Sebastian Hack, Christian Hammer, Jan Reineke

Saarland University

Static Program Analysis

Introduction

Winter Semester 2014

Slides based on:

- H. Seidl, R. Wilhelm, S. Hack: Compiler Design, Volume 3, Analysis and Transformation, Springer Verlag, 2012
- F. Nielson, H. Riis Nielson, C. Hankin: Principles of Program Analysis, Springer Verlag, 1999
- R. Wilhelm, B. Wachter: Abstract Interpretation with Applications to Timing Validation. CAV 2008: 22-36
- Helmut Seidl's slides

A Short History of Static Program Analysis

- Early high-level programming languages were implemented on very small and very slow machines.
- Compilers needed to generate executables that were extremely efficient in space and time.
- Compiler writers invented efficiency-increasing program transformations, wrongly called optimizing transformations.
- Transformations must not change the semantics of programs.
- Enabling conditions guaranteed semantics preservation.
- Enabling conditions were checked by static analysis of programs.

Theoretical Foundations of Static Program Analysis

- Theoretical foundations for the solution of recursive equations: Kleene (30s), Tarski (1955)
- Gary Kildall (1972) clarified the lattice-theoretic foundation of data-flow analysis.
- Patrick Cousot (1974) established the relation to the programming-language semantics.

Static Program Analysis as a Verification Method

- Automatic method to derive invariants about program behavior, answers questions about program behavior:
  - will index always be within bounds at program point p?
  - will memory access at p always hit the cache?
- answers of sound static analysis are correct, but approximate: don't know is a valid answer!
- analyses proved correct wrt. language semantics,

1 Introduction

a simple imperative programming language with:

• variables	//	registers
• $R = e;$	//	assignments
• $R = M[e];$	//	loads
• $M[e_1] = e_2;$	//	stores
• if $(e)\ s_1$ else s_2	//	conditional branching
• goto $L;$	//	no loops

An intermediate language into which (almost) everything can be translated. In particular, no procedures. So, only intra-procedural analyses!

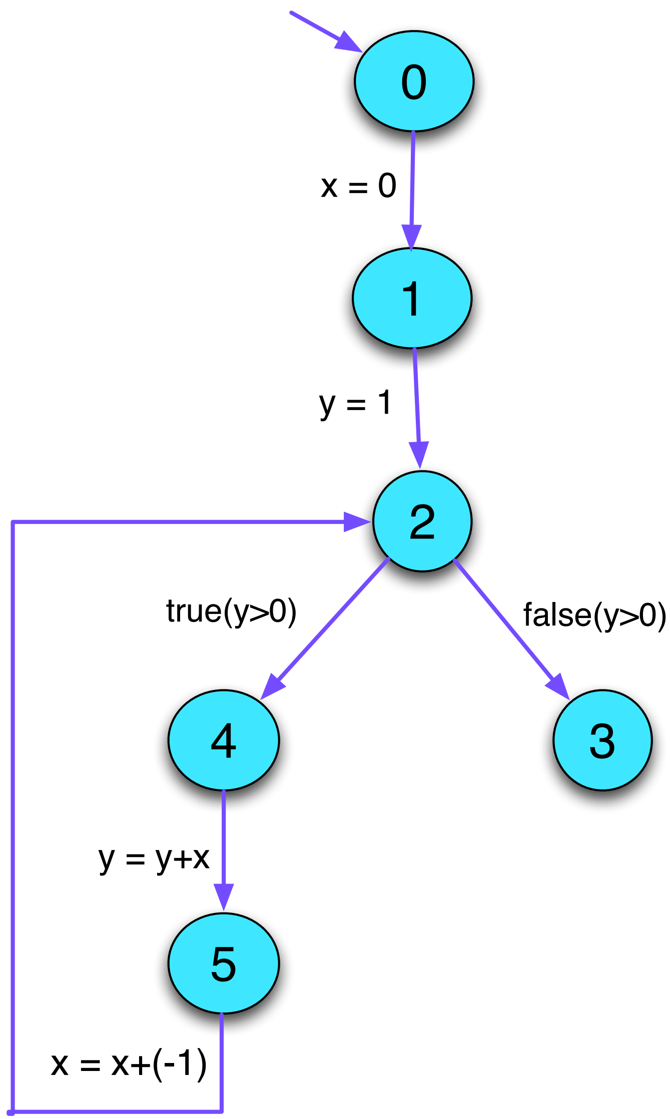
2 Example — Rules-of-Sign Analysis

Problem: Determine at each program point the sign of the values of all variables of numeric type.

Example program:

```
x = 0;
y = 1;
while (y > 0) do
      y = y + x;
      x = x + (-1);
```

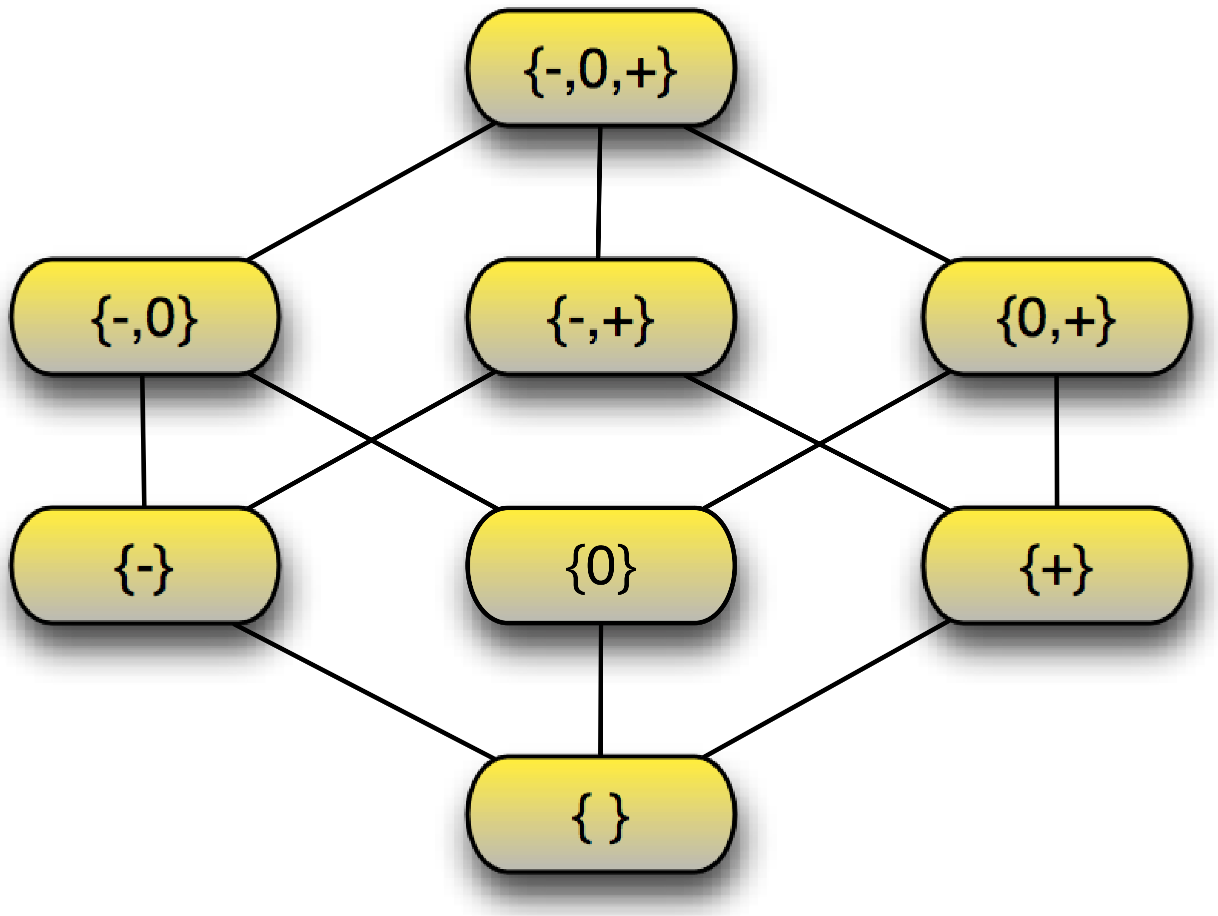
Program representation as *control-flow graphs*



We need the following ingredients:

- a set of information elements, each a set of possible signs,
- a partial order, “$\sqsubseteq$”, on these elements, specifying the ”relative strength” of two information elements,
- these together form the abstract domain, a lattice,
- functions describing how signs of variables change by the execution of a statement, abstract edge effects,
- these need an abstract arithmetic, an arithmetic on signs.

We construct the abstract domain for single variables starting with the lattice $\quad Signs = 2^{\{-,0,+\}}$ with the relation "$\sqsubseteq$" ="$\subseteq$".



The analysis should ”bind” program variables to elements in $Signs$.

So, the abstract domain is $\mathbb{D} = (\, Vars \to Signs)_{\perp}$, a Sign-environment.

$\perp \in \mathbb{D}$ is the function mapping all arguments to $\{\}$.

The partial order on $\mathbb{D}$ is $\quad D_1 \sqsubseteq D_2 \quad$ iff

$D_1 = \perp \qquad$ or

$D_1\, x \subseteq D_2\, x \qquad (x \in Vars)$

Intuition?

The analysis should "bind" program variables to elements in $Signs$.

So, the abstract domain is $\mathbb{D} = (\,Vars \to Signs)_\perp$. a Sign-environment.

$\perp \in \mathbb{D}$ is the function mapping all arguments to $\{\}$.

The partial order on $\mathbb{D}$ is $\qquad D_1 \sqsubseteq D_2 \quad$ iff

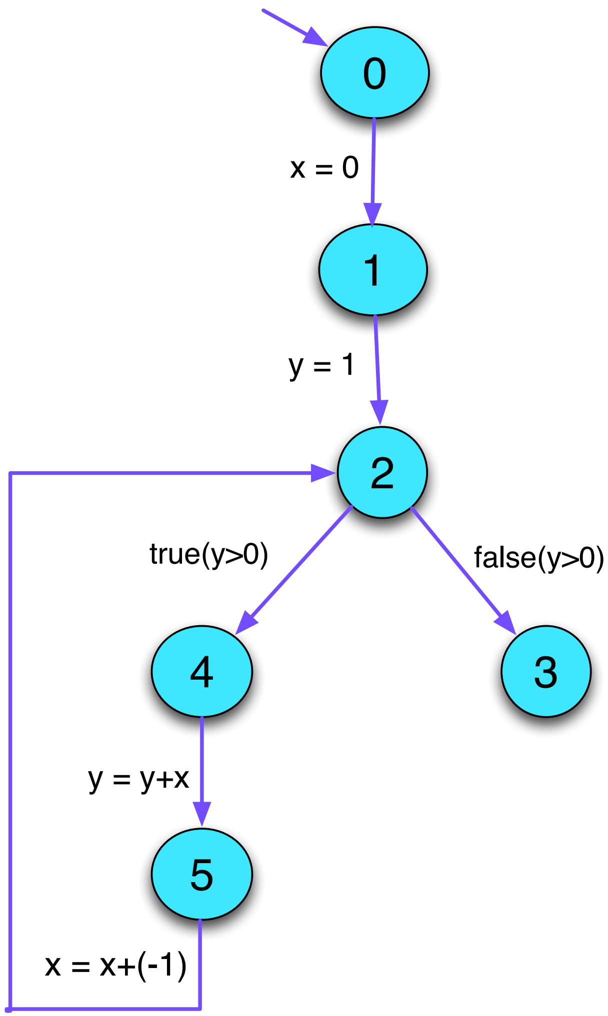
$D_1 = \perp \qquad\qquad$ or

$D_1\,x \subseteq D_2\,x \qquad (x \in Vars)$

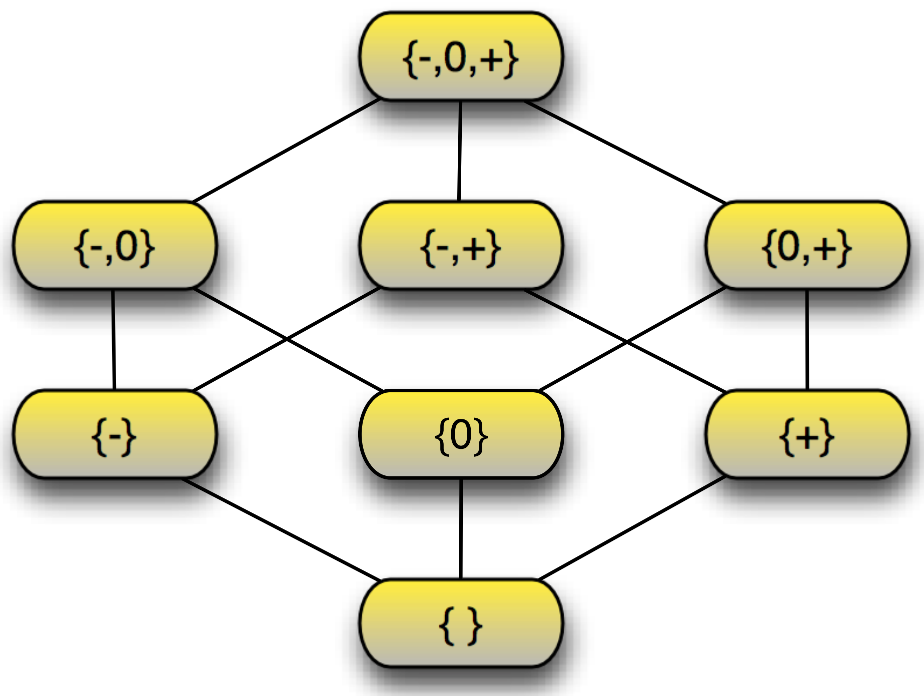
Intuition?

D_1 is at least as precise as D_2 since D_2 admits at least as many signs as D_1

How did we analyze the program?

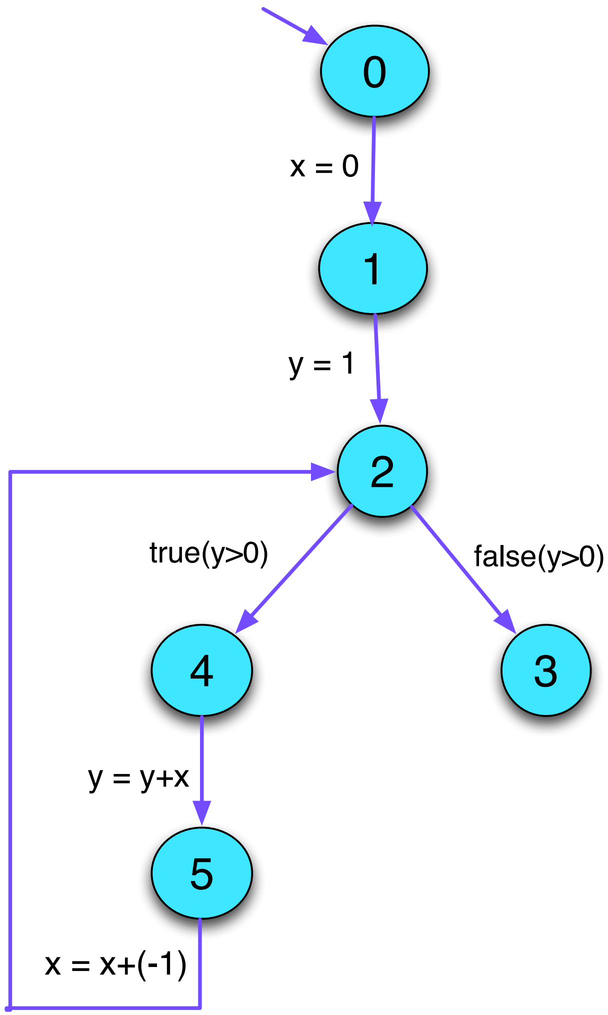


In particular, how did we walk the lattice for y at program point 5?



How is a solution found?

Iterating until a fixed-point is reached



0		1		2		3		4		5	
x	y	x	y	x	y	x	y	x	y	x	y

Idea:

- We want to determine the sign of the values of expressions.

Idea:

- We want to determine the sign of the values of expressions.
- For some sub-expressions, the analysis may yield $\{+, -, 0\}$, which means, it couldn't find out.

Idea:

- We want to determine the signs of the values of expressions.
- For some sub-expressions, the analysis may yield $\{+, -, 0\}$, which means, it couldn't find out.
- We replace the concrete operators $\Box$ working on values by abstract operators $\Box^\sharp$ working on signs:

Idea:

- We want to determine the signs of the values of expressions.
- For some sub-expressions, the analysis may yield $\{+, -, 0\}$, which means, it couldn't find out.
- We replace the concrete operators $\Box$ working on values by abstract operators $\Box^\sharp$ working on signs:
- The abstract operators allow to define an abstract evaluation of expressions:

$$[\![e]\!]^\sharp \ : \ (\mathit{Vars} \to \mathit{Signs}) \to \mathit{Signs}$$

Determining the sign of expressions in a Sign-environment works as follows:

$$
\begin{array}{lcl}
[\![c]\!]^\sharp\, D & = & \begin{cases} \{+\} & \text{if c > 0} \\ \{-\} & \text{if c < 0} \\ \{0\} & \text{if c = 0} \end{cases} \\
[\![v]\!]^\sharp & = & D(v) \\
[\![e_1 \,\square\, e_2]\!]^\sharp\, D & = & [\![e_1]\!]^\sharp\, D \,\square^\sharp\, [\![e_2]\!]^\sharp\, D \\
[\![\square e]\!]^\sharp\, D & = & \square^\sharp [\![e]\!]^\sharp\, D
\end{array}
$$

Abstract operators working on signs (Addition)

$+^{\#}$	{0}	{+}	{-}	{-, 0}	{-, +}	{0, +}	{-, 0, +}
{0}	{0}	{+}					
{+}							
{-}							
{-, 0}							
{-, +}							
{0, +}							
{-, 0, +}	{-, 0, +}						

Abstract operators working on signs (Multiplication)

$\times^{\#}$	{0}	{+}	{-}	{-, 0}	{-, +}	{0, +}	{-, 0, +}
{0}	{0}	{0}					
{+}							
{-}							
{-, 0}							
{-, +}							
{0, +}							
{-, 0, +}	{0}						

Abstract operators working on signs (unary minus)

$-^{\#}$	{0}	{+}	{-}	{-, 0}	{-, +}	{0, +}	{-, 0, +}
	{0}	{-}	{+}	{+, 0}	{-, +}	{0, -}	{-, 0, +}

Working an example: $D = \{x \mapsto \{+\}, y \mapsto \{+\}\}$

$$
\begin{array}{lcl}
[\![x+7]\!]^\sharp\, D & = & [\![x]\!]^\sharp\, D \;+^\sharp\; [\![7]\!]^\sharp\, D \\
 & = & \{+\} \;+^\sharp\; \{+\} \\
 & = & \{+\} \\
[\![x+(-y)]\!]^\sharp\, D & = & \{+\} \;+^\sharp\; (-^\sharp [\![y]\!]^\sharp\, D\,) \\
 & = & \{+\} \;+^\sharp\; (-^\sharp \{+\}) \\
 & = & \{+\} \;+^\sharp\; \{-\} \\
 & = & \{+, -, 0\}
\end{array}
$$

$[\![lab]\!]^\sharp$ is the abstract edge effects associated with edge k.

It depends only on the label *lab*:

$$
\begin{array}{lll}
[\![;]\!]^\sharp\, D & = & D \\
[\![\text{true}\,(e)]\!]^\sharp\, D & = & D \\
[\![\text{false}\,(e)]\!]^\sharp\, D & = & D \\
[\![x = e;]\!]^\sharp\, D & = & D \oplus \{x \mapsto [\![e]\!]^\sharp\, D\} \\
[\![x = M[e];]\!]^\sharp\, D & = & D \oplus \{x \mapsto \{+, -, 0\}\} \\
[\![M[e_1] = e_2;]\!]^\sharp\, D & = & D
\end{array}
$$

... whenever $\qquad D \neq \bot$

These edge effects can be composed to the effect of a path $\pi = k_1 \ldots k_r$:

$$[\![\pi]\!]^\sharp = [\![k_r]\!]^\sharp \circ \ldots \circ [\![k_1]\!]^\sharp$$

Consider a program node v:

- $\rightarrow$ For every path π from program entry *start* to v the analysis should determine for each program variable x the set of all signs that the values of x may have at v as a result of executing π.
- $\rightarrow$ Initially at program start, no information about signs is available.
- $\rightarrow$ The analysis computes a superset of the set of signs as safe information.

$\Longrightarrow$ For each node v, we need the set:

$$\mathcal{S}[v] \quad = \quad \bigcup\{[\![\pi]\!]^\sharp\bot \mid \pi : \mathit{start} \rightarrow^* v\}$$

Question:

How do we compute $\mathcal{S}[u]$ for every program point u?

Question:

How can we compute $\mathcal{S}[u]$ for every program point u?

Collect all constraints on the values of $\mathcal{S}[u]$ into a system of constraints:

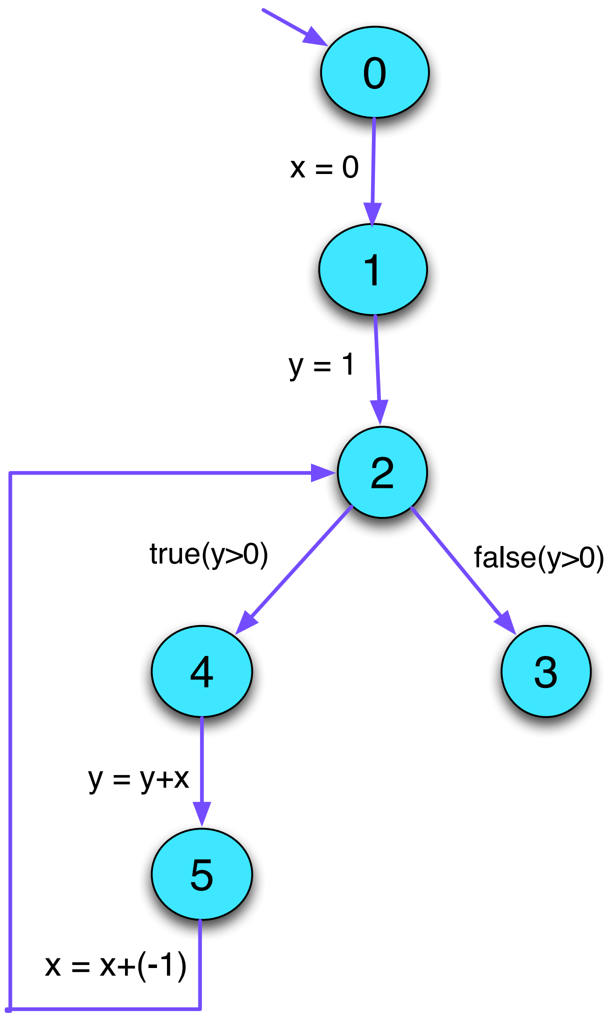
$$
\begin{array}{lcll}
\mathcal{S}[start] & \supseteq & \perp & \\
\mathcal{S}[v] & \supseteq & [\![k]\!]^\sharp\,(\mathcal{S}[u]) & \qquad k = (u, \_, v) \quad \text{edge}
\end{array}
$$

Why $\supseteq$?

Wanted:

- a least solution (why least?)
- an algorithm that computes this solution

Example:



$$\begin{array}{lcl}
\mathcal{S}[0] & \supseteq & \bot \\
\mathcal{S}[1] & \supseteq & \mathcal{S}[0] \oplus \{x \mapsto \{0\}\} \\
\mathcal{S}[2] & \supseteq & \mathcal{S}[1] \oplus \{y \mapsto \{+\}\} \\
\mathcal{S}[2] & \supseteq & \mathcal{S}[5] \oplus \{x \mapsto [\![x + (-1)]\!]^\sharp\, \mathcal{S}[5]\} \\
\mathcal{S}[3] & \supseteq & \mathcal{S}[2] \\
\mathcal{S}[4] & \supseteq & \mathcal{S}[2] \\
\mathcal{S}[5] & \supseteq & \mathcal{S}[4] \oplus \{y \mapsto [\![y + x]\!]^\sharp\, \mathcal{S}[4]\}
\end{array}$$