Intermediate Code Generation

Transforming the abstract syntax tree (AST) into intermediate representation is called lowering.

We will modify the stack/register machine that we saw already by adding types to it.

The resulting intermediate representation is similar to LLVM (Low Level Virtual Machine), but not completely the same. In the end, we decided not to base the complete course on LLVM, because some of its features are ugly. It is close enough to our representation, so you will have no problems understanding it.

The big advantage of teaching LLVM is that it is used by real compilers.

Requirements on Intermediate Representation

Most compilers translate different languages into different machines.

In order to avoid quadratic explosion, one should use a single common intermediate representation.

The intermediate representation must be far enough from the languages to be suitable for all of them, and must be also far enough from the machines to enable translation into all of them.

The representation should have a simple type system, and be suitable for optimizations.

Simple/Non-simple Types

We distinguish between simple and non-simple types.

A simple type is a type that can be bytewise copied, and whose size is known at compile time.

Primitive types are always simple.

Classes that have non-trivial copy-constructor/assignment/destructor or that can inherited from, are non-simple.

Simple is also called POD (plain old data).

Simple types can be put in registers if they are not too big. We assume that all primitive types fit into registers.

In CLANG, a complex number fits into a register, while a quaternion does not any more.

Typed Register Machine

Before we can discuss any form of typed translation, we need to develop a typed version of the stack/register machine.

We add types to register assignments, so that they now have form:

$$\begin{aligned} \alpha_1\text{: } & \textbf{int} & \leftarrow \quad & 3 \\ \alpha_2\text{: } & \textbf{real} & \leftarrow \quad & 2.7182818 \\ \alpha_3\text{: } & \textbf{real} & \leftarrow \quad & \textbf{conv } \alpha_1 \\ \alpha_4\text{: } & \textbf{real} & \leftarrow \quad & \textbf{sum } \alpha_2\ \alpha_3 \end{aligned}$$

Note that, even though we attached the types to the registers, we are still typing the operations (as would be in assembly language). This is because every operation works only on registers, and creates a register.

Explicit Addresses, More Complicated Calculations

In the untyped register machine, load and store had the following forms:

$$\alpha_1 \leftarrow \text{load } \#1$$

$$\text{store } \alpha_1 \ \#2$$

Because C has pointers, which can be calculated at run time, the address of the load/store is not always fixed in advance. We will use # to create a pointer to the local variable, and let load/store use the pointer. After the #, we list the types of the variables being skipped:

$$\alpha_1\text{: pntr(int)} \leftarrow \#\text{char}$$

$$\alpha_2\text{: int} \leftarrow \text{load } \alpha_1$$

$$\alpha_3\text{: pntr(int)} \leftarrow \#\text{char, int}$$

$$\text{store } \alpha_2 \ \alpha_3$$

Real Hardware is Complicated

On real hardware, the address calculations are more complicated, because types have alignment requirements. On most hardware, **int** and **double** must start at an address that is a multiple of 4. This implies that if the first local variable is **char**, 3 bytes will be wasted.

This also applies to fields in structs.

We ignore these problems, because in every meaningful program, the sequence of local variables is always fixed at each point in the program. Hence, the exact address can be computed, if one wishes to produce real assembly. Also, sizes are hardware dependent, so we don't want to include them.

Alloc and Dealloc

We need to add a type to the alloc statement. Otherwise, it would be impossible to translate to hardware.

We also add types to dealloc,

alloc **double**

alloc **int**

alloc **char**

...

dealloc **char**, **int**, **double**.

Sequence of Local Variables is Always Fixed

We require that, when two possible computation paths merge, the types of the allocated variables on these paths are the same.

This means that the following situation is not allowed:

$$
\begin{array}{ll}
 & \text{iffalse } \alpha \text{ goto } L_1 \\
 & \text{alloc } \textbf{double} \\
 & \text{goto } L_2 \\
L_1: & \text{alloc } \textbf{int} \\
L_2: & \alpha{:}?? \leftarrow \# \\
 & \text{dealloc } ?
\end{array}
$$

At L_2, we don't know the type of #, and we don't know how much space to deallocate.

Memcopy

We assumed that registers are limited in size. They can only hold built-in types and pointers.

In order to assign **struct**s and pass them to functions, we need a **memcopy** statement. It has form **memcopy** α_1 α_2, where α_1 α_2 must have form **pntr**(T). It copies from left to right.

On the following slides, we give some examples. The function call/return mechanism is not changed.

The translation function will be a bit more complicated (but not much) because of possible **struct**s that don't fit into a register.

Consider the `fact` function again, this time properly typed:

```
double fact( unsigned n )
{
   double res = 1.0;
   while( n != 0 )
   {
      res = res * n;
      n = n - 1;
   }
   return res;
}
```

It starts with

unsigned	**double**
n	(uninitialized)

Translation of `fact`

Create local variable `res` and initialize it with 1.0 :

alloc **double**

α_0: **pntr**(**double**) $\leftarrow$ #

α_1: **double** $\leftarrow$ 1.0

store α_1 α_0

The condition: `while( n != 0 )`

loop : α_2: **pntr**(**unsigned**) $\leftarrow$ #**double**

α_3: **unsigned** $\leftarrow$ load α_2

α_4: **unsigned** $\leftarrow 0$

α_5: **bool** $\leftarrow$ ne α_3 α_4

iffalse α_5 goto exit

Statement `res = res * n`

α_6: **pntr**(**double**) $\leftarrow$ #

α_7: **double** $\leftarrow$ load α_6

α_8: **pntr**(**unsigned**) $\leftarrow$ #**double**

α_9: **unsigned** $\leftarrow$ load α_8

α_{10}: **double** $\leftarrow$ conv α_9

α_{11}: **double** $\leftarrow$ mult α_7 α_{10}

α_{12}: **pntr**(**double**) $\leftarrow$ #

store α_{11} α_{12}

Statement `n = n - 1`

α_{13}: **pntr**(**unsigned**) $\leftarrow$ #**double**

α_{14}: **unsigned** $\leftarrow$ load α_{13}

α_{15}: **unsigned** $\leftarrow$ 1

α_{16}: **unsigned** $\leftarrow$ minus α_{14} α_{15}

α_{17}: **pntr**(**unsigned**) $\leftarrow$ #**double**

store α_{16} α_{17}

goto loop

Copy local variable `res` into the result, and clean up `res` and `n`:

exit : α_{18}: **pntr**(**double**) $\leftarrow$ #

α_{19}: **double** $\leftarrow$ load α_{18}

dealloc 2

α_{20}: **pntr**(**double**) $\leftarrow$ #

store α_{19} α_{20}

return

strlen and lstlen

We study two more examples, which are given on this and the next slide. The aim is to explain how pointer arithmetic works.

```
unsigned int strlen( char* p )
{
   unsigned int len = 0;
   while( *p )
   {
      ++ p;                 // pointer arithmetic
      ++ len;
   }
   return len;
}
```

lstlen

```
struct list { int first;   struct list* next; };

unsigned int listlen( struct list* p )
{
   unsigned int len = 0;
   while(p)
   {
      p = ( p -> next );   // pointer arithmetic
      ++ len;
   }
   return len;
}
```

The while condition `while( *p )` is not difficult to translate, it contains an implicit conversion to **bool**:

For simplicity, we assume that counting of registers starts at 0.

α_0: **pntr**(**pntr**(**char**)) $\leftarrow$ #**unsigned**

α_1: **pntr**(**char**) $\leftarrow$ load α_0

α_2: **char** $\leftarrow$ load α_1

α_3: **bool** $\leftarrow$ conv α_2

iffalse α_3 goto exit

Translation of Control Statements

Let's start doing systematic translations.

Statements of form **if** B **then** S_1 **else** S_2 are translated as follows:

$$
\begin{array}{ll}
 & \alpha\text{: } \mathbf{bool} \leftarrow \cdots \\
 & \text{iffalse } \alpha \text{ goto } L_1 \\
 & \text{translation of } S_1 \\
 & \text{goto } L_2 \\
L_1 : & \text{translation of } S_2 \\
L_2 : &
\end{array}
$$

Statements of form **while** B **do** S are translated into:

L_1 α: **bool** $\leftarrow \cdots$

iffalse α goto L_2

translation of S

goto L_1

L_2 :

Understanding Pointers

1. A pointer can point to an array element. In that case, one can add integers to it, which will move the pointer to another position in the same array. This type of addition does not change the type of the pointer. As a consequence, array pointers can be subtracted.
2. A pointer can point to a struct. In that case, one can add an unsigned int to it, which will move the pointer into a field. This addition will change the type of the pointer.
3. A pointer can point to an array. In that case, one can add an unsigned to it, which will move the pointer to an array element. This addition will change the type of the pointer.

Due to implicit conversions, (3) never occurs in *C*, but it exists in reality, and other languages may have it.

I believe that the intended type of pointer addition, can always derived from the types. Additions of type 2,3, will always change the type, while addition of type 1 does not.

Hence, it is possible to use the name **sum** for all of them. We will also allow **diff** on type 1 pointers.

LLVM, which we will see later, uses another approach: It defines a ternary (or higher) operator `getelementpntr` (`gep`), where the first operation is always of type 1, and later arguments are type 2,3 additions. If you want only a 2 or 3, you have to insert 0 as first argument.

Statement `p ++`

α_0: **pntr**(**pntr**(**char**)) $\leftarrow$ #**unsigned**

α_1: **pntr**(**char**) $\leftarrow$ load α_0

α_2: **unsigned** $\leftarrow 1$

α_3: **pntr**(**char**) $\leftarrow$ sum α_1 α_2

α_4: **pntr**(**pntr**(**char**)) $\leftarrow$ #**unsigned**

store α_3 α_4

The sum in α_3 is a type 1 addition. Because of this, α_1 and α_3 have the same type.

Note that in concrete hardware, type 1 addition must take the length of the data into account.

Translation of `p = ( p -> next )`

α_0: **pntr**(**pntr**(**list**)) $\leftarrow$ #**unsigned**

α_1: **pntr**(**list**) $\leftarrow$ load α_0

α_2: **unsigned** $\leftarrow$ 1

α_3: **pntr**(**pntr**(**list**)) $\leftarrow$ sum α_1 α_2

α_4: **pntr**(**list**) $\leftarrow$ load α_3

α_5: **pntr**(**pntr**(**list**)) $\leftarrow$ #**unsigned**

store α_4 α_5

The addition in α_3 is a type 2 addition. It changed the type of α_1, and α_1 is a pointer to a struct type.

Two Translation Functions

We are ready to translate C-expressions. Since we have to deal with non-simple **struct** types that do not fit in a register, we need two versions.

The first version **translate**(AST t) returns a register, and emits code that writes the value of t into this register as before. The second version **memtranslate**(AST t) emits code that pushes the value of t on the memory stack, and returns nothing.

In principle **memtranslate**(t) would be sufficient to obtain a complete translation function, but it would be very inefficient on simple operations on simple data, because of too many **load**s and **store**s.

Treatment of Lvalues

Lvalues are in fact pointers, so we define the following function:

Let t be an AST. If $t.\text{lr} = \textbf{lval}$, then $t.\text{truetype} = \textbf{pntr}(t.\text{type})$.
Otherwise, $t.\text{truetype} = t.\text{type}$.

Beginning is always hard

Assume that we want to translate expression t.

Let $T = t$.truetype. If T is **void**, nothing needs to be done. If T is simple, we can call **translate**(t) and ignore the result.

Otherwise, call **memtranslate**(t) and emit **dealloc** T.

If t is the condition of an **if** or **while** statement it must be the case that t.truetype = **bool**.

We assume that **return** t has been replaced by an assignment to the return variable.

Translation Function

Translation into a register is not really different from the untyped translation function, but C-operators need to be replaced by their definitions in the process.

Function **translate**(AST t) returns a register, and emits code that writes the value of t into this register.

- If t is a constant c, then let T be the type of c. Invent a new register name α. Emit α: $T \leftarrow c$ and return α.
- If t is a variable v, then let T be the type with which v was declared. Invent a new register name α. Let $T_1, \ldots, T_n$ be the types of the variables that need to be skipped in order to access v on the stack. Emit: α: **pointer**$(T) \leftarrow \#T_1, \ldots, T_n$ and return α.

if the AST has form $f(t_1, \ldots, t_n)$, with f a user defined function, we proceed as follows:

First call **memtranslate**($f(t_1, \ldots, t_n)$). This emits code that pushes the value of $f(t_1, \ldots, t_n)$ on the stack.

But we need it in a register. Invent two register names α, β, and emit

$$\beta\text{: } \mathbf{pntr}(T) \leftarrow \#$$

$$\alpha\text{: } T \leftarrow \text{load } \beta$$

$$\text{dealloc } T$$

Now we can return α.

If t has form $\mathbf{op}(t_1, \ldots, t_n)$ with **op** a primitive operator (i.e. part of the language), s.t. each T_i fits into memory, we proceed as follows:

First, recursively call

$$\alpha_1 = \mathbf{translate}(t_1), \ldots, \alpha_n = \mathbf{translate}(t_n).$$

For each t_i, define $T_i = t_i.\text{truetype}$.

On the slides slides that follow, we explain how **translate** continues for the different primitive operators **op**.

Emitting Code for Built-In Functions

- If **op** = load, $n = 1$, then T_1 must have form **pntr**(T). Invent a new register name α, and emit

$$\alpha\text{: } T \leftarrow \text{load } \alpha_1.$$

  After that, return α.

- If **op** = sum, $n = 2$, and T_1, T_2 are types on which addition is defined, then invent a new register name α. Let T be the return type of the addition. Emit

$$\alpha\text{: } T \leftarrow \text{sum } \alpha_1 \ \alpha_2$$

  and return α.

- The case for **op** = diff is similar.

Definitions of Built-In Functions

- If **op** = **conv** and $n = 1$, let T be the type of the result. If $T = T_1$, then don't emit anything and return α_1.

  If $T \neq T_1$, then invent a new register name α. Emit

  $$\alpha\colon T \ \leftarrow \textbf{conv}\ \alpha_1,$$

  and return α.

  A conversion between identical types emits no code, and returns the register of the original value.

- If $\mathbf{op} = \text{xpp},\ n = 1$, then we know that T_1 has form $\mathbf{pntr}(U)$, and the return type is U. Invent register names α, β_1, β_2, and emit

$$\alpha\text{: } U \leftarrow \text{load } \alpha_1$$
$$\beta_1\text{: } U \leftarrow 1$$
$$\beta_2\text{: } U \leftarrow \text{sum } \alpha\ \beta_1$$
$$\text{store } \beta_2\ \alpha_1$$

  After that, return α.

- The definition for ppx is similar, but we return β_2 instead of α.

- The definitions for xmm and mmx are similar, but we use diff instead of sum.

Definitions of Assignment Operators

- If **op** = assign, $n = 2$, then T_1 must have form **pntr**(T_2).

  Emit

  $$\text{store } \alpha_2 \ \alpha_1,$$

  and return α_1.

Definitions of Assignment Operators

- If op = **iadd**, $n = 2$, then T_1 must have form **pntr**(T_2).

  Invent two register names β_1 and β_2. Emit

  $$\beta_1\text{: } T_1 \leftarrow \text{load } \alpha_1$$
  $$\beta_2\text{: } T_2 \leftarrow \text{sum } \beta_1\ \alpha_2$$
  $$\text{store } \beta_2\ \alpha_1$$

  and return α_1.

- The definitions of the other modifying assignment operators are similar.

Lvalue Field Selection

We give the definition for lvalue field selection. Assume that t has form $\mathbf{select}_f(t_1)$.

Let $T_1 = t_1$.truetype. Since we are assuming that t_1.lr = **lval**, T_1 must have form $\mathbf{pntr}(U)$, where U is the type of the struct. Let i be the index of field f in U. Let T be its type in U.

First call $\alpha_1 = \text{translate}(t_1)$.

Invent a new register name α, and emit

$$\beta\text{: } \mathbf{unsigned} \leftarrow i$$

$$\alpha\text{: } \mathbf{pntr}(T) \leftarrow \text{sum } \alpha_1 \ \beta$$

After that, return α.

Field selection adds the offset of the field to the pointer α_1. The addition is of type 2. (See slide 22)

Assignment of a Non-Simple Value

We consider the case where t has form assign(t_1, t_2), and the type of t_2 is not simple. Because t_1.lr = **lval**, the result will be still simple.

We have t_1.truetype = **pntr**(T_2), and T_1 is also the type of t.

First call **memtranslate**(t_2). This emits code that creates a value of type T_2 on the memory stack.

After that, call $\alpha =$ **translate**(t_1). This is possible, because pointer types are simple.

Create a new register name α_1, and emit

$$\alpha_1\text{: } \mathbf{pntr}(T_2) \leftarrow \#$$

$$\text{memcopy } \alpha_1 \ \alpha$$

$$\text{dealloc } T_2$$

After that, return α.

Rvalue field selection

Assume that t has form $\mathbf{select}_f(t_1)$, and $t_1.\text{lr} = \mathbf{rval}$. In that case, also $t.\text{lr} = \mathbf{lval}$. Since we are translating into a register, we assume that $t.\text{type}$ is simple. Since t_1 is a struct, its type T_1 is not simple.

Let i be the relative position of field f in T_1, and let T be its type.

First call $\mathbf{memtranslate}(t_1)$. This creates code that pushes the value of t_1 on the stack. Create new register names $\alpha, \alpha_1, \alpha_2, \alpha_3$ and emit:

$$\alpha_1\text{: } \mathbf{pntr}(T_1) \leftarrow \#$$

$$\alpha_2\text{: } \mathbf{unsigned} \leftarrow i$$

$$\alpha_3\text{: } \mathbf{pntr}(T) \leftarrow \text{sum } \alpha_1\ \alpha_2$$

$$\alpha\text{: } T \leftarrow \text{load } \alpha_3$$

$$\text{dealloc } T_1$$

After that, return α.

Translation into Memory

We define **memtranslate**(AST t).

It emits code that pushes the value of t on the stack.

Let $T = t$.truetype.

If t has form $f(t_1, \ldots, t_n)$ with f a user defined function, then emit

$$\text{alloc } T.$$

After that, for $i := n$ to 1, call **memtranslate**(t_i). This creates code that pushes the values of t_i on the stack.

Finally, emit

$$\text{call } f.$$

Translation into Memory (2)

Assume that t has form $\mathbf{select}_f(t_1)$, with $t.\text{lr} = \mathbf{rval}$, and $t.\text{type}$ is non-simple. Let $T = t.\text{truetype}$. Let i be the index of field f in T_1

Emit

$$\text{alloc } T.$$

Call $\mathbf{memtranslate}(t_1)$. This emits code that creates a value of type T_1 on the memory stack. Invent register names $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, and emit

α_1: $\mathbf{pntr}(T_1) \leftarrow \#$

α_2: $\mathbf{unsigned} \leftarrow i$

α_3: $\mathbf{pntr}(T) \leftarrow \text{sum } \alpha_1 \ \alpha_2$

α_4: $\mathbf{pntr}(T) \leftarrow \#T_1$

memcopy α_3 α_4

dealloc T_1

Translation into Memory (3)

If t has form $\mathbf{load}(t_1)$, with $t.\text{type} = t_1.\text{type}$ a non-simple type T, $t.\text{lr} = \mathbf{rval}$, and $t_1.\text{lr} = \mathbf{lval}$, then call $\alpha = \text{translate}(t_1)$.

Create a new register β and emit

$$\begin{aligned} &\text{alloc } T \\ &\beta\text{: } \mathbf{pntr}(T) \leftarrow \# \\ &\text{memcopy } \alpha\ \beta \end{aligned}$$

Now we have made a copy of the data that α points to.

Translation into Memory (4)

If the previous cases cannot be applied, then t.truetype must be simple. We call **translate** and store the result in memory.

t must have form $\mathbf{op}(t_1, \ldots, t_n)$ with **op** a simple operator.

Call $\alpha = \mathbf{translate}(t)$.

Create a new register β and emit:

alloc T

β: $\mathbf{pntr}(T) \leftarrow \#$

store α β

Translation of Boolean ? Expression1 : Expression2

Assume that the AST has form $t = ?(t_1, t_2, t_3)$. We have t_1.truetype = **bool**, and t.truetype = t_1.truetype = t_2.truetype. Define $T = t$.truetype.

Translation of $t_1 ? \; t_2 : t_3$

We first define **translate** : Let $\alpha =$ **translate**(t_1). Create two new labels L_1, L_2, and emit

$$\text{iffalse } \alpha \text{ goto } L_1.$$

Call $\alpha_1 =$ **translate**(t_2). Emit

$$\text{goto } L_2$$
$$L_1 :$$

Call $\alpha_2 =$ **translate**(t_3). Emit

$$\alpha_1 \colon T \leftarrow \alpha_2$$
$$L_2 :$$

Return α_1.

Translation of $t_1 ?\ t_2 : t_3$

We define **memtranslate**($?(t_1, t_2, t_3)$) :

Let $\alpha =$ **translate**(t_1). Create two new labels L_1, L_2. Emit

$$\text{iffalse } \alpha \text{ goto } L_1$$

Call $\alpha_1 =$ **memtranslate**(t_2). Emit

$$\text{goto } L_2$$

$$L_1 :$$

Call **memtranslate**(t_3). Emit

$$L_2 :$$

Return.

Examples

This is a possible implementation of **strcpy**:

```
void strcopy( char* p, const char* q )
{
   size_t i = 0;
   while( q[i] )
   {
      p[i] = q[i];
      i ++ ;
   }
   p[i] = 0;
}
```

Length of a C-string

```
size_t length( char* p )
{
   size_t length = 0;
   while( *( p++ ))
      ++ length;
   return length;
}
```

Example with Twodimensional Arrays

```
void fill( p[][10], unsigned int n )
{
   for( unsigned int i = 0; i < n; ++ i )
      for( unsigned int j = 0; j < 10; ++ j )
         p[i][j] = i * j;
}

main( )
{
   int squares[5][10];
   fill( squares, 5 );
   return 0;
}
```

C^{++}

We have shown how to translate expression AST for C into intermediate language.

If one wants to translate C^{++}, then one needs to make the following modifications:

- **class** types can have overloaded copy constructors and assignment operators. For such types, **memcopy** cannot be used. Instead on needs to call the user defined constructors and assignment operator.
- In addition, **class** types can have destructors. When a type has a destructor, it cannot be simply deallocated. One first has to call the destructor.

Since constructors, assignment and destructors are user defined, one has to follow the usual calling convention.

C^{++} : Explicit References

In C^{++}, the lr-value distinction was replaced by the reference/value distinction.

This means that **ref** becomes a type constructor, similar to **pntr**.

References complicate things a bit, but not much.

The real problem comes on the next slides.

C^{++} : Presence of Temporaries

When a value is returned but a reference is needed, C^{++} creates a temporary:

```
std::ostream& operator << ( std::ostream& , const X& x );
X( int x ); // constructor
std::cout << X(3) << "\n";
```

The C^{++} standard guarantees that temporaries exist until the expression is completely evaluated. This is needed for example in:

```
const X& max( const X& x1, const X& x2 )
   { return either x1 or x2; }


std::cout << max( X(i), max( X(j), X(k) ) ) << "\n";
```

There are three temporaries. Since we don't know in advance which one will be printed, all three of them have to exist to the end.

Presence of Temporaries (2)

In order to add to the pleasure, temporaries can have destructors, and may be conditionally created:

```
X x(1)
std::cout << ( randombool ? x : X(2) ) << "\n";
```

Here is a sketch of the modifications needed:

1. During semantic analysis, create a new conversion, which converses T to **constref**(T).
2. Before translation, create, for each conversion to a temporary, a short-lived stack variable. If the type has a destructor, then also create a register of type **bool**. Initialize it with false.
3. Translate the temporary conversions into an assignment to the corresponding local variable. At the same time make the corresponding **bool** true.

Differences with LLVM

Our register/stack machine is close to LLVM, which is used by the Clang compiler. The main differences are:

1. LLVM allows direct function calls, e.g. without writing and reading parameters on the stack.
2. LLVM strictly forbids multiple assignments to the same variable, even on different branches. This results in a normal form, called static single assignment (SSA). So called ϕ-functions are introduced at merge points.
3. LLVM uses one single pointer instruction `getelementpntr` for all pointer calculations.
4. In LLVM, the instructions have typed names, instead of the registers. It makes no difference.

Forgotten Topics

- I did not say anything about inheritance. It doesn't have impact on the compilation algorithm, as far as I can see. It makes method definitions a bit more complicated.
- The exception mechanism in C^{++} is very complicated. It has has to be implemented in such a way that exeptions that are not thrown result in zero cost.

  As a consequence, one cannot check exceptions by **ifs** after returning. The exception mechanism has to unwind the call stack and call destructors, when an exception is thrown.

  B.t.w. many people, I also, believe that the exception mechanism (and the concept of exception) in C^{++} is a failed design.