供的一个语法的糖衣。通过 typedef 可以为复杂类型提供较为方便使用的别名。一个常见的做法就是将一个记录名 typedef 为对应的记录类型。

虽然在实际中很少会有人将 typedef 过的类型名再重新使用为变量名，但这对于编译器而言是必须要考虑的 corner case。

LoliCCompiler 对于 typedef 的实现主要是基于环境的 lexer hacking 和语义检查阶段的支持。

3.5.1 Lexer Hacking

分词器(Lexer)无法判断一个 Identifier 是否是被 typedef 过的 Typename。然而语法分析器(Paser)却依赖分词器所给出的符号来进行语法分析。于是，必须有某种方法让语法分析器“告诉”分词器现在某个特定的 Identifier 是否是被 typedef 过的类型名。这种反信息流而行的方法被称为 Lexer hacking。

LoliCCompiler 也实现了 Lexer hacking 的技术。在分词器和语法分析器之间搭建了一个环境，即一张符号表，分词器根据这一符号表的信息判断某一个 Identifier 是类型名还是标识符，而语法分析器会根据语法规则的语义添加删除名字空间，并将某一标识符判定为类型名加入到符号表中。小心地实现这一过程，可以解决类型名和标识符的混淆问题。

3.5.2 语义检查支持

经过语法分析，typedef 会被看成声明被放入 AST 中，而在语义检查阶段，建立了基于语义的符号表来对 DefinedType 进行解定义，还原回其原类型。因此 typedef 对解释器及后面的编译过程就不再产生影响了。

3.6 特性：C 语言解释器

LoliCCompiler 的一项特色功能是其内置的 C 语言的解释器 LoliCInterpreter。可以直接运行 C 语言源程序。其将源程序整个读入，在进行完语义检查之后得到的 IRT 上对其运行过程进行模拟，通过模拟控制台的 GUI 与用户进行交互。

LoliCInterpreter 本身是 IRT 的一个 visitor，由于对于表达式的计算封装在了运算符类之中，解释器主要实现控制流、堆栈存储和 STL。

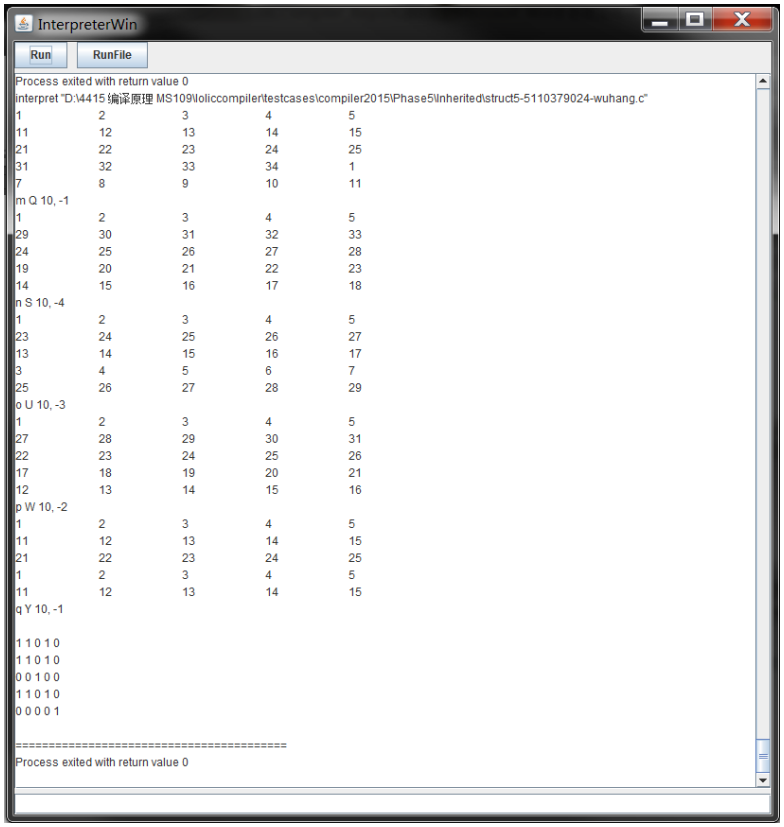

图 12：解释器 GUI
(图中显示的是 struct5.c 的运行结果，正确无误)

3.6.1 控制流实现

LoliCInterpreter 本身是 IRT 的 visitor，对于某一 IRT 的节点，让 Interpreter 对其进行访问即是进行对这一节点的解释。因此，函数调用的控制流恰好由系统的调用来实现，而分支和循环的实现也相对简单，对于 break 和 continue 则在 Interpreter 设置了一个中断信号，将不会进行解释，而特定语句能够改变信号并对其作出反应。

3.6.2 模拟堆栈

LoliCInterpreter 使用了一个大小为 4MB 的 byte 类型的数组来模拟实际运行时的堆栈。处理上有三个要点：

1、正确处理函数调用和返回。模拟了运行时环境的栈帧，在调用和返回时会进行帧的分配和删除。

2、栈空间和堆空间。栈和堆分别从 byte 数组的两端开始。malloc 将从末端开始分配空间。而栈则由栈顶指针从头端开始进行空间的分配。

3、地址和指针。取地址操作将会得到虚拟堆栈上的地址，而指针运算则按照实际在内存中那样进行，解释器能够正确实现包括函数指针在内的指针访问。

3.6.3 语言特性支持：未定大小数组/函数指针/STL(含 scanf)支持

typedef 的问题在语义检查阶段就已经解决了。而未定大小数组则仅剩下实际是对分配空间和寻址访问的支持，在模拟堆栈的基础上也是较为容易的。

对于函数指针，LoliCInterpreter 无法像实际的函数指针那样保存一个指令在内存中的地址，而是保存的所调用的函数的 id，当通过函数指针引发调用时，将在环境中寻找对应 id 的函数来进行访问。

对于 STL 的支持，LoliCInterpreter 用 Java 严格按照 C 的标准进实现了 printf/ getchar/ malloc/ scanf。其输入和输出都建立了一层缓冲区。malloc 移动并返回堆顶指针。getchar 取出缓冲区的下一个字符。printf 则根据格式描述符来依次输出每个参数。scanf 则从输入读入字符串并按照格式描述符进行解释放到作为参数的每个地址中。

而为了模拟实时响应的控制台，采用了 Java 多线程技术，在不同的线程中运行解释器和处理 IO 中断。解释器同样支持一些较为简单的运行时错误，会显示在模拟控制台中。

4 Intermediate Representation

将中间表示树转化为中间表示，即三地址代码，是 LoliCCompiler 的第三个模块。中间表示已经完全消除了源代码的语法特性的细节，复杂操作也被分解为可以被翻译为汇编代码的多个简单操作，开始接近可以实际运行的机器代码。而大部分的优化实质上则是从一种中间代码到同一种或另一种中间代码的映射，所以这一步骤对于生成高质量的机器代码而言是必要的。我在这一阶段除了使用三地址代码以外，还实现了静态单赋值(SSA)作为一种辅助中间表示。

4.1 中间代码设计和转化

4.2 特性：控制流优化

4.3 特性：函数指针

4.4 特性：内嵌函数和高阶函数

4.5 特性：静态单赋值(SSA)

4.5.1 控制流分析、控制流图、边切分性质

4.5.2 最近支配点计算

4.5.3 支配边界计算

4.5.4 插入 phi 函数

4.5.5 变量重命名

4.5.6 活性分析修剪

4.5.7 转换回中间代码

4.5.8 别名处理

5 Code Generation

LoliCCompiler 的最后一个模块是将中间表示转化为汇编代码。这一阶段所要解决的问题是代码选择、寄存器分配和指派。我除了实现了最为基本的不使用寄存器而依赖内存的暴力代码生成以外，还实现了较为复杂但效果卓越的跨过程全局寄存器指派和分配。

5.1 暴力代码生成

5.2 高效代码生成

5.2.1 寄存器合并

5.2.2 活性分析剪枝

5.2.3 STL 内联展开

5.2.4 别名处理

5.3 特性：全局寄存器指派和分配

5.3.1 循环分析

5.3.2 过程间数据流/控制流分析

5.3.3 符号寄存器代码选择

5.3.4 图染色寄存器分配

5.3.5 改进溢出处理

5.4 特性：scanf 支持

5.5 特性：高阶函数（续）

6 Optimizations

实现一些较为简单的优化算法是 LoliCCompiler 项目的一大目标。除了前面提到的全局寄存器分配外。LoliCCompiler 还实现了以下的优化。

6.1 特性：公共子表达式消除

6.2 特性：冗余复制消除(基于 SSA)

6.3 特性：死代码消除(基于 SSA)

7 Conclusion

LoliCCompiler 虽然只是在两个月的时间内由我一个人仓促完成的一个学生项目，和 gcc 等功能强大的现代编译器毫无可比性，但它仍然实现了完成一个 C 编译器的项目目标，从中可以看出很多现代编译器特性的影子。并且它在实现基本的编译功能之外大胆探索创新，支持了许多超过课程要求的独特特性，例如用户友好型 GUI、C 语言解释器 LoliCInterpreter 等等。对于我这么一个对于计算机并无太多了解的大二学生来说，LoliCCompiler 不失为一件值得称赞的十分优秀的作品。

References

[1] Alfred V. Aho, Monica S. Lam, Ravi Sethi, and Jeffrey D. Ullman Compilers: Principles, Techniques and Tools (Second Edition). Pearson Education, 2006.

[2] Appel, Andrew W. and Ginsburg, Maia. Modern Compiler Implementation in Java (Second Edition). Cambridge University Press, 1997.

[3] Steven S. Muchnick. Advanced Compiler Design and Implementation. Elsevier Science, 1997.

[4] ISO/IEC 9899:1999 - Programming languages - C (C99)

[5] Gerwin Klein, Steve Rowe, and Regis Decamps. JFlex User's Manual. 2015.

[6] Scott Hudson. JCup User's Manual. 2014.

[7] James R. Larus. Appendix A: Assemblers, Linkers, and the SPIM Simulator.

[8] Kathy Sierra, Bert Bates. Head First Java (Second Edition). O'Reilly Media, 2005.

[9] Lots of Authors. Static Single Assignment Book. Not published yet.

[10] GCC, the GNU Compiler Collection, online documentation.

[11] Ted Yin. CIBIC: C Implemented Bare and Ingenuous Compiler. 2014.

Remarks

最大工程 通宵 10 次左右 时间估计 350h 代码量估计 2.5 万行 收获 3 编译器 java/工程 个人管理 还有很多不足 语言特性匮乏 缺少汇编器/连接器 跨平台 今后的维护 full of buges 更多优化 跨平台集成 更多的学习

感谢课程助教 学长 同学