

Compilerconstructie

najaar 2012

<http://www.liacs.nl/home/rvvliet/coco/>

Rudy van Vliet

kamer 124 Snellius, tel. 071-527 5777

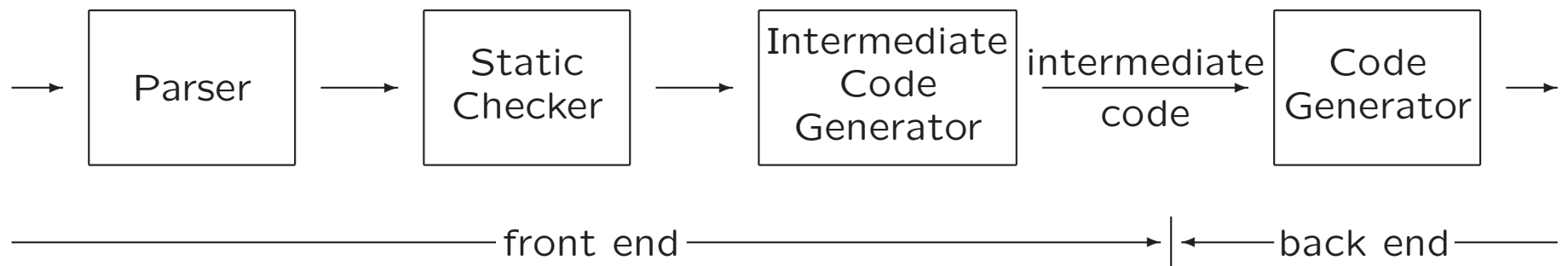
rvvliet(at)liacs.nl

college 6, dinsdag 23 oktober 2012

Intermediate Code Generation

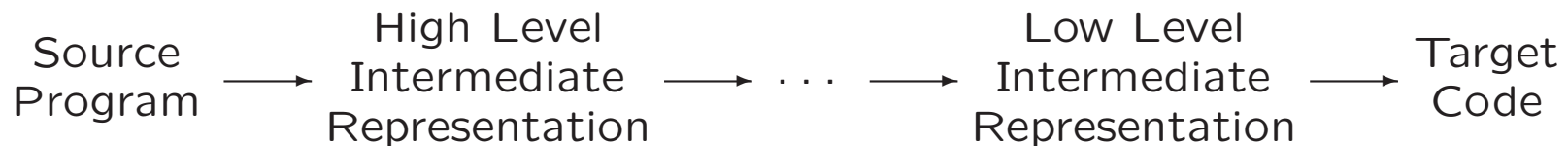
6. Intermediate Code Generation

- Front end: generates intermediate representation
- Back end: generates target code



Intermediate Representation

- Facilitates efficient compiler suites: $m + n$ instead of $m * n$
- Different types, e.g.,
 - syntax trees
 - three-address code: $x = y \text{ op } z$
- High-level vs. low-level
- C for C++

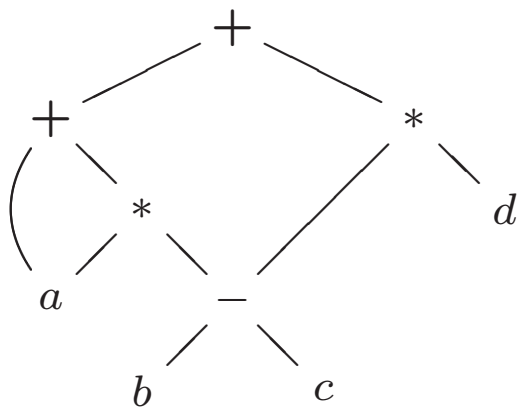


6.2 Three-Address Code

- Linearized representation of syntax tree / syntax DAG
- Sequence of instructions: $x = y \text{ op } z$

Example: $a + a * (b - c) + (b - c) * d$

Syntax DAG



Three-address code

```
t1 = b - c
t2 = a * t1
t3 = a * t2
t4 = t1 * d
t5 = t3 + t4
```

Addresses

At most three addresses per instruction

- Name: source program name / symbol-table entry
- Constant
- Compiler-generated temporary: distinct names

Three-Address Instructions

1	Assignment instructions	$x = y \text{ op } z$
2	Assignment instructions	$x = \text{op } y$
3	Copy instructions	$x = y$
4	Unconditional jumps	goto L
5	Conditional jumps	if x goto L / ifFalse x goto L
6	Conditional jumps	if $x \text{ relop } y$ goto L / ifFalse...
7	Procedure calls and returns	param x_1 param x_2 ... param x_n call p, n return y
8	Indexed copy instructions	$x = y[i]$ / $x[i] = y$
9	Address and pointer assignments	$x = \&y,$ $x = *y,$ $*x = y$

Symbolic lable L represents index of instruction

Three-Address Instructions (Example)

```
do i = i+1; while (a[i] < v);
```

Syntax tree...

Two examples of possible translations:

Symbolic labels

```
L:  t1 = i+1
    i = t1
    t2 = i * 8
    t3 = a [ t2 ]
    if t3 < v goto L
```

Position numbers

```
100: t1 = i+1
101: i = t1
102: t2 = i * 8
103: t3 = a [ t2 ]
104: if t3 < v goto 100
```

Implementation of Three-Address Instructions

Quadruples: records *op*, *vararg1*, *vararg2*, *result*

Example: $a = b * - c + b * - c$

Syntax tree...

Three-address code

t1 = minus c

t2 = b * t1

t3 = minus c

t4 = b * t3

t5 = t2 + t4

a = t5

	<i>op</i>	<i>vararg1</i>	<i>vararg2</i>	<i>result</i>
0	minus	<i>c</i>		<i>t1</i>
1	*	<i>b</i>	<i>t1</i>	<i>t2</i>
2	minus	<i>c</i>		<i>t3</i>
3	*	<i>b</i>	<i>t3</i>	<i>t4</i>
4	+	<i>t2</i>	<i>t4</i>	<i>t5</i>
5	=	<i>t5</i>		<i>a</i>
			...	

Implementation of Three-Address Instructions

Three-address code

```
t1 = minus c
t2 = b * t1
t3 = minus c
t4 = b * t3
t5 = t2 + t4
a = t5
```

	<i>op</i>	<i>vararg1</i>	<i>vararg2</i>	<i>result</i>
0	minus	<i>c</i>		<i>t1</i>
1	*	<i>b</i>	<i>t1</i>	<i>t2</i>
2	minus	<i>c</i>		<i>t3</i>
3	*	<i>b</i>	<i>t3</i>	<i>t4</i>
4	+	<i>t2</i>	<i>t4</i>	<i>t5</i>
5	=	<i>t5</i>		<i>a</i>
			...	

Exceptions

1. minus, =
2. param
3. jumps

Field *result* mainly for temporaries...

Implementation of Three-Address Instructions

Triples: records *op*, *vararg1*, *vararg2*

Example: $a = b * - c + b * - c$

Syntax tree...

Three-address code

t1 = minus c

t2 = b * t1

t3 = minus c

t4 = b * t3

t5 = t2 + t4

a = t5

	<i>op</i>	<i>vararg1</i>	<i>vararg2</i>
0	minus	<i>c</i>	
1	*	<i>b</i>	(0)
2	minus	<i>c</i>	
3	*	<i>b</i>	(2)
4	+	(1)	(3)
5	=	<i>a</i>	<i>t5</i>
		...	

Implementation of Three-Address Instructions

Three-address code

t1 = minus c

t2 = b * t1

t3 = minus c

t4 = b * t3

t5 = t2 + t4

a = t5

	<i>op</i>	<i>vararg1</i>	<i>vararg2</i>
0	minus	<i>c</i>	
1	*	<i>b</i>	(0)
2	minus	<i>c</i>	
3	*	<i>b</i>	(2)
4	+	(1)	(3)
5	=	<i>a</i>	<i>t5</i>
		...	

Equivalent to DAG

Special case: $x[i] = y$ or $x = y[i]$

Pro: temporaries are implicit

Con: difficult to rearrange code

Implementation of Three-Address Instructions

Indirect triples: pointers to triples

Example: $a = b * - c + b * - c$

Syntax tree...

Three-address code

```
t1 = minus c
t2 = b * t1
t3 = minus c
t4 = b * t3
t5 = t2 + t4
a = t5
```

instruction

35	(0)
36	(1)
37	(2)
38	(3)
39	(4)
40	(5)
	...

	<i>op</i>	<i>vararg1</i>	<i>vararg2</i>
0	minus	<i>c</i>	
1	*	<i>b</i>	(0)
2	minus	<i>c</i>	
3	*	<i>b</i>	(2)
4	+	(1)	(3)
5	=	<i>a</i>	(4)
		...	

6.3.3 Declarations

- Three-address code is simplistic
It assumes that names of variables can be easily resolved by the back end in global or local variables
- We need symbol tables to record global and local declarations in procedures, blocks, and structs to resolve names
- Symbol table contains type and relative address of names

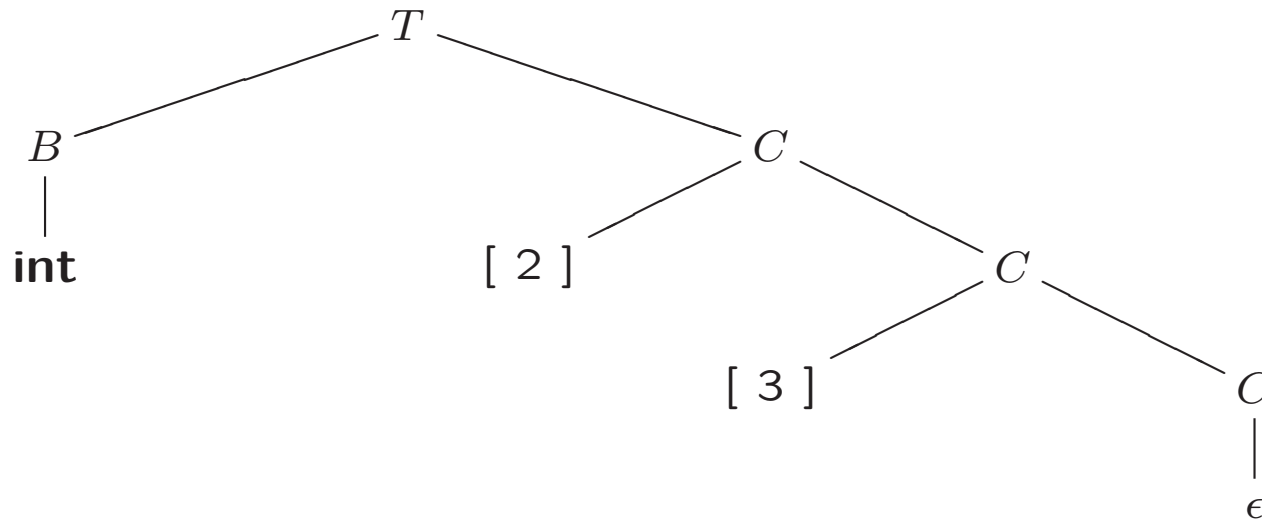
Example:

$$\begin{aligned} D &\rightarrow T \text{ id}; D \mid \epsilon \\ T &\rightarrow B C \mid \text{record } \{ D \} \\ B &\rightarrow \text{int} \mid \text{float} \\ C &\rightarrow \epsilon \mid [\text{num}] C \end{aligned}$$

Structure of Types (Example)

$$\begin{aligned} T &\rightarrow B C \mid \text{record } \{ D \} \\ B &\rightarrow \text{int} \mid \text{float} \\ C &\rightarrow \epsilon \mid [\text{num}] C \end{aligned}$$

int [2] [3]



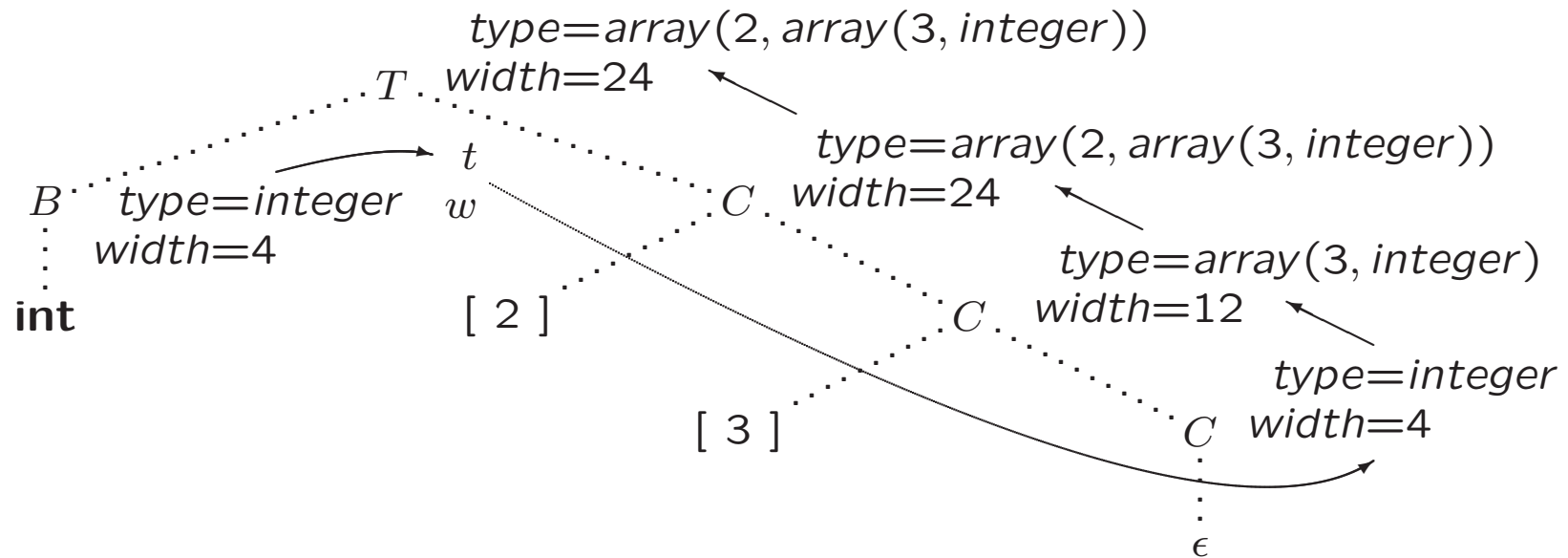
Storage Layout at Compile Time

- Storage comes in blocks of contiguous bytes
- **Width** of type is number of bytes needed

$$\begin{array}{ll} T \rightarrow B & \{ t = B.type; w = B.width; \} \\ & C \quad \{ T.type = C.type; T.width = C.width; \} \\ B \rightarrow \mathbf{int} & \{ B.type = integer; B.width = 4; \} \\ B \rightarrow \mathbf{float} & \{ B.type = float; B.width = 8; \} \\ C \rightarrow \epsilon & \{ C.type = t; C.width = w; \} \\ C \rightarrow [\mathbf{num}] C_1 & \{ C.type = array(\mathbf{num.value}, C_1.type); \\ & C.width = \mathbf{num.value} \times C_1.width; \} \end{array}$$

Types and Their Widths (Example)

$T \rightarrow B$ $\{ t = B.type; w = B.width; \}$
 $\quad \quad C$ $\{ T.type = C.type; T.width = C.width; \}$
 $B \rightarrow \mathbf{int}$ $\{ B.type = \mathit{integer}; B.width = 4; \}$
 $B \rightarrow \mathbf{float}$ $\{ B.type = \mathit{float}; B.width = 8; \}$
 $C \rightarrow \epsilon$ $\{ C.type = t; C.width = w; \}$
 $C \rightarrow [\mathbf{num}] C_1$ $\{ C.type = \mathit{array}(\mathbf{num.value}, C_1.type);$
 $C.width = \mathbf{num.value} \times C_1.width; \}$



Sequences of Declarations

$$D \rightarrow T \mathbf{id}; D \mid \epsilon$$

Use *offset* as next available address

$$\begin{aligned} P &\rightarrow \{ \textit{offset} = 0; \} \\ D &\rightarrow T \mathbf{id}; \{ \textit{top.put}(\mathbf{id.lexeme}, T.type, \textit{offset}); \\ &\quad \textit{offset} = \textit{offset} + T.width; \} \\ D &\rightarrow \epsilon \end{aligned}$$

Fields in Records and Classes

Example

```
float x;  
record { float x; float y; } p;  
record { int tag; float x; float y; } q;  
x = p.x + q.x;
```

$$\begin{aligned} D &\rightarrow T \text{ id}; D \mid \epsilon \\ T &\rightarrow \mathbf{record} \{ D \} \end{aligned}$$

- Fields are specified by sequence of declarations
 - Field names within record must be distinct
 - Relative address for field is relative to data area for that record

Fields in Records and Classes

Stored in separate symbol table t

Record type has form $record(t)$

```
 $T \rightarrow$  record '{' { Env.push(top); top = new Env();  
                      Stack.push(offset); offset = 0; }  
       $D$  '}' { T.type = record(top); T.width = offset;  
            top = Env.pop(); offset = Stack.pop(); }
```

6.4 Translation of Expressions

- Temporary names are created
 $E \rightarrow E_1 + E_2$ yields $t = E_1 + E_2$, e.g.,

t5 = t2 + t4
a = t5

- If expression is identifier, then no new temporary
- Nonterminal E has two attributes:
 - $E.addr$ – address that will hold value of E
 - $E.code$ – three-address code sequence

Syntax-Directed Definition

To produce three-address code for assignments

Production	Semantic Rules
$S \rightarrow \mathbf{id} = E;$	$S.code = E.code \parallel$ $\quad gen(top.get(\mathbf{id}.lexeme) ' = ' E.addr)$
$E \rightarrow E_1 + E_2$	$E.addr = \mathbf{new Temp}()$ $E.code = E_1.code \parallel E_2.code \parallel$ $\quad gen(E.addr ' = ' E_1.addr ' + ' E_2.addr)$
$-E_1$	$E.addr = \mathbf{new Temp}()$ $E.code = E_1.code \parallel$ $\quad gen(E.addr ' = ' 'minus' E_1.addr)$
(E_1)	$E.addr = E_1.addr$ $E.code = E_1.code$
\mathbf{id}	$E.addr = top.get(\mathbf{id}.lexeme)$ $E.code = ''$

Translation scheme

To incrementally produce three-address code for assignments

S	\rightarrow	id = E;	{	$gen(top.get(\mathbf{id}.lexeme) \ ' \ = \ ' \ E.addr);$	}
E	\rightarrow	$E_1 + E_2$	{	$E.addr = \mathbf{new} \ Temp();$ $gen(E.addr \ ' \ = \ ' \ E_1.addr \ ' \ + \ ' \ E_2.addr);$	}
		$-E_1$	{	$E.addr = \mathbf{new} \ Temp();$ $gen(E.addr \ ' \ = \ ' \ \mathbf{minus} \ E_1.addr);$	}
		(E_1)	{	$E.addr = E_1.addr;$	}
		id	{	$E.addr = top.get(\mathbf{id}.lexeme);$	}

Addressing Array Elements

- Array $A[n]$ with elements at positions $0, 1, \dots, n - 1$
- Let
 - w be width of array element
 - $base$ be relative address of storage allocated for A ($= A[0]$)

Element $A[i]$ begins in location $base + i \times w$

- In two dimensions, let
 - w_1 be width of row,
 - w_2 be width of element of row

Element $A[i][j]$ begins in location $base + i \times w_1 + j \times w_2$

- In k dimensions $base + i_1 * w_1 + i_2 * w_2 + \dots + i_k * w_k$

Translation of Array References

L generates array name followed by sequence of index expressions

$$L \rightarrow L[E] \quad | \quad \mathbf{id}[E]$$

Three synthesized attributes

- $L.addr$: temporary used to compute location in array
- $L.array$: pointer to symbol-table entry for array name
 - $L.array.base$: base address of array
- $L.type$: type of **sub**array generated by L
 - For type t : $t.width$
 - For array type t : $t.elem$

Translation of Array References

$S \rightarrow \mathbf{id} = E;$	{	$gen(top.get(\mathbf{id}.lexeme) ' = ' E.addr);$	}
$S \rightarrow L = E;$	{	$gen(L.array.base '['L.addr ']' ' = ' E.addr);$	}
$E \rightarrow E_1 + E_2$	{	$E.addr = \mathbf{new} Temp();$ $gen(E.addr ' = ' E_1.addr ' + ' E_2.addr);$	}
$E \rightarrow \mathbf{id}$	{	$E.addr = top.get(\mathbf{id}.lexeme);$	}
$E \rightarrow L$	{	$E.addr = \mathbf{new} Temp();$ $gen(E.addr ' = ' L.array.base '['L.addr ']);$	}
$L \rightarrow \mathbf{id} [E]$	{	$L.array = top.get(\mathbf{id}.lexeme);$ $L.type = L.array.type.elem;$ $L.addr = \mathbf{new} Temp();$ $gen(L.addr ' = ' E.addr ' * ' L.type.width);$	}
$L \rightarrow L_1[E]$	{	$L.array = L_1.array;$ $L.type = L_1.type.elem;$ $t = \mathbf{new} Temp();$ $L.addr = \mathbf{new} Temp();$ $gen(t ' = ' E.addr ' * ' L.type.width);$ $gen(L.addr ' = ' L_1.addr ' + ' t);$	}

Translation of Array References

$E \rightarrow \mathbf{id}$ { $E.addr = top.get(\mathbf{id}.lexeme);$ }

$L \rightarrow \mathbf{id} [E]$ { $L.array = top.get(\mathbf{id}.lexeme);$
 $L.type = L.array.type.elem;$
 $L.addr = \mathbf{new} Temp();$
 $gen(L.addr ' = ' E.addr ' * ' L.type.width);$ }

$L \rightarrow L_1[E]$ { $L.array = L_1.array;$
 $L.type = L_1.type.elem;$
 $t = \mathbf{new} Temp();$
 $L.addr = \mathbf{new} Temp();$
 $gen(t ' = ' E.addr ' * ' L.type.width);$
 $gen(L.addr ' = ' L_1.addr ' + ' t);$ }

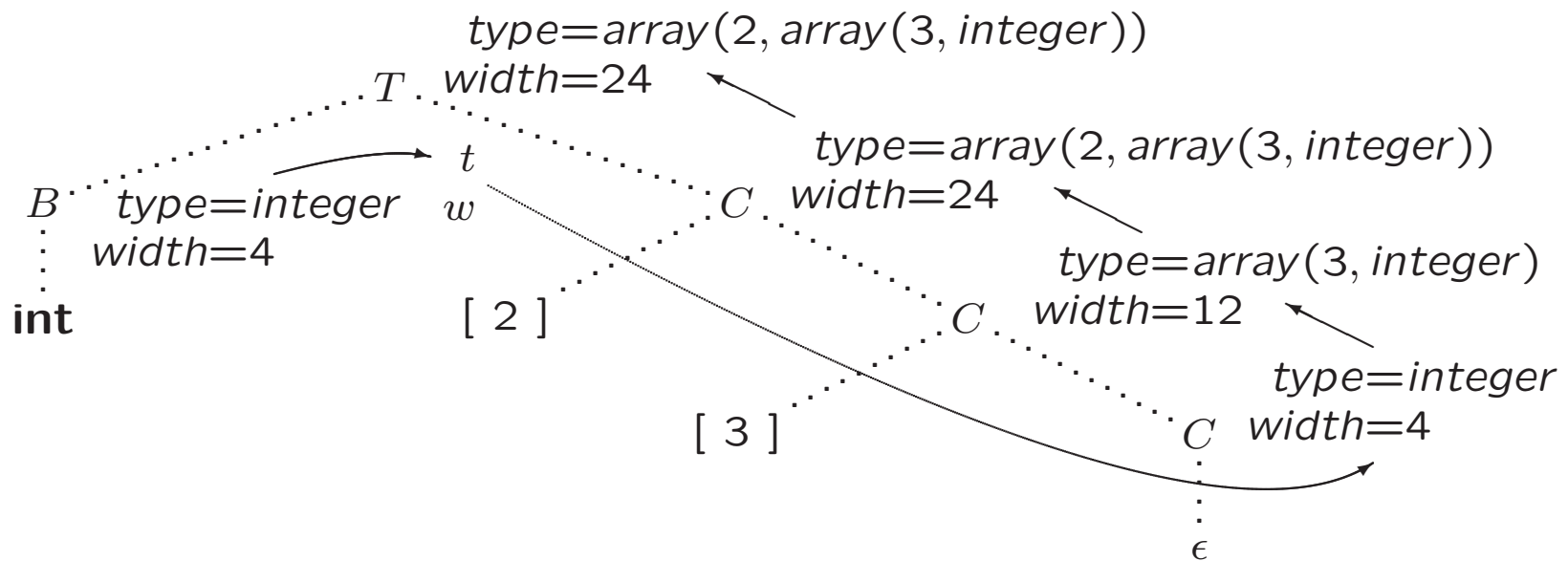
$E \rightarrow L$ { $E.addr = \mathbf{new} Temp();$
 $gen(E.addr ' = ' L.array.base '[' L.addr ']);$ }

$E \rightarrow E_1 + E_2$ { $E.addr = \mathbf{new} Temp();$
 $gen(E.addr ' = ' E_1.addr ' + ' E_2.addr);$ }

$S \rightarrow \mathbf{id} = E;$ { $gen(top.get(\mathbf{id}.lexeme) ' = ' E.addr);$ }

$S \rightarrow L = E;$ { $gen(L.array.base '[' L.addr ']' ' = ' E.addr);$ }

Types and Their Widths (Example)



Translation of Array References (Example)

- Let a be 2×3 array of integers
- Let c , i and j be integers
- Annotated parse tree for expression $c + a[i][j]$

6.6 Control Flow

- Boolean expressions used to
 1. Alter flow of control: **if** (E) S
 2. Compute logical values, cf. arithmetic expressions

- Generated by

$B \rightarrow B||B \mid B\&\&B \mid !B \mid (B) \mid E \text{ rel } E \mid \text{true} \mid \text{false}$

- In $B_1||B_2$, if B_1 is true, then expression is true
In $B_1\&\&B_2$, if ...

Short-Circuit Code

or jumping code

Boolean operators `||`, `&&` and `!` translate into jumps

Example

```
if ( x < 100 || x > 200 && x!=y ) x = 0;
```

Precedence: `||` < `&&` < `!`

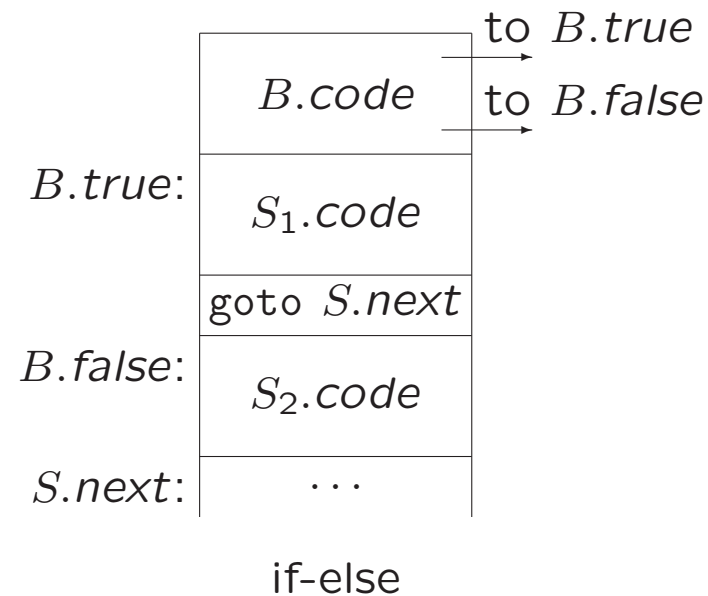
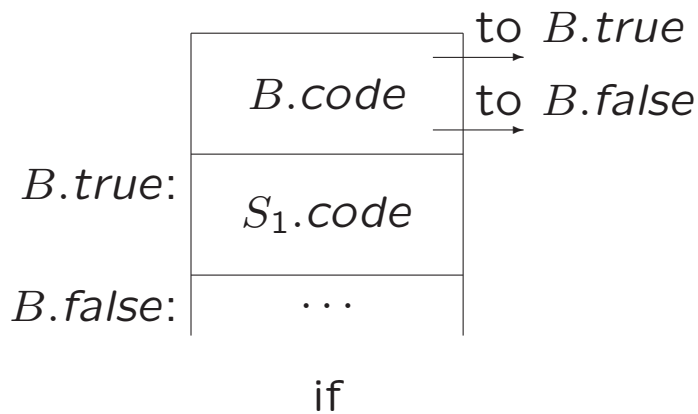
```
    if x < 100 goto L2
    ifFalse x > 200 goto L1
    ifFalse x != y goto L1
L2:  x = 0
L1:
```

Flow-of-Control Statements

$S \rightarrow \text{if } (B) S_1$

$S \rightarrow \text{if } (B) S_1 \text{ else } S_2$

$S \rightarrow \text{while } (B) S_1$



Translation using

- synthesized attributes `B.code` and `S.code`
- inherited attributes (labels) `B.true`, `B.false` and `S.next`

Syntax-Directed Definition

Production	Semantic Rules
$P \rightarrow S$	$S.next = newlabel()$ $P.code = S.code \parallel label(S.next)$
$S \rightarrow \mathbf{if} (B) S_1$	$B.true = newlabel()$ $B.false = S_1.next = S.next$ $S.code = B.code \parallel label(B.true) \parallel S_1.code$
$B \rightarrow B_1 \parallel B_2$	$B_1.true = B.true$ $B_1.false = newlabel()$ $B_2.true = B.true$ $B_2.false = B.false$ $B.code = B_1.code \parallel label(B_1.false) \parallel B_2.code$
$B_1 \rightarrow E_1 \mathbf{rel} E_2$	$B_1.code = E_1.code \parallel E_2.code$ $\parallel gen('if' E_1.addr \mathbf{rel.op} E_2.addr 'goto' B_1.true)$ $\parallel gen('goto' B_1.false)$
$B_2 \rightarrow B_3 \&\& B_4$	$B_3.true = newlabel()$ $B_3.false = B_2.false$ $B_4.true = B_2.true$ $B_4.false = B_2.false$ $B_2.code = B_3.code \parallel label(B_3.true) \parallel B_4.code$

Avoiding Redundant Gotos

```
    if x < 100 goto L2
    goto L3
L3:  if x > 200 goto L4
    goto L1
L4:  if x != y goto L2
    goto L1
L2:  x = 0
L1:
```

Versus

```
    if x < 100 goto L2
    ifFalse x > 200 goto L1
    ifFalse x != y goto L1
L2:  x = 0
L1:
```

6.7 Backpatching

- Code generation problem:
 - Labels (addresses) that control must go to may not be known at the time that jump statements are generated
- One solution:
 - Separate pass to bind labels to addresses
- Other solution: backpatching
 - Generate jump statements with empty target
 - Add such statements to a list
 - Fill in labels when proper label is determined

Backpatching

- **Synthesized** attributes *B.truelist*, *B.falselist*, *S.nextlist* containing lists of jumps
- Three functions
 1. *makelist(i)* creates new list containing index *i*
 2. *merge(p₁, p₂)* concatenates lists pointed to by *p₁* and *p₂*
 3. *backpatch(p, i)* inserts *i* as target label for each instruction on list pointed to by *p*

Grammars for Backpatching

- Grammar for boolean expressions:

$$\begin{aligned} B &\rightarrow B_1 || M B_2 \mid B_1 \&\& M B_2 \mid ! B_1 \mid (B_1) \\ &\quad \mid E_1 \text{ rel } E_2 \mid \text{true} \mid \text{false} \\ M &\rightarrow \epsilon \end{aligned}$$

M is marker nonterminal

- Grammar for flow-of-control statements

$$\begin{aligned} S &\rightarrow \text{if } (B) S_1 \mid \text{if } (B) S_1 \text{ else } S_2 \\ &\quad \mid \text{while } (B) S_1 \mid \{ L \} \mid A ; \\ N &\rightarrow \epsilon \\ L &\rightarrow L_1 S \mid S \end{aligned}$$

Example: `if (x < 100 || x > 200 && x != y) x = 0;`

Translation Scheme for Backpatching

$B \rightarrow E_1 \text{ rel } E_2$	{	$B.\text{truelist} = \text{makelist}(\text{nextinstr});$ $B.\text{falselist} = \text{makelist}(\text{nextinstr} + 1);$ $\text{gen}(\text{'if' } E_1.\text{addr rel.op } E_2.\text{addr 'goto -'});$ $\text{gen}(\text{'goto -'});$	}
$M \rightarrow \epsilon$	{	$M.\text{instr} = \text{nextinstr};$	}
$B_2 \rightarrow B_3 \&\& M B_4$	{	$\text{backpatch}(B_3.\text{truelist}, M.\text{instr});$ $B_2.\text{truelist} = B_4.\text{truelist};$ $B_2.\text{falselist} = \text{merge}(B_3.\text{falselist}, B_4.\text{falselist});$	}
$B \rightarrow B_1 M B_2$	{	$\text{backpatch}(B_1.\text{falselist}, M.\text{instr});$ $B.\text{truelist} = \text{merge}(B_1.\text{truelist}, B_2.\text{truelist});$ $B.\text{falselist} = B_2.\text{falselist};$	}
$S \rightarrow A$	{	$S.\text{nextlist} = \text{null};$	}
$S \rightarrow \text{if } (B) M S_1$	{	$\text{backpatch}(B.\text{truelist}, M.\text{instr});$ $S.\text{nextlist} = \text{merge}(B.\text{falselist}, S_1.\text{nextlist});$	}

6.8 Switch-Statements

```
switch (  $E$  )  
{  
    case  $V_1$ :  $S_1$   
    case  $V_2$ :  $S_2$   
        ...  
    case  $V_{n-1}$ :  $S_{n-1}$   
    default  $S_n$   
}
```

Translation:

1. Evaluate expression E
2. Find value V_j in list of cases that matches value of E
3. Execute statement S_j

Translation of Switch-Statement

```
        code to evaluate E into t
        goto test
L1:   code for S1
        goto next
L2:   code for S2
        goto next
        ...
L_{n-1}: code for S_{n-1}
        goto next
L_{n}:   code for S_n
        goto next
test: if t = V1 goto L1
        if t = V2 goto L2
        ...
        if t = V_{n-1} goto L_{n-1}
        goto L_{n}
next:
```

Volgende week

- Maandag 29 oktober: inleveren opdracht 2
- Practicum over opdracht 3
- Eerst naar 403, daarna naar 302/304
- Inleveren 19 november

Compiler constructie

college 6

Intermediate Code Generation

Chapters for reading:

6.intro, 6.2–6.2.3, 6.3.3–6.3.6, 6.4, 6.6–6.8