

Intra-procedural Data Flow Analysis

Introduction

Hanne Riis Nielson and Flemming Nielson

email: {riis,nielson}@imm.dtu.dk

Informatics and Mathematical Modelling
Technical University of Denmark

©Hanne Riis Nielson & Flemming Nielson

Copyright

This set of transparencies is mainly based on

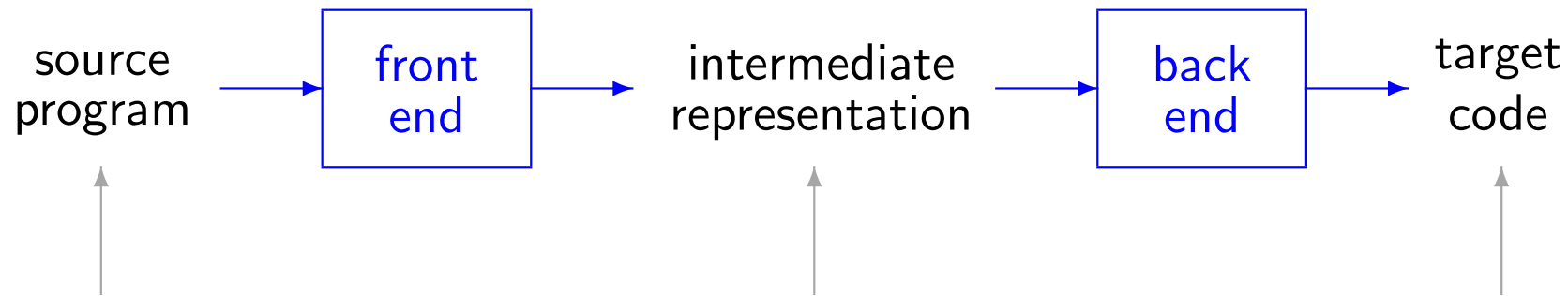
Flemming Nielson, Hanne Riis Nielson, Chris Hankin:
Principles of Program Analysis.
Springer, 1999; ISBN 3-540-65410-0.

They may be used by instructors for presentations based on the book but may not be copied and distributed electronically or by other means.

A handout version with four slides per page may be used for producing paper copies of the slides for students; the students may be charged no fee beyond reasonable production cost.

The handout version of the transparencies is available from the webpage
<http://www.imm.dtu.dk/~riis/ppa.htm>.

The Setting: Optimising Compilers



the programmer can:

- profile program
- change algorithm
- transform program

beyond the compiler

the compiler can:

- improve loops
- procedure calls
- address calculations

architecture independent

the compiler can:

- use registers
- instruction selection
- peephole optimisation

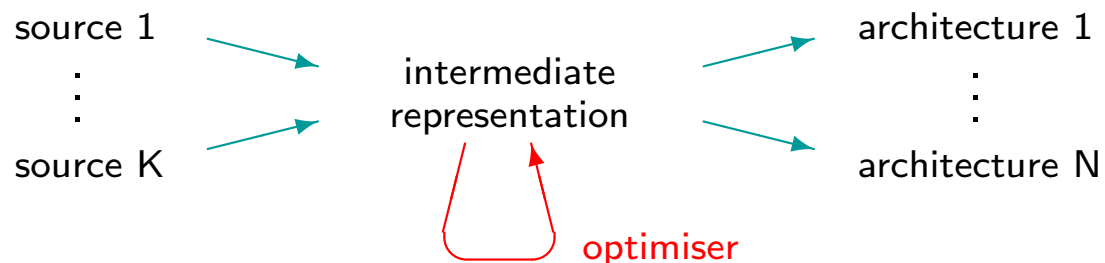
architecture dependent

Criteria for Optimisations

- an optimisation must preserve the meaning of programs
 - must not change the output for any input
 - must not cause errors (e.g. division by 0) that were not present before
- an optimisation must speed up programs by a measurable amount
 - occasionally one optimise for space
 - not every optimisation succeeds in improving every program
 - occasionally an “optimisation” may slow down a program slightly but this is acceptable as long as it improves the average speed up
- an optimisation must be worth the effort
 - the additional time spend by the compiler must be repayed when running the target program

The Level of Optimisations

- source level: **language dependent**
 - e.g. references to array elements give rise to redundant computations
 - * Pascal and Fortran programs cannot be optimised at the source levels
 - * C programs allow both kinds of references and can be optimised
- intermediate level: **language and architecture independent**



- target/low level: **architecture dependent**
 - less likely to be portable

What Optimisations are Worthwhile?

Folklore: most programs spend 90% of their execution time in 10% of the code.

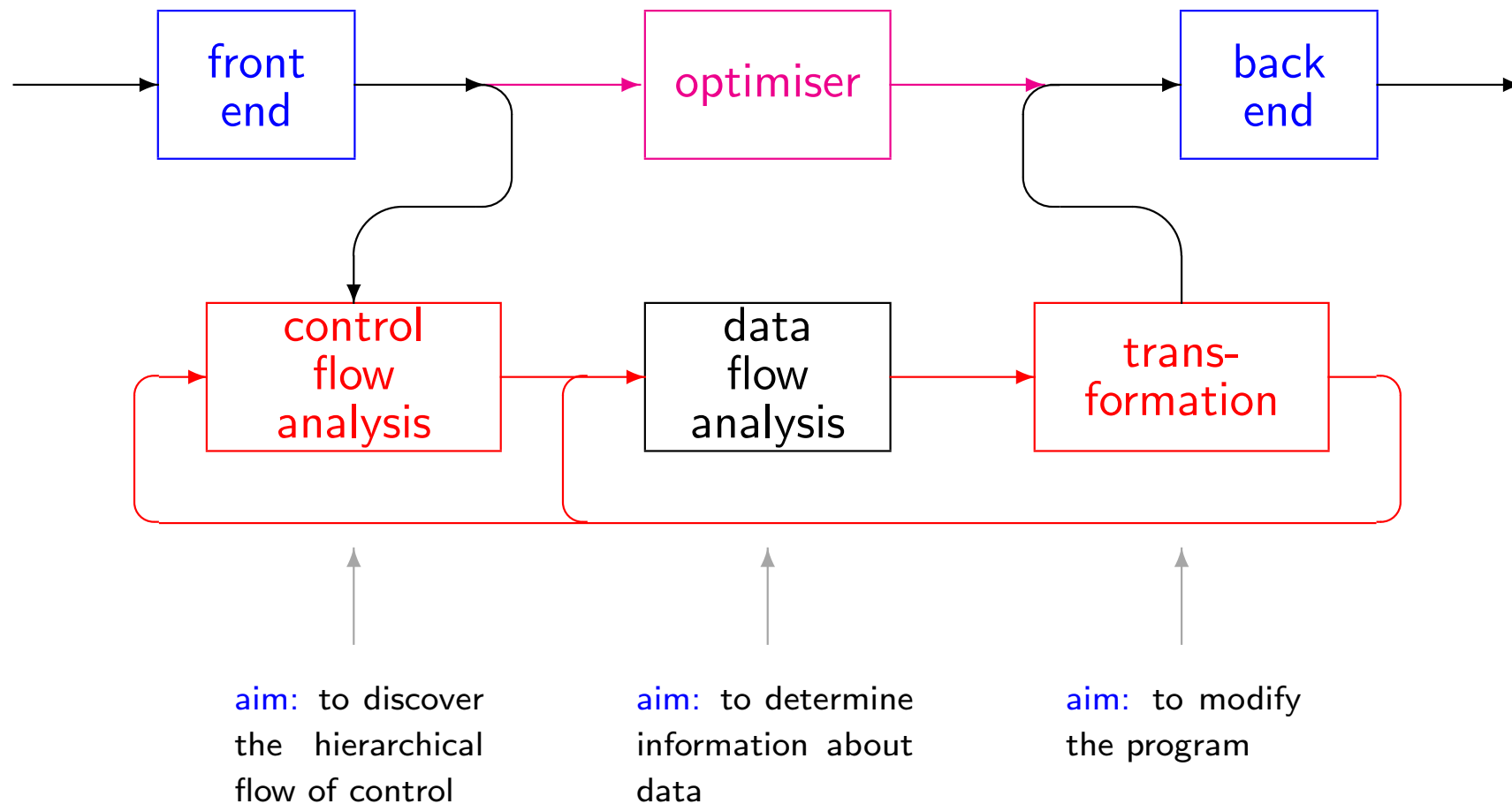
- Inner loops are good candidates
- Experiment, profiling, collect statistics, . . .
- Build on other people's experience, see e.g.

Steven S. Muchnick:

Advanced Compiler Design and Implementation

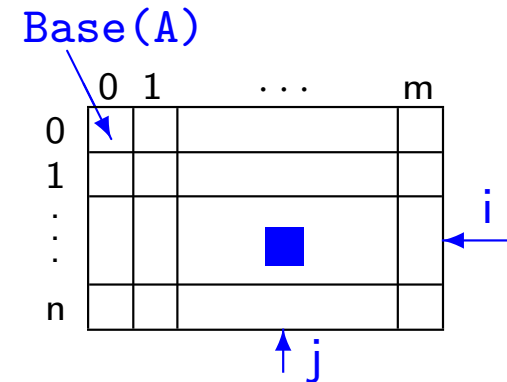
Morgan Kaufmann, 1997

Organisation of an Optimiser



Example

(ack. Reinhard Wilhelm)



Algol-like arrays:

```

i := 0;
while i <= n do
  j := 0;
  while j <= m do
    A[i,j] := B[i,j] + C[i,j];
    j := j+1
  od;
  i := i+1
od

```

C-like arrays:

```

i := 0;
while i <= n do
  j := 0;
  while j <= m do
    temp := Base(A) + i * (m+1) + j;
    Cont(temp) := Cont(Base(B) + i * (m+1) + j)
                + Cont(Base(C) + i * (m+1) + j);
    j := j+1
  od;
  i := i+1
od

```


Typical Optimisations

- **Avoid redundant computations**
 - reuse available results
 - move loop invariant computations out of loops
- **Avoid superfluous computations**
 - results known not to be needed
 - results known already at compile time

```
i := 0;
while i <= n do
  j := 0;
  while j <= m do
    temp := Base(A) + i * (m+1) + j;
    Cont(temp) := Cont(Base(B)
                    + i * (m+1) + j)
                + Cont(Base(C)
                    + i * (m+1) + j);
    j := j+1
  od;
  i := i+1
od
```

Available Expressions Analysis

```

i := 0;
while i <= n do
  j := 0;
  while j <= m do
    temp := Base(A) + i*(m+1) + j;
    Cont(temp) := Cont(Base(B) + i*(m+1) + j)
                 + Cont(Base(C) + i*(m+1) + j);
    j := j+1
  od;
  i := i+1
od

```

Diagram annotations:

- A red box highlights the two lines of code: `temp := Base(A) + i*(m+1) + j;` and `Cont(temp) := Cont(Base(B) + i*(m+1) + j) + Cont(Base(C) + i*(m+1) + j);`
- A teal arrow labeled "first computation" points to the expression `i*(m+1) + j` in the first line of the box.
- Two teal arrows labeled "re-computations" point to the expression `i*(m+1) + j` in the second line of the box.

Common
subexpression
elimination:

```

t1 := i * (m+1) + j;
temp := Base(A) + t1;
Cont(temp) := Cont(Base(B)+t1)
             + Cont(Base(C)+t1);

```

Detection of Loop Invariants

```
i := 0;
while i <= n do
  j := 0;
  while j <= m do
    t1 := i * (m+1) + j;
    temp := Base(A) + t1;
    Cont(temp) := Cont(Base(B) + t1)
                 + Cont(Base(C) + t1);
    j := j+1
  od;
  i := i+1
od
```

loop invariant

Invariant code motion:

```
t2 := i * (m+1);
while j <= m do
  t1 := t2 + j;
  temp := ...
  Cont(temp) := ...
  j := ...
od
```

Detection of Induction Variables

```

i := 0;
while i <= n do
  j := 0;
  t2 := i * (m+1);
  while j <= m do
    t1 := t2 + j;
    temp := Base(A) + t1;
    Cont(temp) := Cont(Base(B) + t1)
                 + Cont(Base(C) + t1);
    j := j+1
  od;
  i := i+1
od

```

induction variable

Strength reduction:

```

i := 0;
t3 := 0;
while i <= n do
  j := 0;
  t2 := t3;
  while j <= m do ... od
  i := i + 1;
  t3 := t3 + (m+1)
od

```

Copy Analysis

```
i := 0;
t3 := 0;
while i <= n do
  j := 0;
  t2 := t3;
  while j <= m do
    t1 := t2 + j;
    temp := Base(A) + t1;
    Cont(temp) := Cont(Base(B) + t1)
                + Cont(Base(C) + t1);
    j := j+1
  od;
  i := i+1;
  t3 := t3 + (m+1)
od
```

The diagram shows a box containing the text `t2 = t3`. Two arrows originate from this box: one points to the assignment `t2 := t3;` in the outer while loop, and the other points to the expression `t2` in the assignment `t1 := t2 + j;` within the inner while loop. The inner while loop is enclosed in a red rectangular box.

Copy propagation:

```
while j <= m do
  t1 := t3 + j;
  temp := ...;
  Cont(temp) := ...;
  j := ...
od
```

Live Variables Analysis

```

i := 0;
t3 := 0;
while i <= n do
  j := 0;
  t2 := t3;
  while j <= m do
    t1 := t3 + j;
    temp := Base(A) + t1;
    Cont(temp) := Cont(Base(B) + t1)
                + Cont(Base(C) + t1);

    j := j+1
  od;
  i := i+1;
  t3 := t3 + (m+1)
od

```

dead variable

Dead code elimination:

```

i := 0;
t3 := 0;
while i <= n do
  j := 0;
  while j <= m do
    t1 := t3 + j;
    temp := Base(A) + t1;
    Cont(temp) := Cont(Base(B) + t1)
                + Cont(Base(C) + t1);

    j := j+1
  od;
  i := i+1;
  t3 := t3 + (m+1)
od

```

Analyses and Transformations

Data Flow Analysis

Transformation

Available expressions analysis

Common subexpression elimination

Detection of loop invariants

Invariant code motion

Detection of induction variables

Strength reduction

Copy analysis

Copy propagation

Live variables analysis

Dead code elimination

...

...

The Essence of Program Analysis

Program analysis offers techniques for predicting

statically at compile-time

safe & efficient **approximations**

to the set of configurations or behaviours arising

dynamically at run-time

we cannot expect
exact answers!

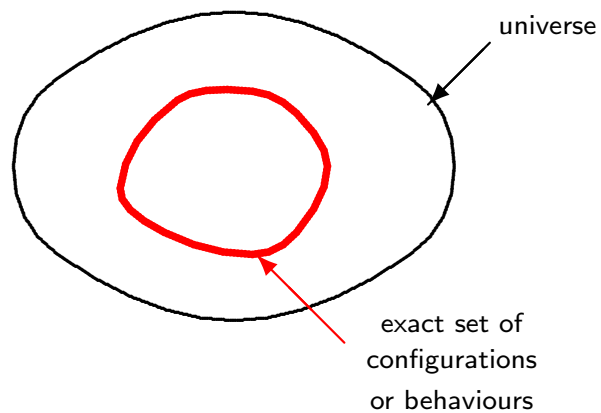
Safe: faithful to the semantics

Efficient: implementation with

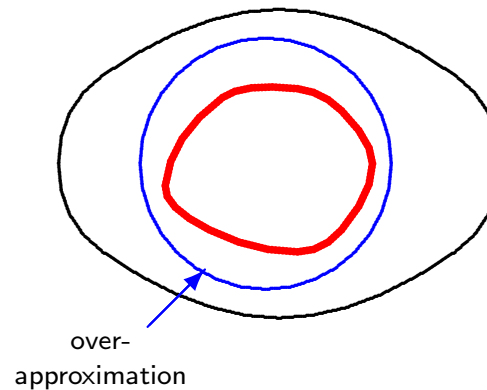
- good time performance and
- low space consumption

The Nature of Approximation

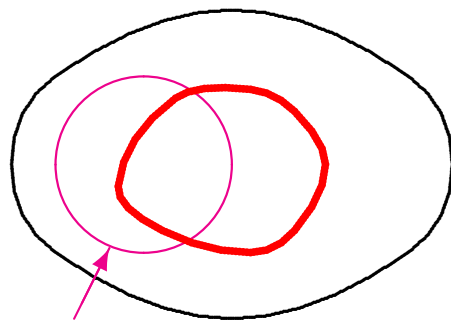
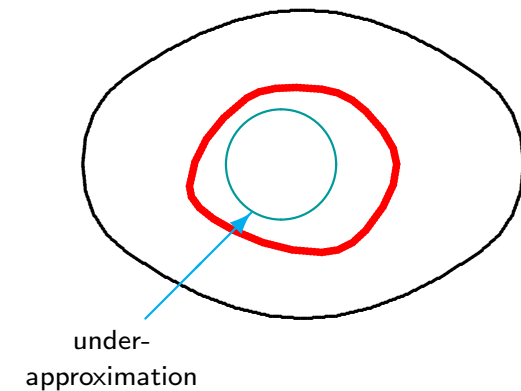
The exact world



Over-approximation



Under-approximation



unacceptable!

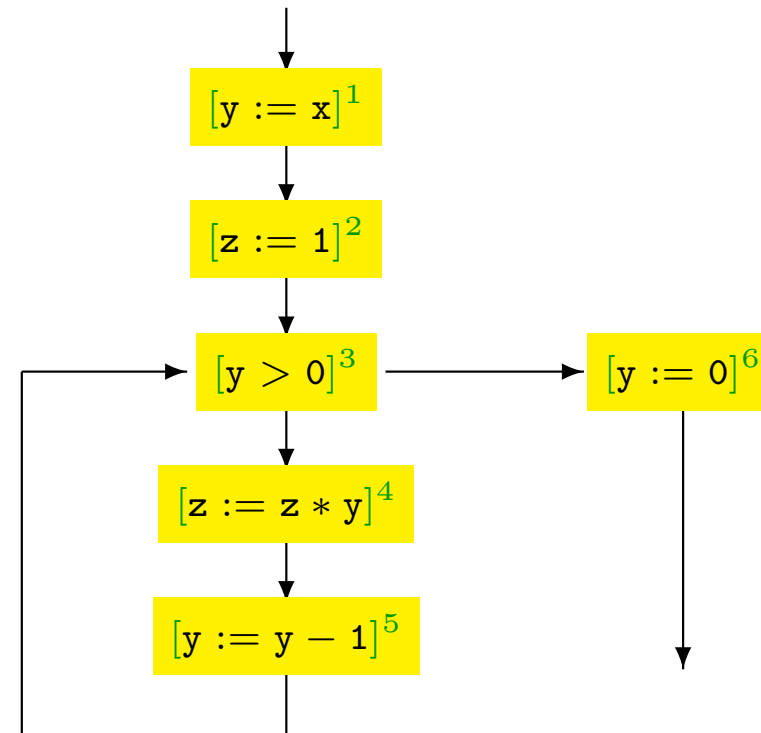
Slogans: Err on the safe side!
Trade precision for efficiency!

Example

Program with labels for elementary blocks:

```
[y := x]1;  
[z := 1]2;  
while [y > 0]3 do  
  [z := z * y]4;  
  [y := y - 1]5  
od;  
[y := 0]6
```

Flow graph:

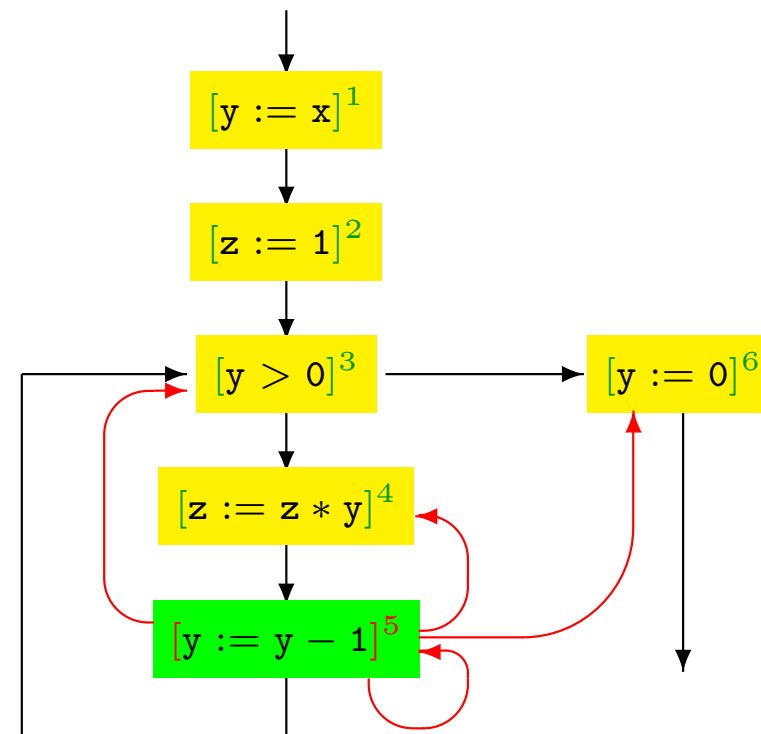


Example: Reaching Definitions Analysis

Problem: which definitions reach which program points

For a simple while language:
a definition of a variable x is an assignment $[x := a]^{\ell}$ to x

The assignment $[x := a]^{\ell}$ **reaches** ℓ' if there is an execution where x was last assigned at ℓ



Analysing the Program by Hand (1)

<code>[y := x]¹;</code>	←	<code>{(x, ?), (y, ?), (z, ?)}</code>
<code>[z := 1]²;</code>	←	<code>{(x, ?), (y, 1), (z, ?)}</code>
<code>while [y > 0]³ do</code>	←	<code>{(x, ?), (y, 1), (z, 2)}</code>
<code>[z := z * y]⁴;</code>	←	<code>{(x, ?), (y, 1), (z, 2)}</code>
<code>[y := y - 1]⁵</code>	←	
<code>od;</code>	←	
<code>[y := 0]⁶</code>	←	<code>{(x, ?), (y, 1), (z, 2)}</code>
	←	

Analysing the Program by Hand (2)

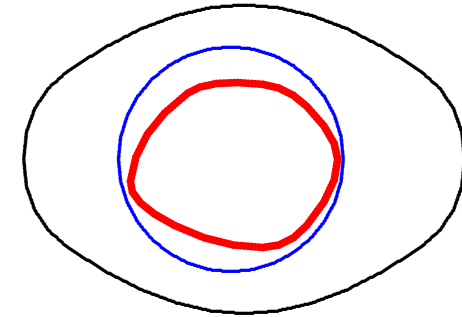
<code>[y := x]¹;</code>	←	$\{(x, ?), (y, ?), (z, ?)\}$
<code>[z := 1]²;</code>	←	$\{(x, ?), (y, 1), (z, ?)\}$
<code>while [y > 0]³ do</code>	←	$\{(x, ?), (y, 1), (z, 2)\} \cup \{(y, 5), (z, 4)\}$
<code>[z := z * y]⁴;</code>	←	$\{(x, ?), (y, 1), (z, 2)\}$
<code>[y := y - 1]⁵</code>	←	$\{(x, ?), (y, 1), (z, 4)\}$
<code>od;</code>	←	$\{(x, ?), (y, 5), (z, 4)\}$
<code>[y := 0]⁶</code>	←	$\{(x, ?), (y, 1), (z, 2)\}$
	←	

Analysing the Program by Hand (3)

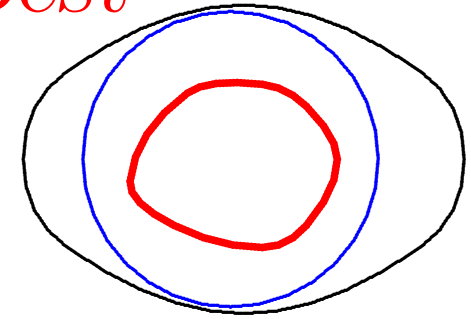
$[y := x]^1;$	←	$\{(x, ?), (y, ?), (z, ?)\}$
$[z := 1]^2;$	←	$\{(x, ?), (y, 1), (z, ?)\}$
while $[y > 0]^3$ do	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\} \cup \{(y, 5), (z, 4)\}$
$[z := z * y]^4;$	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\}$
$[y := y - 1]^5$	←	$\{(x, ?), (y, 1), (y, 5), (z, 4)\}$
od;	←	$\{(x, ?), (y, 5), (z, 4)\}$
$[y := 0]^6$	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\}$
	←	$\{(x, ?), (y, 6), (z, 2), (z, 4)\}$

The Best Solution

<code>[y := x]¹;</code>	←	$\{(x, ?), (y, ?), (z, ?)\}$
<code>[z := 1]²;</code>	←	$\{(x, ?), (y, 1), (z, ?)\}$
<code>while [y > 0]³ do</code>	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\}$
<code>[z := z * y]⁴;</code>	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\}$
<code>[y := y - 1]⁵</code>	←	$\{(x, ?), (y, 1), (y, 5), (z, 4)\}$
<code>od;</code>	←	$\{(x, ?), (y, 5), (z, 4)\}$
<code>[y := 0]⁶</code>	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\}$
	←	$\{(x, ?), (y, 6), (z, 2), (z, 4)\}$

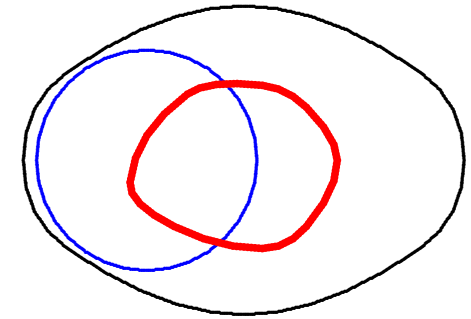


A Safe Solution — but not the Best



<code>[y := x]¹;</code>	←	$\{(x, ?), (y, ?), (z, ?)\}$
<code>[z := 1]²;</code>	←	$\{(x, ?), (y, 1), (z, ?)\}$
<code>while [y > 0]³ do</code>	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\}$
<code>[z := z * y]⁴;</code>	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\}$
<code>[y := y - 1]⁵</code>	←	$\{(x, ?), (y, 1), (y, 5), (z, 2), (z, 4)\}$
<code>od;</code>	←	$\{(x, ?), (y, 1), (y, 5), (z, 2), (z, 4)\}$
<code>[y := 0]⁶</code>	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\}$
	←	$\{(x, ?), (y, 6), (z, 2), (z, 4)\}$

An Unsafe Solution

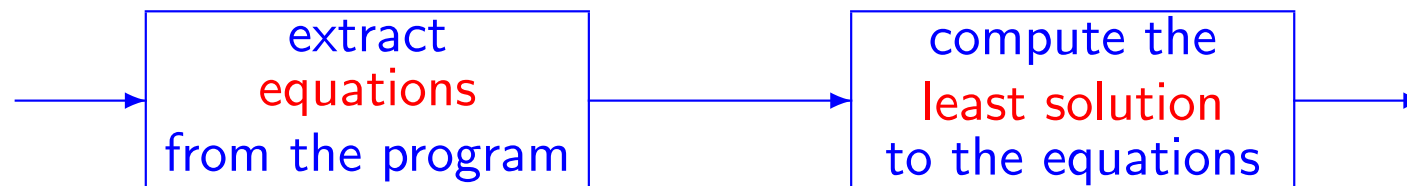


<code>[y := x]¹;</code>	←	$\{(x, ?), (y, ?), (z, ?)\}$
<code>[z := 1]²;</code>	←	$\{(x, ?), (y, 1), (z, ?)\}$
<code>while [y > 0]³ do</code>	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\}$
<code>[z := z * y]⁴;</code>	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\}$
<code>[y := y - 1]⁵</code>	←	$\{(x, ?), (y, 1), (y, 5), (z, 4)\}$
<code>od;</code>	←	$\{(x, ?), (y, 5), (z, 4)\}$
<code>[y := 0]⁶</code>	←	$\{(x, ?), (y, 1), (z, 2), (y, 5), (z, 4)\}$
	←	$\{(x, ?), (y, 6), (z, 2), (z, 4)\}$

Formalising the Development

Problem: which definitions reach which program points

The assignment $[x := a]^{\ell}$ reaches ℓ' if there is an execution where x was last assigned at ℓ



- the programming language of interest
- abstract flow graphs
- extract and solve the equations

The WHILE Language:

$$a ::= x \mid n \mid a_1 \ op_a \ a_2$$
$$b ::= \text{true} \mid \text{false} \mid \text{not } b \mid b_1 \ op_b \ b_2 \mid a_1 \ op_r \ a_2$$
$$S ::= [x := a]^\ell \mid [\text{skip}]^\ell \mid S_1; S_2 \mid \\ \text{if } [b]^\ell \text{ then } S_1 \text{ else } S_2 \mid \text{while } [b]^\ell \text{ do } S \text{ od}$$

Assignments and tests are (uniquely) labelled to allow the analyses to refer to these program fragments — the labels corresponds to pointers into the syntax tree.

We use abstract syntax and insert paranthesis to disambiguate the syntax — again this corresponds to thinking in terms of syntax trees.

OBS: the control structure is known — there is no need for a control flow analysis!

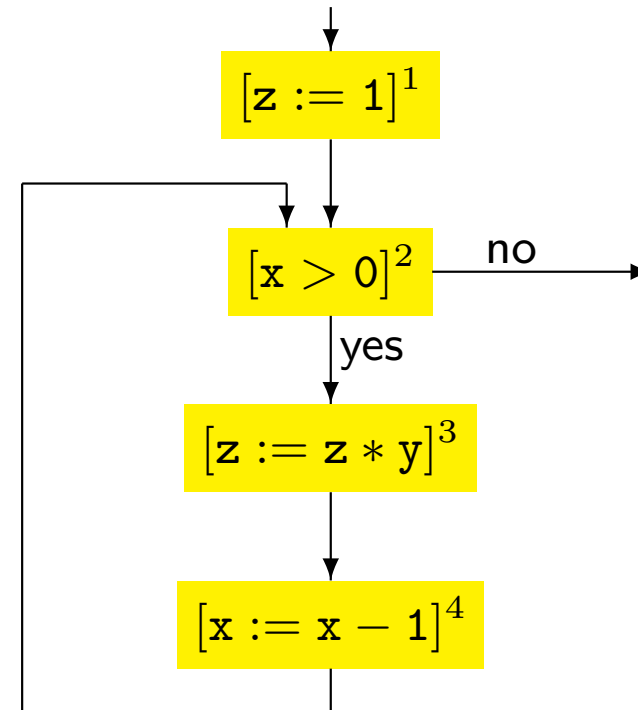
Talking about Flow Graphs

- $\text{labels}(S)$ the **set of nodes** of the flow graphs of S
- $\text{init}(S)$ the **initial node** of the flow graph of S ; this is the unique node where the execution of the program starts
- $\text{final}(S)$ the **final nodes** of the flow graphs for S ; this is the set of nodes where the execution of the program may terminate
- $\text{flow}(S)$ the **edges** of the flow graphs for S ; there is an edge from one node to another if the flow of control may go from the first node to the second — used for **forward** analyses
- $\text{flow}^R(S)$ the **reversed edges** of the flow graphs of S ; there is an edge from one node to another if the flow of control may go from the second node to the first — used for **backwards** analyses

The Abstract Flow Graph

Example: $[z:=1]^1; \text{while } [x>0]^2 \text{ do } [z:=z*y]^3; [x:=x-1]^4 \text{ od}$

$\text{labels}(\dots) = \{1, 2, 3, 4\}$
 $\text{init}(\dots) = 1$
 $\text{final}(\dots) = \{2\}$
 $\text{flow}(\dots) = \{(1, 2), (2, 3), (3, 4), (4, 2)\}$
 $\text{flow}^R(\dots) = \{(2, 1), (2, 4), (3, 2), (4, 3)\}$



Computing the Information (1)

S	$\text{labels}(S)$	$\text{init}(S)$	$\text{final}(S)$
$[x := a]^\ell$	$\{\ell\}$	ℓ	$\{\ell\}$
$[\text{skip}]^\ell$	$\{\ell\}$	ℓ	$\{\ell\}$
$S_1; S_2$	$\text{labels}(S_1) \cup \text{labels}(S_2)$	$\text{init}(S_1)$	$\text{final}(S_2)$
$\text{if } [b]^\ell \text{ then } S_1$ $\quad \text{else } S_2$	$\{\ell\} \cup \text{labels}(S_1)$ $\quad \cup \text{labels}(S_2)$	ℓ	$\text{final}(S_1)$ $\cup \text{final}(S_2)$
$\text{while } [b]^\ell \text{ do } S \text{ od}$	$\{\ell\} \cup \text{labels}(S)$	ℓ	$\{\ell\}$

Computing the Information (2)

S	$\text{flow}(S)$	$\text{blocks}(S)$
$[x := a]^\ell$	\emptyset	$\{[x := a]^\ell\}$
$[\text{skip}]^\ell$	\emptyset	$\{[\text{skip}]^\ell\}$
$S_1; S_2$	$\text{flow}(S_1) \cup \text{flow}(S_2)$ $\cup \{(\ell, \text{init}(S_2)) \mid \ell \in \text{final}(S_1)\}$	$\text{blocks}(S_1)$ $\cup \text{blocks}(S_2)$
$\text{if } [b]^\ell \text{ then } S_1$ $\quad \text{else } S_2$	$\text{flow}(S_1) \cup \text{flow}(S_2)$ $\cup \{(\ell, \text{init}(S_1)), (\ell, \text{init}(S_2))\}$	$\{[b]^\ell\} \cup \text{blocks}(S_1)$ $\cup \text{blocks}(S_2)$
$\text{while } [b]^\ell \text{ do } S \text{ od}$	$\{(\ell, \text{init}(S))\} \cup \text{flow}(S)$ $\cup \{(\ell', \ell) \mid \ell' \in \text{final}(S)\}$	$\{[b]^\ell\} \cup \text{blocks}(S)$

$$\text{flow}^R(S) = \{(\ell, \ell') \mid (\ell', \ell) \in \text{flow}(S)\}$$

Simplifying assumptions

The program of interest S_\star is often assumed to satisfy:

- S_\star has **isolated entries** if there are no edges leading into $\text{init}(S_\star)$:

$$\forall \ell : (\ell, \text{init}(S_\star)) \notin \text{flow}(S_\star)$$

- S_\star has **isolated exits** if there are no edges leading out of labels in $\text{final}(S_\star)$:

$$\forall \ell \in \text{final}(S_\star), \forall \ell' : (\ell, \ell') \notin \text{flow}(S_\star)$$

- S_\star is **label consistent** if

$$\forall [B_1^{\ell_1}], [B_2^{\ell_2}] \in \text{blocks}(S_\star) : \ell_1 = \ell_2 \Rightarrow B_1 = B_2$$

This holds if S_\star is uniquely labelled.

Reaching Definitions Analysis

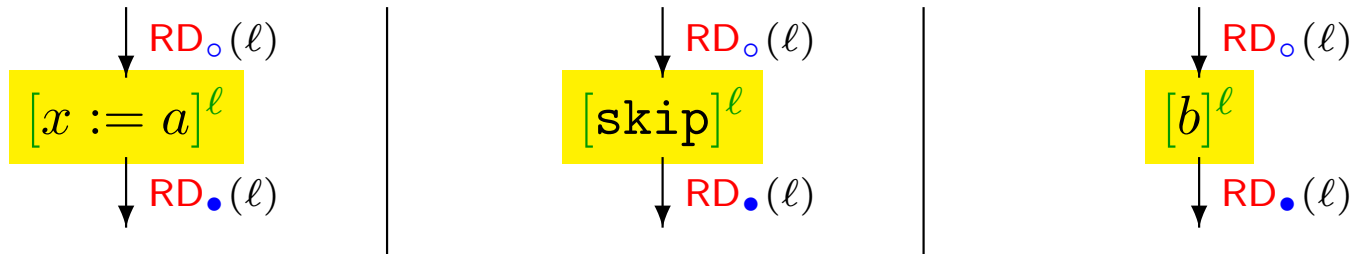
The aim of the **Reaching Definitions Analysis** is to determine

For each program point, which assignments may have been made and not overwritten, when program execution reaches this point along some path.

Example

$[y := x]^1; [z := 1]^2; \text{while } [y > 0]^3 \text{ do } [z := z * y]^4; [y := y - 1]^5 \text{ od}; [y := 0]^6$

The Basic Idea



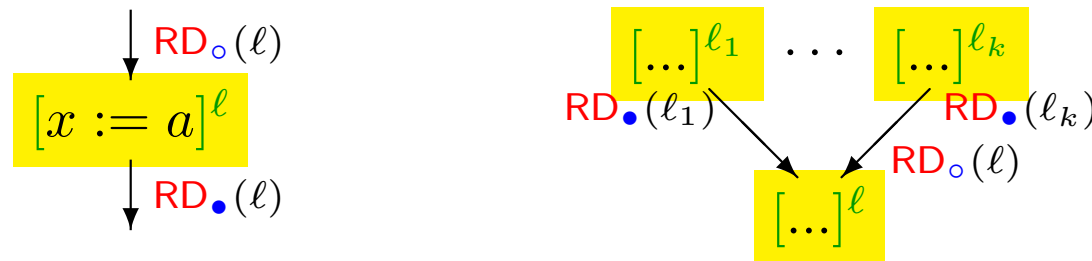
Analysis information:

- $RD_\circ(\ell)$: the definitions that reach the **entry** of block ℓ
- $RD_\bullet(\ell)$: the definitions that reach the **exit** of block ℓ

Auxiliary information:

- $\text{kill}_{RD}(\ell)$: the definitions that are **killed** by an elementary block
- $\text{gen}_{RD}(\ell)$: the definitions that are **generated** by an elementary block

Analysis of the program



$$RD_{\bullet}(\ell) = (RD_{\circ}(\ell) \setminus \text{kill}_{RD}(B^{\ell})) \cup \text{gen}_{RD}(B^{\ell}) \quad \text{if } B^{\ell} \in \text{blocks}(S_{\star})$$

$$RD_{\circ}(\ell) = \begin{cases} \{(x, ?) \mid x \in FV(S_{\star})\} & \text{if } \ell = \text{init}(S_{\star}) \\ \cup \{RD_{\bullet}(\ell') \mid (\ell', \ell) \in \text{flow}(S_{\star})\} & \text{otherwise} \end{cases}$$

$$\text{kill}_{RD}([x := a]^{\ell}) = \{(x, ?)\} \cup \{(x, \ell') \mid B^{\ell'} \text{ is an assignment to } x\}$$

$$\text{gen}_{RD}([x := a]^{\ell}) = \{(x, \ell)\}$$

$$\text{kill}_{RD}(B^{\ell}) = \text{gen}_{RD}(B^{\ell}) = \emptyset \quad \text{otherwise}$$

Example

$[y := x]^1; [z := 1]^2; \text{while } [y > 0]^3 \text{ do } [z := z * y]^4; [y := y - 1]^5 \text{ od}; [y := 0]^6$

Equations:

$$RD_{\bullet}(1) = RD_{\circ}(1) \setminus \{(y, ?), (y, 1), (y, 5), (y, 6)\} \cup \{(y, 1)\}$$

$$RD_{\bullet}(2) = RD_{\circ}(2) \setminus \{(z, ?), (z, 2), (z, 4)\} \cup \{(z, 2)\}$$

$$RD_{\bullet}(3) = RD_{\circ}(3) \setminus \emptyset \cup \emptyset$$

$$RD_{\bullet}(4) = RD_{\circ}(4) \setminus \{(z, ?), (z, 2), (z, 4)\} \cup \{(z, 4)\}$$

$$RD_{\bullet}(5) = RD_{\circ}(5) \setminus \{(y, ?), (y, 1), (y, 5), (y, 6)\} \cup \{(y, 5)\}$$

$$RD_{\bullet}(6) = RD_{\circ}(6) \setminus \{(y, ?), (y, 1), (y, 5), (y, 6)\} \cup \{(y, 6)\}$$

$$RD_{\circ}(1) = \{(x, ?), (y, ?), (z, ?)\}$$

$$RD_{\circ}(2) = RD_{\bullet}(1)$$

$$RD_{\circ}(3) = RD_{\bullet}(2) \cup RD_{\bullet}(5)$$

$$RD_{\circ}(4) = RD_{\bullet}(3)$$

$$RD_{\circ}(5) = RD_{\bullet}(4)$$

$$RD_{\circ}(6) = RD_{\bullet}(3)$$

Compute the least solution to these equations.

Solving the equations

1. $W := \text{nil};$

for all (ℓ, ℓ') in $\text{flow}(S_\star)$ do $W := \text{cons}((\ell, \ell'), W);$

for all ℓ in $\text{labels}(S_\star)$ do

if $\ell = \text{init}(S_\star)$ then $\text{RD}_\circ(\ell) := \{(x, ?) \mid x \in \text{FV}(S_\star)\}$ else $\text{RD}_\circ(\ell) := \emptyset$

2. while $W \neq \text{nil}$ do

$(\ell, \ell') := \text{head}(W); W := \text{tail}(W);$

if $\underbrace{(\text{RD}_\circ(\ell) \setminus \text{kill}_{\text{RD}}(\ell)) \cup \text{gen}_{\text{RD}}(\ell)}_{\text{RD}_\bullet(\ell)} \not\subseteq \text{RD}_\circ(\ell')$

then $\text{RD}_\circ(\ell') := \text{RD}_\circ(\ell') \cup \underbrace{(\text{RD}_\circ(\ell) \setminus \text{kill}_{\text{RD}}(\ell)) \cup \text{gen}_{\text{RD}}(\ell)}_{\text{RD}_\bullet(\ell)};$

for all ℓ'' with (ℓ', ℓ'') in $\text{flow}(S_\star)$ do $W := \text{cons}((\ell', \ell''), W);$