**中间表示的代码生成**

先说总体工作打算吧，先进行规划中间表示的具体数据结构，随后，针对每个语句节点进行代码生成工作。也就是说，这篇文档主要描述的是中间表示的规划设计以及由前端语法树生成中间表示的工作。

对于中间代码生成工作是否和语义分析放在一趟里面的问题。其实吧，理论上来说，如果能够混合在语义分析的遍历过程中是最好的，因为这会少很多代码。

但是问题一样存在，本来expr的语义分析上上下下2k行，如果再将中间代码生成放在里面，单个文件里面内容太多了。Declaration里面虽然东西多，但是真的需要考虑的无非是声明的变量的size而已，实际上中间代码生成并不会放进去太多内容，但是expr里面则不同。

目标是什么？

没必要完全对标LLVM，因为这是一个专门为了C编译器设计的框架，没必要兼容更多的东西，一些语法糖同样无关紧要。但是，中间表示的设计，需要尽可能的能让一些主流的编译优化的算法能够运行起来，而不出现一些因为中间表示设计的方案问题，引起的大规模的重构等等问题。

首先，因为源语言和目标平台是基本确定的（当然你非要说什么32位和64位的差别，我也没办法咯），所以完全可以针对性的，比如指令合法化（就是检测IR的指令是否能够直接翻译成目标平台的指令，不合法就用比如说多条目标平台指令来代替一条这样子的办法来解决掉）这样的pass本身可以在针对性的IR设计中去消除的，另外，由于设计目的是在后端使用gcc的asmer和linker，那么code emission这样的事情，由于本编译器的设计实现，直接就没有必要了。因此，虽然可以参考LLVM的后端设计，也可以留下接口，但是，没必要太苛刻的去完全“学习”。

其次，设计的可扩展性。因为确实要赶工期了。害。这么说吧，如果可以，那么最好在IR之后，直接就可以不经过任何优化，就翻译成目标平台代码，然后通过模块化的各种优化pass，当成插件一样，去进行优化。当然这是最最理想的情况了。

**中间表示的代码表**

如果说前端核心是抽象语法树，那前端到后端的核心部件就是中间表示。

但是到底用那些中间表示的代码，这是个问题。

先说整体框架，我的打算确实是类似于LLVM的，就是module到function到BB到instruction。但是图的表示也依然需要。什么Ud du链也都是需要的。

整体的中间表示的语句，我也会一定程度的参考的，但是并不完全等同，起码会进行一定的删减。

总之，对于AST到IR的翻译，我认为先构建起中间表示才是最重要的。

* 支持的各个操作的表示

中间表示的操作选择相当复杂，从理论和直观来说，由于目前我的编译器的目标平台是x86，所以完全可以针对x86的模型去做，但是中间表示的设计又要求其本身尽可能泛用性，很有可能多条IR指令对应一条x86或者一条IR需要多条x86去翻译。

1/load/store

2/int类型的运算，加减乘除

3/int类型的位运算

4/real类型的运算（没有位运算这种）

5/控制转移指令，跳转，包括无条件跳转和条件跳转，function call ret等等

6/内存分配

|  |  |  |  |
| --- | --- | --- | --- |
| 指令 | Dst | Src1 | Src2 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

* 中间表示的寄存器和基本类型

LLVM拥有无限多的寄存器，对于表达式的翻译，其中间结果，肯定是存放在寄存器中的。而上面的想法所不同的是，对于内存变量，LLVM同样将他们视作寄存器。而在后面的pass中进行内存布局的确定。换句话说，那些可以确定的变量是Static Alloca，而中间结果，则是Dynamic Alloca。

由于死变量消除等等问题，在中间表示不直接使用内存分配，而在经过优化之后再确定如何使用，我认为是非常合理的。否则会造成内存空间的浪费。所以就这样了。

这里面需要分配寄存器标号。

对于寄存器占用的大小，我反正直接用64bit和32bit两种寄存器，一方面，无论是long double 还是long long int都是64位表示的，这是最大的数据量了，另一方面，考虑的目标平台是x86-64，基本上64位和32位寄存器能够表达除了complex的绝大多数东西了，而且complex其实就是一个长度为2的数组，我也懒得说。就这么定了。

问，如何保证能够在目标平台上跑，因为目标平台只有一种长度可能，所以这里面依然需要合法化的步骤。如果64位跑到32位的，那么需要拆分计算，如果32位跑到64位的，那么需要扩展。

1. **存储绑定**

存储绑定也就是前面说的寄存器无限分配的设计方面。

一个变量需要给定其存储单元。

­­方法一：变量分为 全局变量，全局栈变量，栈变量，栈静态变量。

全局变量和静态存储类的变量分配固定可重定位地址，或者是一个相对于基址寄存器的偏移。

栈变量则分配为相对于栈帧的地址。

也就是说，给出当前栈帧的相对地址即可。

方法二：也就是上面说的LLVM所使用的方法

给他们分配名义的符号寄存器，这种名义寄存器本身是无限的，并且可以按照累加方式分配。不过一般不用于数组或者聚合类型。

使用这种方法需要配合后续的图着色寄存器分配算法等算法，在IR到目标指令的过程中指定寄存器。

另一个问题在于类型系统，LLVM存在更多的类型，但是就C这门语言的翻译来说，所要求的类型并没有那么夸张和复杂，~~我打算直接按照占用内存大小来区分。~~我对此的看法是，对于非聚合类型的变量可以用寄存器的方法，但是聚合类型的还是在栈上面分配吧。

1. 还有就是**边界对齐**。

边界对齐的方法，对于普通变量，既可以按照声明的顺序去分配他们。也可以先分配64位的变量，再分配32位的，以此类推。或者先分配最小的，再分配最大的。

对于较大的数据结构，比如说数组。一个问题是，立即寻址的长度位数是有限的，通常达不到32位（我要是没记错，MIPS是20位立即数好像），有可能数组的长度太长了。一种方法是，使用一个指针指向其位置，再使用加法来进行偏移寻址。

1. **寄存器使用的优化**

不再简单的使用的时候从内存取东西到寄存器，算好立即存储。

一种改进是追踪每个寄存器的内容，如果已经在合适类型的寄存器中，不再冗余的取它，而是直接使用，如果类型不符，但是在某个寄存器，那么可以进行类型传送，比如浮点到整数的传送指令。

如果用完了寄存器，可以选择一个新的（就是说，存储一个值到内存，然后空出来一个寄存器，就可以了。

不是算好一个值就立即存储，而是推迟到基本块结束时或者寄存器用完的时候再存储。

对于跨基本块的寄存器分配，如果只有一个前驱块，那么前驱的出口寄存器状态就是后继的入口状态。多个前驱可以取交集。

更多的可以考虑全局寄存器分配算法。

1. **中间表示的形式以及我打算采用的形式**

四元式：不想解释

三元式：相比于四元式，结果并不显式的给出。而是使用对于该三元式的标号来使用其结果

树：不想解释，由四元式生成树应当是很简单很基本的事情。

DAG：在树中，假设同一个变量被使用，会使用两个叶子结点。但是在DAG中，则是两个引用指针即可。通常来说，这被用于基本块之内。

前缀波兰表示：也就是树的遍历形式而已，其实质并没有太大的差别。

关于我个人对于中间表示想要采用的形式：

首先，针对于源语言翻译成中间表示来说，肯定是下面描述的这种线性结构更为合适的。但是在这一步之后，我认为中间表示应当在后面的步骤中被转换为DAG形式。这也是LLVM所实现的。

* 中间表示的module function BB instruction结构

本来就这四个层次的结构没什么好说的，但是有那么几点，所以拿来说

首先是BB，这里面需要构建每个BB（其实也有function来做的）的DAG，这毫无疑问。而经过指令选择之后，需要将BB（或function）下的ins链转化为DAG图，这个数据结构如何处理？也就是说，instruction结构如何能在表示指令链条的同时表示DAG图。我个人其实比较喜欢function来做DAG。因为这样可以有全局的指令优化空间。问，这样一来BB这部分怎么处理呢？

其次是，IR的表示，和经过指令选择之后，实现的目标平台代码之间的兼容性，因为我才懒得再去做一个数据结构来表示MachineInst和MCInst呢，自找麻烦是不是，那么这样一个兼容从IR到目标平台代码的数据结构和表示，设计起来，也不算简单。

其实就我的看法，说到底，无非这两样，

1指令类型（以及对应的操作数的数量，0操作数，1操作数，2操作数，3操作数）这个四元组最经典的东西，虽然我不会用就是了。

2，指令的操作数从哪里来，分别有立即数（也就是常量了），寄存器，内存位置，指令位置（比如跳转之类的）这四种。

其中，内存位置这个，只有load store这两个操作可以使用，并且其操作数数量是相同且确定的，因此，尽管x86的寻址方式复杂，却可以在IR生成的时候就写在里面，并且能够封装好。

立即数的常量，那没啥说的，直接写就好了。我在语义分析折腾了半天的常量表达式，难道是用来看的嘛（但是仍然要考虑常量折叠）

指令位置，也就是一般用于跳转到某个标号那里这种东西呢，一定存在于转移指令（包括无条件转移和条件转移指令对吧。

其他地方都应当而且必须是寄存器。在这里，由于铁了心要弄ssa，所以寄存器（因为无限数量嘛），只会被赋值一次，所以，也其实没那么麻烦就是了。

* 目标平台寄存器和可选指令集

这里面因为平台的32和64位的不同，所以依然要实现指令选择。（叹气）

首先，无论如何都要感谢不知道叫什么的老哥，对x86-64指令集的整理。我把链接贴出来哈：

<http://ref.x86asm.net/#column_op>

这里有较为详细的x86 32位及64位指令集的汇总整理——总不能再为了这个项目去翻intel手册吧。

当然，无论如何我都会为32和64位的代码表使用机器而不是人工方式来进行整理——2k多个代码，人工整理，那我岂不是智障。

* IEEE-754-2008

在中间表示的类型中，对于浮点数的类型表示，参照IEEE 754 2008的描述进行。但是相比于普通学到的754标准，确实有点东西哈，我简单描述一下，网上参考资料也不少，反正我是直接找了一个参考资料的，懒得详细说了。

https://pan.baidu.com/s/10soDctgCJ84MDs3PyhJBcw?pwd=lzku

* X86 amx和mmx类型

既然是要做x86，而且没说不考虑向量矩阵运算，这两个类型也是需要考虑的，当然如果是其他目标架构，理论上可以就不管它。

* 函数调用

函数调用，我认为需要提一下的，一个是plt和got，这在对应的md文档中写了，举的是linux下plt和got的例子。另一个则是动态链和静态链，这个概念本身是相对于前端语言转换成中间语言的过程的，因此放在这边说。

动态链是直接调用当前活动栈帧的一个调用链，指向上一个调用者，我认为没什么好说的，看过C函数调用的都明白有一个指向上一个栈帧的指针。而静态链，其目的是用于解析非局部符号表。

先说说前提，首先静态链本身针对的是能够在函数中嵌套函数定义的语言，C显然不是这样一种语言，因此，如果只是想看C编译器怎么搞，这部分别管了。

静态链指向的是其直接外层调用者，也就是说，自递归的静态链并不是指向自己，而是指向自己外面那一层。因为其目的是找非局部符号，在这样一个函数嵌套定义的过程中，上层函数的符号，在下一层函数中也是可以使用的，因此，才有了这样一个静态链。

在相同的嵌套层次，那么从上一层的静态链复制，如果被调用过程高n层，那么底层的局部变量是不能被看见的，所以需要回溯n层来寻找。

嵌套层次显示表display表，我看了一下就是对每一层的嵌套都指出其最新的函数的静态链（最新的含义是，类似于树的深搜，而每一层都有一个当前路径的节点这样）