G53CMP: Lecture 13 & 14

Code Generation II

Henrik Nilsson

University of Nottingham, UK
Let us generate code for:

```plaintext
let
  var f: Integer := 1;
  var i: Integer := 1
in
  while i <= 10 do begin
    f := f * i;
    putint(f);
    i := i + 1
  end
```
And for this program using arrays and a procedure:

```plaintext
let

proc swap(var x: Integer, var y: Integer)
  let
    var t: Integer
  in begin
    t := x; x := y; y := t
  end;

var a: Integer[5] := [7,3,1,9,2];
var i: Integer;
var j: Integer
```
in begin
    i := 0;
    while i < 4 do begin
        j := i + 1;
        while j < 5 do begin
            if a[i] > a[j] then
                swap(a[i], a[j])
            else skip();
            j := j + 1
        end;
        i := i + 1
    end;
end;
i := 0;
while i <= 4 do begin
  putint(a[i]);
  i := i + 1
end
end
Specifying Code Selection (1)

- Code selection is specified *inductively* over the phrases of the source language:

\[
\text{Command} \rightarrow \underline{\text{Identifier}} := \text{Expression} \\
\mid \text{Command} ; \text{Command} \\
\ldots
\]
Specifying Code Selection (1)

- Code selection is specified *inductively* over the phrases of the source language:

  \[
  \text{Command} \rightarrow \text{Identifier} := \text{Expression} \\
  | \text{Command} ; \text{Command}
  \]

- **Code Function**: maps a source phrase to an instruction sequence. For example:

  \[
  \text{execute} : \text{Command} \rightarrow \text{Instruction}^* \\
  \text{evaluate} : \text{Expression} \rightarrow \text{Instruction}^* \\
  \text{elaborate} : \text{Declaration} \rightarrow \text{Instruction}^*
  \]
Specifying Code Selection (2)

Note:
Specifying Code Selection (2)

Note:

- `execute` generates code for executing a command (it does not execute a command directly);
Specifying Code Selection (2)

Note:

- `execute` generates code for executing a command (it does not execute a command directly);
- `evaluate` generates code for evaluating an expression, leaving the result on the top of the stack.
Specifying Code Selection (2)

Note:

- **execute** generates code for executing a command (it does not execute a command directly);

- **evaluate** generates code for evaluating an expression, leaving the result on the top of the stack.

- **elaborate** generates code for reserving storage for declared constants and variables, evaluating any initialisation expressions, and for declared procedures and functions.
Specifying Code Selection (3)

- Code functions are specified by means of **code templates**:

  \[
  \text{execute} \left[ \begin{array}{c} C_1 \\ ; \\ C_2 \end{array} \right] = \\
  \text{execute} \ C_1 \\
  \text{execute} \ C_2
  \]
Specifying Code Selection (3)

- Code functions are specified by means of **code templates**:

\[
\text{execute } \left[ C_1 ; C_2 \right] = \text{execute } C_1 ; \text{execute } C_2
\]

- The brackets `[` and `]` enclose pieces of **concrete syntax** and **meta variables**.
Specifying Code Selection (3)

- Code functions are specified by means of code templates:

\[
\text{execute } \left[ \begin{array}{c}
C_1 \\
; \\
C_2
\end{array} \right] =
\]

- The brackets \([ \] and \]) enclose pieces of concrete syntax and meta variables.

- Note the recursion; i.e. inductive definition over the underlying phrase structure.
Specifying Code Selection (3)

- Code functions are specified by means of code templates:

\[
\text{execute} \left[ C_1 \ ; \ C_2 \right] = \text{execute} \ C_1 \ ; \text{execute} \ C_2
\]

- The brackets [ ] and ] enclose pieces of concrete syntax and meta variables.
- Note the recursion; i.e. inductive definition over the underlying phrase structure.

(Think of \([ \cdot ]\) as a map from concrete to abstract syntax as specified by the abstract syntax grammars.)
Specifying Code Selection (4)

In a simple language, the code template for assignment might be:

\[
\text{execute } \left[ I := E \right] = \\
\text{evaluate } E \\
\text{STORE } addr(I)
\]

where

\[
addr : \text{Identifier} \rightarrow \text{Address}
\]
In a simple language, the code template for assignment might be:

\[
\text{execute } [ I := E ] = \\
\text{evaluate } E \\
\text{STORE } addr(I)
\]

where

\[addr : \text{Identifier} \rightarrow \text{Address}\]

Note that the instruction sequences and individual instructions in the RHS of the defining equation are implicitly \textit{concatenated}.
Specifying Code Selection (4)

In a simple language, the code template for assignment might be:

\[
\text{execute} \ [ \ I := E ] = \\
\text{evaluate } E \\
\text{STORE } addr(I)
\]

where

\[
addr : \text{Identifier} \rightarrow \text{Address}
\]

Note that the instruction sequences and individual instructions in the RHS of the defining equation are implicitly \textit{concatenated}.

Note: meta variables range over \textit{abstract syntax}. 
Exercise: Code Templates

Generate code for the fragment

\[ f := f \times n; \]
\[ n := n - 1 \]

using the following two templates:

- \text{execute} \left[ C_1 ; C_2 \right] = \text{execute} C_1 \text{execute} C_2
- \text{execute} \left[ I := E \right] = \text{evaluate} E \text{STORE} \text{addr}(I)

and \( \text{addr}(f) = [\text{SB} + 11], \text{addr}(n) = [\text{SB} + 17] \).

Expand as far as the above templates allow.
Not Quite that Simple …

However, something is clearly missing! Recall:

\[
\begin{align*}
\text{execute} & : \text{Command} \rightarrow \text{Instruction}^* \\
\text{evaluate} & : \text{Expression} \rightarrow \text{Instruction}^* \\
\text{elaborate} & : \text{Declaration} \rightarrow \text{Instruction}^* \\
\text{addr} & : \text{Identifier} \rightarrow \text{Address}
\end{align*}
\]

and consider again:

\[
\text{execute} \left[ I := E \right] = \\
\text{evaluate} \ E \\
\text{STORE} \ \text{addr}(I)
\]
Not Quite that Simple …

However, something is clearly missing! Recall:

execute : Command → Instruction*
evaluate : Expression → Instruction*
elaborate : Declaration → Instruction*
addr : Identifier → Address

and consider again:

execute \([ I := E ] \) =
evaluate \( E \)
STORE addr\((I)\)

How can the function \( \text{addr} \) possibly map an identifier (a name) to an address?
In more detail:
Not Quite that Simple … (2)

In more detail:

• *elaborate* is responsible for *assigning* addresses to variables
In more detail:

- *elaborate* is responsible for *assigning* addresses to variables
- a function like *addr* needs *access* to the addresses assigned by *elaborate*
Not Quite that Simple … (2)

In more detail:

- *elaborate* is responsible for **assigning** addresses to variables
- a function like *addr* needs **access** to the addresses assigned by *elaborate*
- but the given type signatures for the code functions do **not permit** this communication!
Not Quite that Simple … (3)

Consequently:
Consequently:

- The code functions need an additional stack environment argument, associating variables with addresses.
Consequently:

- The code functions need an additional *stack environment argument*, associating variables with addresses.
- The code function *elaborate* must *return an updated stack environment*. 
Not Quite that Simple … (3)

Consequently:

- The code functions need an additional *stack environment argument*, associating variables with addresses.

- The code function *elaborate* must *return an updated stack environment*.

- Need to keep track of the *current stack depth* (with respect to LB) to allow *elaborate* to determine the address (within activation record) for a new variable.
Not Quite that Simple … (5)

- Need to keep track of the current scope level as the difference of current scope level and the scope level of a variable is needed in addition to its address to access it (see later lecture on run-time organisation and *static links*).
Not Quite that Simple … (5)

- Need to keep track of the current scope level as the difference of current scope level and the scope level of a variable is needed in addition to its address to access it (see later lecture on run-time organisation and static links).

Moreover, need to generate fresh names for jump targets (recall the demo).
To clearly convey the basic ideas first, we will:

- Use simplified MiniTriangle as main example:
  - No user-defined procedures or functions (only predefined, global ones).
  - Consequently, all variables are global (addressed with respect to $SB$).
  - No arrays (only simple variables, all of size 1 word).

- Gloss over the bookkeeping details for the most part.
However:

- Additional details will be given occasionally.
- Will revisit at appropriate points in lectures on run-time organisation.
- Refer to the HMTC (coursework compiler) source code for full details.
In the simplified exposition, we can consider the code functions to have the following types:

- **run**: $\text{Program} \rightarrow \text{Instruction}^*$
- **execute**: $\text{Command} \rightarrow \text{Instruction}^*$
- **execute**: $\text{Command}^* \rightarrow \text{Instruction}^*$
- **evaluate**: $\text{Expression} \rightarrow \text{Instruction}^*$
- **evaluate**: $\text{Expression}^* \rightarrow \text{Instruction}^*$
- **fetch**: $\text{Identifier} \rightarrow \text{Instruction}^*$
- **assign**: $\text{Identifier} \rightarrow \text{Instruction}^*$
- **elaborate**: $\text{Declaration} \rightarrow \text{Instruction}^*$
- **elaborate**: $\text{Declaration}^* \rightarrow \text{Instruction}^*$
A Code Generation Monad

HMTC uses a *Code Generation monad* to facilitate some of the bookkeeping:

```haskell
instance Monad (CG instr)
```
A Code Generation Monad

HMTC uses a *Code Generation monad* to facilitate some of the bookkeeping:

```
instance Monad (CG instr)
```

Takes care of:

- Collation of generated instructions
- Generation of fresh names
A Code Generation Monad

HMTC uses a *Code Generation monad* to facilitate some of the bookkeeping:

```haskell
instance Monad (CG instr)
```

Takes care of:

- Collation of generated instructions
- Generation of fresh names

Typical operations:

- `emit :: instr -> CG instr ()`
- `newName :: CG instr Name`
Some HMTC Code Functions

execute :: Level -> CGEnv -> Depth -> Command
    -> CG TAMInst ()

evaluate :: Level -> CGEnv -> Expression
    -> CG TAMInst ()

elaborateDecls :: Level -> CGEnv -> Depth
    -> [Declaration]
    -> CG TAMInst (CGEnv, Depth)

(In essence: actual signatures differ in minor ways.)
(Simplified: no procedures, functions, arrays)

Program $\rightarrow$ Command

Command $\rightarrow$ $\textbf{Identifier} \ := \ Expression$

| $\textbf{Identifier} \ (\ Expression^* \ )$
| $\textbf{begin} \ Command^* \ \textbf{end}$
| $\textbf{if} \ Expression \ \textbf{then} \ Command$
| $\textbf{else} \ Command$
| $\textbf{while} \ Expression \ \textbf{do} \ Command$
| $\textbf{let} \ Declaration^* \ \textbf{in} \ Command$
Meta Variable Conventions

\[
\begin{align*}
C & \in \text{Command} \\
Cs & \in \text{Command}^* \\
E & \in \text{Expression} \\
Es & \in \text{Expression}^* \\
D & \in \text{Declaration} \\
Ds & \in \text{Declaration}^* \\
I & \in \text{Identifier} \\
O & \in \text{Operator} \\
IL & \in \text{IntegerLiteral} \\
TD & \in \text{TypeDenoter}
\end{align*}
\]
Code Function \textit{execute} (1)

\begin{align*}
\text{run} & \colon \text{Program} \to \text{Instruction}^* \\
\text{execute} & \colon \text{Command} \to \text{Instruction}^*
\end{align*}

\begin{align*}
\text{run} \left[ C \right] &= \\
\text{execute} & C \\
\text{HALT}
\end{align*}

\begin{align*}
\text{execute} \left[ I := E \right] &= \\
\text{evaluate} & E \\
\text{assign} & I
\end{align*}
In detail (pseudo Haskell, code generation monad) the code for assignment looks more like this. Note that the variable actually is represented by an expression that gets evaluated to an address:

\[
\text{execute } l \ env \ n \ [E_v := E] = \text{ do }
\]
\[
\begin{align*}
\text{evaluate } l \ env \ E \\
\text{evaluate } l \ env \ E_v \\
\text{case } \text{sizeof}(E) \text{ of} \\
1 & \rightarrow \text{emit (STOREI 0)} \\
\text{size} & \rightarrow \text{emit (STOREIB size)}
\end{align*}
\]

(Reasons include: array references (a[i]), call by reference parameters.)
Code Function \textit{execute} (3)

\[
\text{execute } \left[ I \left( Es \right) \right] =
\]
\[
\text{evaluate}^* \ Es
\]
\[
\text{CALL addr}(I)
\]

\[
\text{execute } \left[ \text{begin } Cs \text{ end} \right] =
\]
\[
\text{execute}^* \ Cs
\]
Code Function \textit{execute} (4)

\begin{align*}
\text{execute} \left[ \text{if } E \text{ then } C_1 \text{ else } C_2 \right] &= \\
&= \text{evaluate } E \\
&\quad \text{JUMPIFZ } g \\
&\quad \text{execute } C_1 \\
&\quad \text{JUMP } h \\
&\quad g : \text{ execute } C_2 \\
&\quad h : \\
\end{align*}

where \( g \) and \( h \) are \textbf{fresh} names.
Exercise: Code Function \textit{execute}

Given

\begin{align*}
\text{evaluate } [I] &= \\
\text{LOAD } addr(I) &= \\
\text{execute } [I := IL] &= \\
\text{LOADL } IL &= \\
\text{STORE } addr(I) &=
\end{align*}

\begin{align*}
addr(a) &= [SB + 11] \\
addr(b) &= [SB + 12] \\
addr(c) &= [SB + 13]
\end{align*}

generate code for:

\begin{verbatim}
if b then
  if c then a := 1 else a := 2
else
  a := 3
\end{verbatim}
Code Function **execute** \( (5) \)

In detail (pseudo Haskell, code generation monad):

\[
execute \ l \ env \ n \ [ \text{if} \ E \ \text{then} \ C_1 \ \text{else} \ C_2 \] = \ do \\
g \leftarrow \text{newName} \\
h \leftarrow \text{newName} \\
evaluate \ l \ env \ E \\
emit (\text{JUMPIFZ} \ g) \\
execute \ l \ env \ n \ C_1 \\
emit (\text{JUMP} \ h) \\
emit (\text{Label} \ g) \\
execute \ l \ env \ n \ C_2 \\
emit (\text{Label} \ h)
\]
Code Function \textit{execute} (6)

\textit{execute} \left[ \text{while } E \text{ do } C \right] = \begin{align*}
\text{JUMP } h \\
g & : \text{execute } C \\
h & : \text{evaluate } E \\
\text{JUMP IFNZ } g
\end{align*}

where $g$ and $h$ are \textit{fresh} names.
Code Function $execute$ (7)

\[
execute \left[ \text{let } Ds \text{ in } C \right] =
\]
\[
\text{elaborate}^* \ Ds
\]
\[
execute \ C
\]
\[
\text{POP } 0 \ s
\]

where $s$ is the amount of storage allocated by $\text{elaborate}^* \ Ds$. 
Code Function \textit{execute} (8)

In detail (pseudo Haskell, code generation monad):

\[
\text{execute } l \text{ env } n \left[ \text{let } Ds \text{ in } C \right] = \text{do}
\]

\[
(env', n') \leftarrow \text{elaborate}^* l \text{ env } n \text{ Ds} \\
\text{execute } l \text{ env'} n' C \\
\text{emit } (\text{POP } 0 (n' - n))
\]

where:

\[
\text{elaborate}^* : \text{Level} \rightarrow \text{CGEnv} \rightarrow \text{Depth} \\
\rightarrow \text{Declaration}^* \\
\rightarrow \text{CG TAMInst} (\text{Env}, \text{Depth})
\]
The code function $\text{execute}^*$ has the obvious definition:

$$\text{execute}^* \left[ \epsilon \right] = \epsilon$$

$$\text{execute}^* \left[ C ; Cs \right] = \text{execute} C \text{execute}^* Cs$$
**MiniTriangle Abstract Syntax Part II**

\[
\begin{align*}
\text{Expression} & \rightarrow \text{IntegerLiteral} \quad \text{ExpLitInt} \\
& \mid \text{Identifier} \quad \text{ExpVar} \\
& \mid \text{Operator Expression} \quad \text{ExpUnOpApp} \\
& \mid \text{Expression Operator} \quad \text{ExpBinOpApp} \\
\text{Expression} \\
\text{Declaration} & \rightarrow \textbf{const} \text{ Identifier : } \quad \text{DeclConst} \\
& \quad \text{TypeDenoter = Expression} \\
& \mid \textbf{var} \text{ Identifier : TypeDenoter} \quad \text{DeclVar} \\
& \quad (:\= \text{Expression} \mid \epsilon) \\
\text{TypeDenoter} & \rightarrow \text{Identifier} \quad \text{TDBaseType}
\end{align*}
\]
Code Function $evaluate$ (1)

\[
evaluate : \text{Expression} \rightarrow \text{Instruction}^*
\]

Fundamental invariant: all operations take arguments from the stack and writes result back onto the stack.
Consider evaluating $2 + 4 \times 3 - 5$. Plausible instruction sequence:

- **LOADL 2**
  - **Stack:** 2

- **LOADL 4**
  - **Stack:** 4, 2

- **LOADL 3**
  - **Stack:** 3, 4, 2

- **CALL mul**
  - **Stack:** 12, 2

- **CALL add**
  - **Stack:** 14

- **LOADL 5**
  - **Stack:** 5, 14

- **CALL sub**
  - **Stack:** 9

* (mul, add, sub are routines in the MiniTriangle standard library.)
Code Function \textit{evaluate} (3)

\[
evaluate[L] = \text{LOADL } c
\]

where \( c \) is the value of \( IL \).

\[
evaluate[I] = \text{fetch } I
\]
**Code Function** \( \text{evaluate (4)} \)

\[
evaluate [ \ominus E ] =
\]
\[
\text{evaluate } E
\]
\[
\text{CALL } addr(\ominus)
\]

\[
evaluate [ E_1 \otimes E_2 ] =
\]
\[
\text{evaluate } E_1
\]
\[
\text{evaluate } E_2
\]
\[
\text{CALL } addr(\otimes)
\]

(A call to a known function that can be replaced by a short code sequence can be optimised away at a later stage; e.g. \text{CALL add} \Rightarrow \text{ADD}.)

G53CMP: Lecture 13 & 14 – p.36/46
Code Functions *fetch* and *assign* (1)

In simplified MiniTriangle, all constants and variables are *global*. Hence addressing relative to $SB$.

$$
\text{fetch } [I] = \text{LOAD } [SB + d]
$$

where $d$ is offset (or *displacement*) of $I$ relative to $SB$.

$$
\text{assign } [I] = \text{STORE } [SB + d]
$$

where $d$ is offset of $I$ relative to $SB$. 
Code Functions *fetch* and *assign* (2)

In a more realistic language, *fetch* and *assign* would take the current scope level and the scope level of the variable into account:
Code Functions *fetch* and *assign* (2)

In a more realistic language, *fetch* and *assign* would take the current scope level and the scope level of the variable into account:

- Global variables addressed relative to SB.
Code Functions *fetch* and *assign* (2)

In a more realistic language, *fetch* and *assign* would take the current scope level and the scope level of the variable into account:

- Global variables addressed relative to $SB$.
- Local variables addressed relative to $LB$. 
In a more realistic language, `fetch` and `assign` would take the current scope level and the scope level of the variable into account:

- Global variables addressed relative to $SB$.
- Local variables addressed relative to $LB$.
- Non-global variables in enclosing scopes would be reached by following the *static links* (see later lecture) in one or more steps, and `fetch` and `assign` would have to generate the appropriate code.
Exercise: Code Function evaluate

Given

\[
\begin{align*}
\text{addr}(a) &= [SB + 11] \\
\text{addr}(b) &= [SB + 12] \\
\text{addr}(+) &= \text{add} \\
\text{addr}(\times) &= \text{mult}
\end{align*}
\]

generate code for:

\[a + (b \times 2)\]
Elaboration must deposit value/reserve space for value on stack. Also, address (offset) of elaborated entity must be recorded (to be used by *fetch* and *assign*).
Code Function \textit{elaborate} (1)

Elaboration must deposit value/reserve space for value on stack. Also, address (offset) of elaborated entity must be recorded (to be used by \textit{fetch} and \textit{assign}).

\[
elaborate : \text{Declaration} \rightarrow \text{Instruction}^* \\
elaborate [\text{const } I : TD = E] = \\
\text{evaluate } E
\]
Elaboration must deposit value/reserve space for value on stack. Also, address (offset) of elaborated entity must be recorded (to be used by *fetch* and *assign*).

\[
\text{elaborate} : \text{Declaration} \rightarrow \text{Instruction}^* \\
\text{elaborate \left[ \const I : TD = E \right] = evaluate E}
\]

(Additionally, the offset (w.r.t. SB) has to be recorded for the identifier denoted by \( I \).)
**Code Function** `elaborate (2)`

\[
\text{elaborate} \left[ \text{var } I : TD \right] = \\
\text{LOADL } 0
\]

\[
\text{elaborate} \left[ \text{var } I : TD := E \right] = \\
\text{evaluate } E
\]

(Additionally, the offset (w.r.t. SB) has to be recorded for the identifier denoted by \( I \).)

\text{LOADL } 0 \text{ is just used to reserve space on the stack; the value of the literal does not matter. More space must be reserved if the values of the type are big (e.g. record, array).}
For procedures and functions:

- Generate a fresh name for the entry point.
- Elaborate the formal arguments to extend the environment.
- Generate code for the body at a scope level incremented by 1 and in the extended environment.
Identifiers vs. Symbols (1)

- The coursework compiler HMTC uses symbols instead of identifiers in the latter stages.
- Symbols are introduced in the type checker (responsible for identification in HMTC) in place of identifiers (rep. changed from AST to MTIR).
- Symbols carry semantic information (e.g., type, scope level) to make that information readily available to e.g. the code generator.

(Cf. the lectures on identification where applied identifier occurrences were annotated with semantic information.)
Identifiers vs. Symbols (2)

- Two kinds of (term-level) symbols:
  - External: defined outside the current compilation unit (e.g., in a library).
  - Internal: defined in the current compilation unit (in a `let`).

```haskell
type TermSym = Either ExtTermSym IntTermSym

data ExtTermSym = ExtTermSym { ... }

data IntTermSym = IntTermSym { ... }
```
External Symbols

- External symbols are known entities.
- Can thus be looked up once and for all (during identification).
- Have a value, such as a (symbolic) address.

```haskell
data ExtTermSym = ExtTermSym {
    etmsName :: Name,
    etmsType :: Type,
    etmsVal :: ExtSymVal
}
```

```haskell
data ExtSymVal = ESVLbl Name | ESVInt MTInt | ...
```
Internal Symbols

- Internal symbols do not carry any value such as stack displacement because this is not computed until the time of code generation.
- Such “late” information about an entity referred to by an internal symbol thus has to be looked up in the code generation environment.

```haskell
data IntTermSym = IntTermSym {
  itmsLvl :: ScopeLvl,
  itmsName :: Name,
  itmsType :: Type,
  itmsSrcPos :: SrcPos
}
```