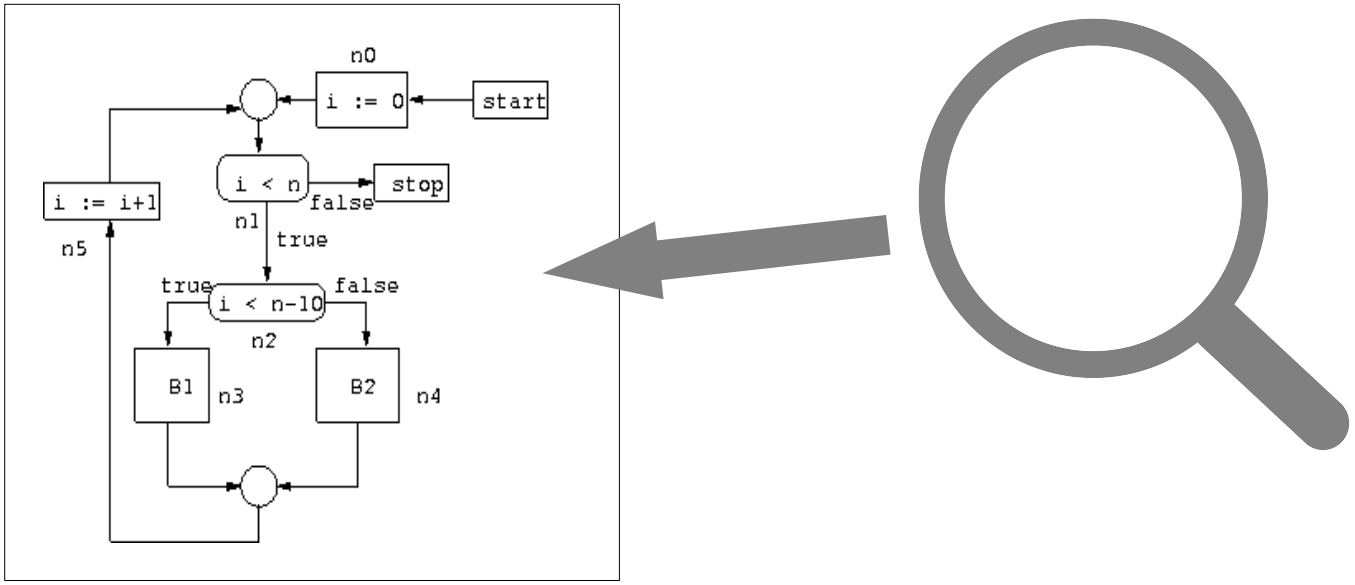
Static Program Analysis
Lecture 4: Dataflow Analysis



Dataflow Analysis

Computes data flow relations between statements in imperative programs

“Classical” family of program analyses, the first to appear

Routinely used in compilers, to find opportunities for code optimization

Can also be used to detect some kinds of bugs

A Dataflow Analysis Example

Reaching Definitions: when a variable is assigned a value, where is that value possibly “active” (not surely overwritten)?

For each point in the program, the analysis computes a set of *reaching definitions* (`x`,`l`)

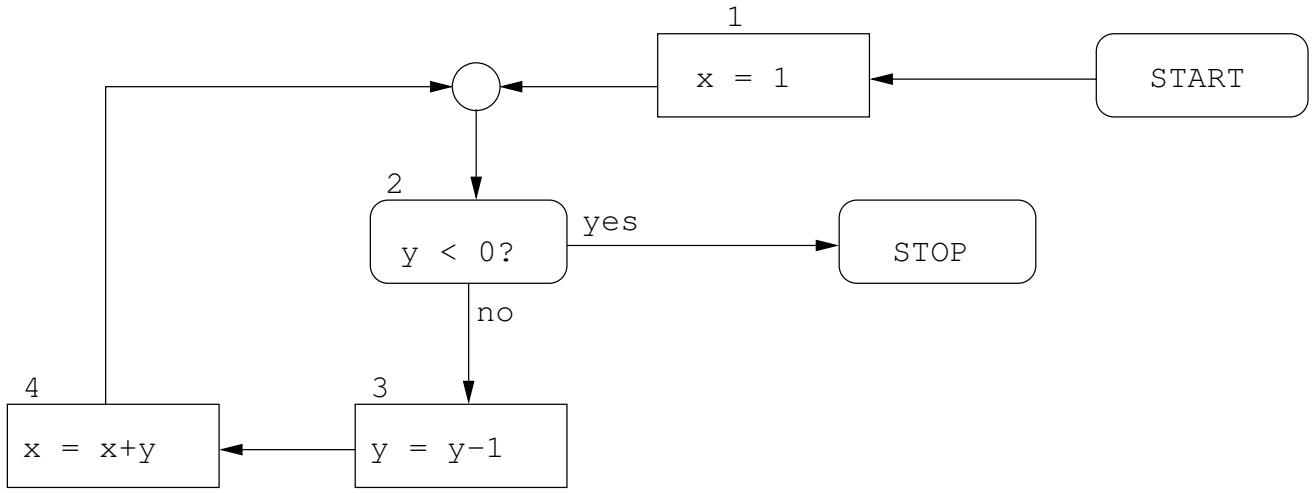
`l` is the location of a statement where `x` is assigned

If the set in program point `p` contains (`x`,`l`), then the assignment of `x` at `l` may *reach* `p`

This information can be used by optimizing compilers

Also to check for possible use of unassigned variables (potential bug)

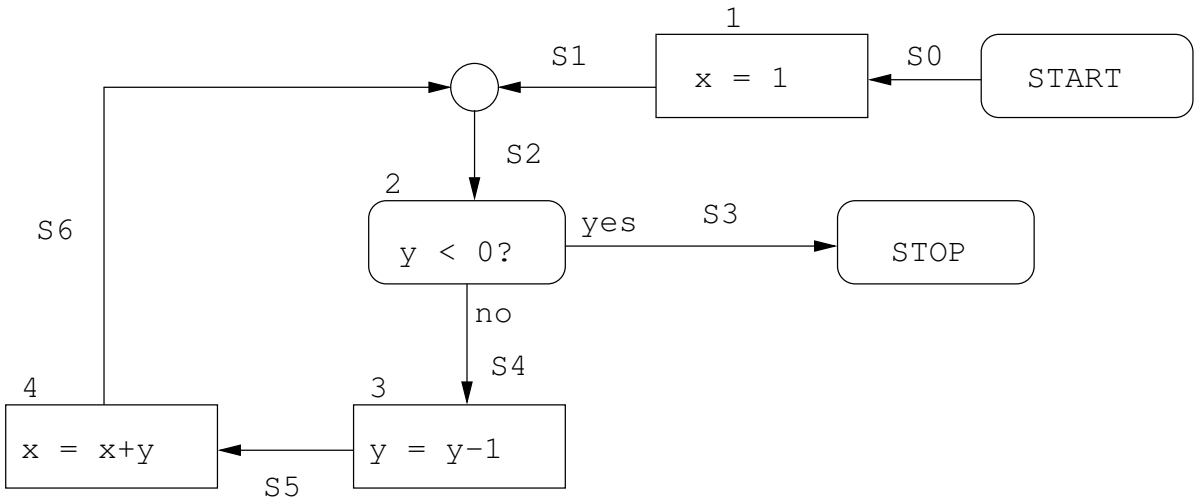
Example Program



We use Control-Flow Graph (CFG) program representation

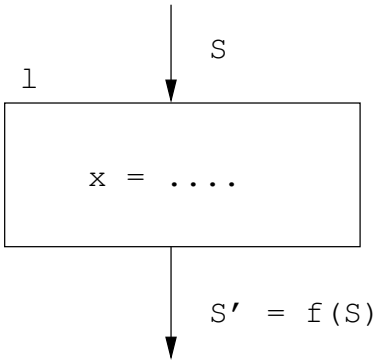
Program points $=$ edges in the CFG (places where you end up after having executed a statement)

Equations for Reaching Definitions Analysis I



For each program point a set of reaching definitions: $S_0, \ldots, S_6$

Equations for Reaching Definitions Analysis II



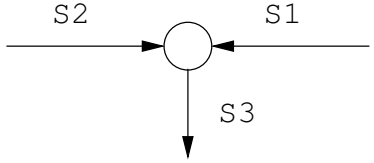
Sets are related through *equations* $S' = f(S)$. For assignments,

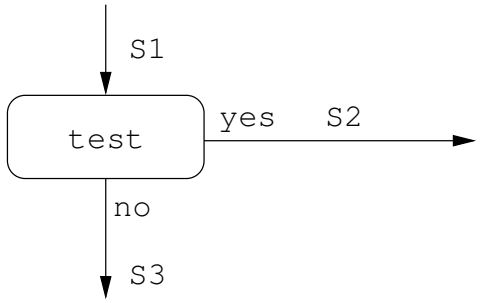
$$f(S) = (S \setminus \{ (\mathtt{x}, \mathtt{l}') \mid \mathtt{l}' \in L \}) \cup \{(\mathtt{x}, \mathtt{l})\}$$

All "old" definitions $(\mathtt{x}, \mathtt{l}')$ are overwritten, and $(\mathtt{x}, \mathtt{l})$ is added

Equations for Reaching Definitions Analysis III

Join nodes, and test nodes:



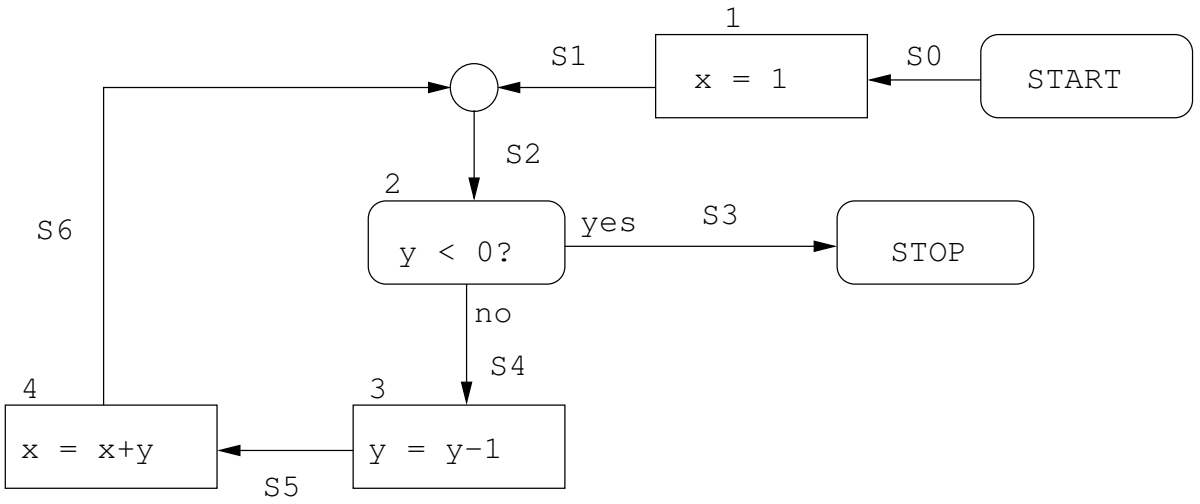


$S_3 = S_1 \cup S_2$

$S_3 = S_1, S_2 = S_1$

Equations for Reaching Definitions Analysis IV

$$
\begin{aligned}
S_0 &= \{(\mathtt{x}, \mathtt{?}), (\mathtt{y}, \mathtt{?})\} \\
S_1 &= f_1(S_0) \\
S_2 &= S_1 \cup S_6 \\
S_3 &= S_2 \\
S_4 &= S_2 \\
S_5 &= f_3(S_4) \\
S_6 &= f_4(S_5)
\end{aligned}
$$

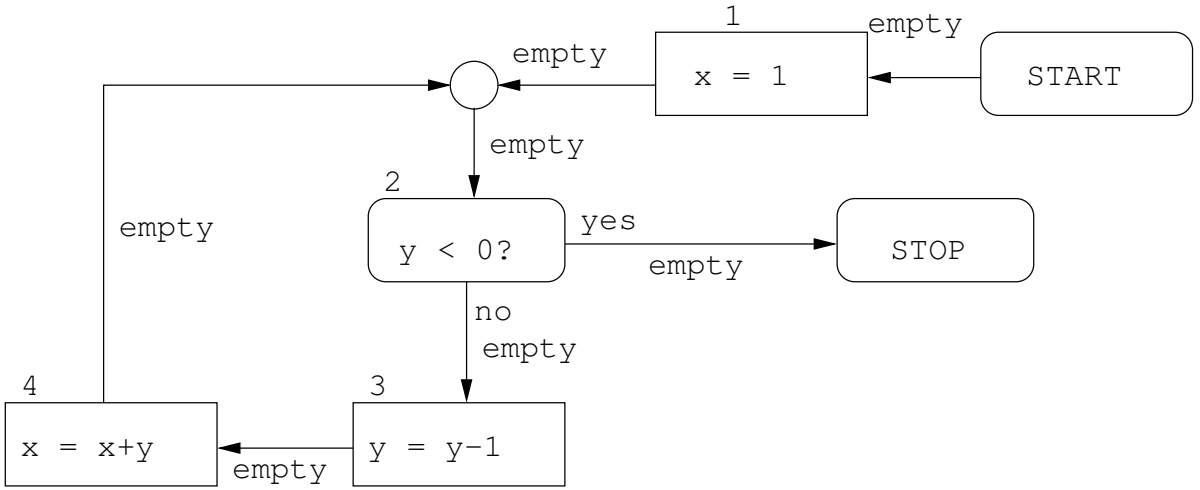


A system of set equations relating $S_0, \ldots, S_6$

Initial definitions $(\mathtt{x},\mathtt{?})$, $(\mathtt{y},\mathtt{?})$ for $\mathtt{x}$ and $\mathtt{y}$, where the program starts

Solving the Equations

$$
\begin{aligned}
S_{0,0} &= \emptyset \\
S_{1,0} &= \emptyset \\
S_{2,0} &= \emptyset \\
S_{3,0} &= \emptyset \\
S_{4,0} &= \emptyset \\
S_{5,0} &= \emptyset \\
S_{6,0} &= \emptyset
\end{aligned}
$$



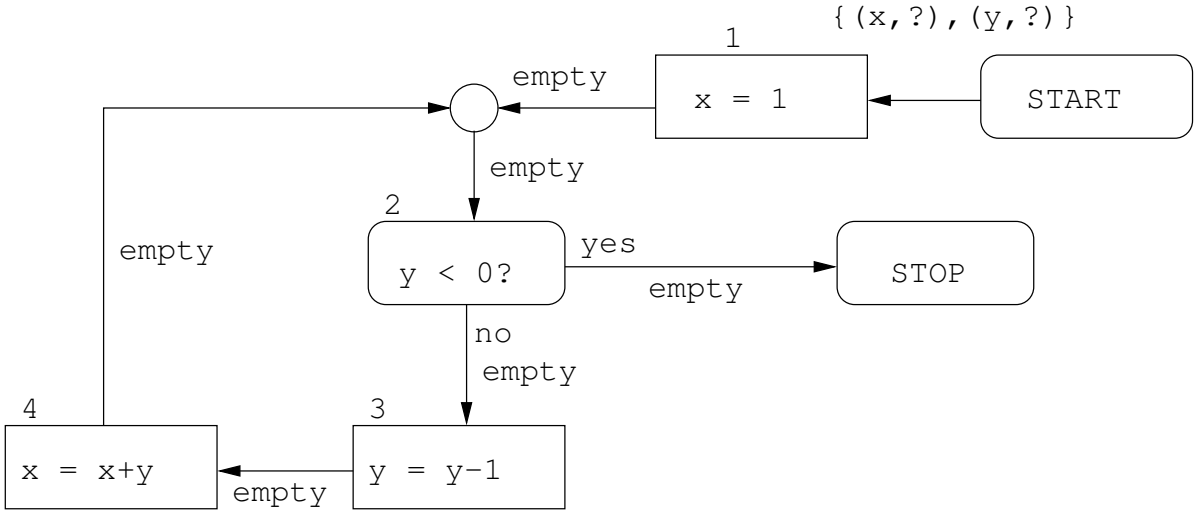
We use *least fixed-point iteration*

Start iteration with *least possible* sets ($\emptyset$), then iterate the equations until no set changes anymore (sets to check for change marked red)

Iteration 1

Update using equation $S_0 = \{(\mathtt{x}, ?), (\mathtt{y}, ?)\}$:

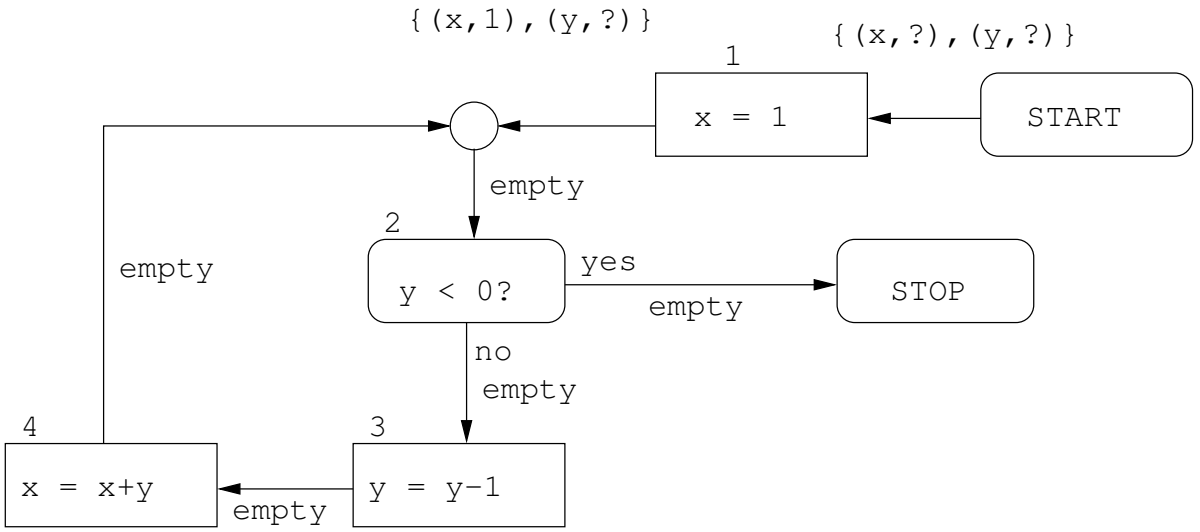
$$
\begin{aligned}
S_{0,1} &= \{(\mathtt{x}, ?), (\mathtt{y}, ?)\} \\
S_{1,0} &= \emptyset \\
S_{2,0} &= \emptyset \\
S_{3,0} &= \emptyset \\
S_{4,0} &= \emptyset \\
S_{5,0} &= \emptyset \\
S_{6,0} &= \emptyset
\end{aligned}
$$



Iteration 2

Update using equation $S_1 = f_1(S_0)$:

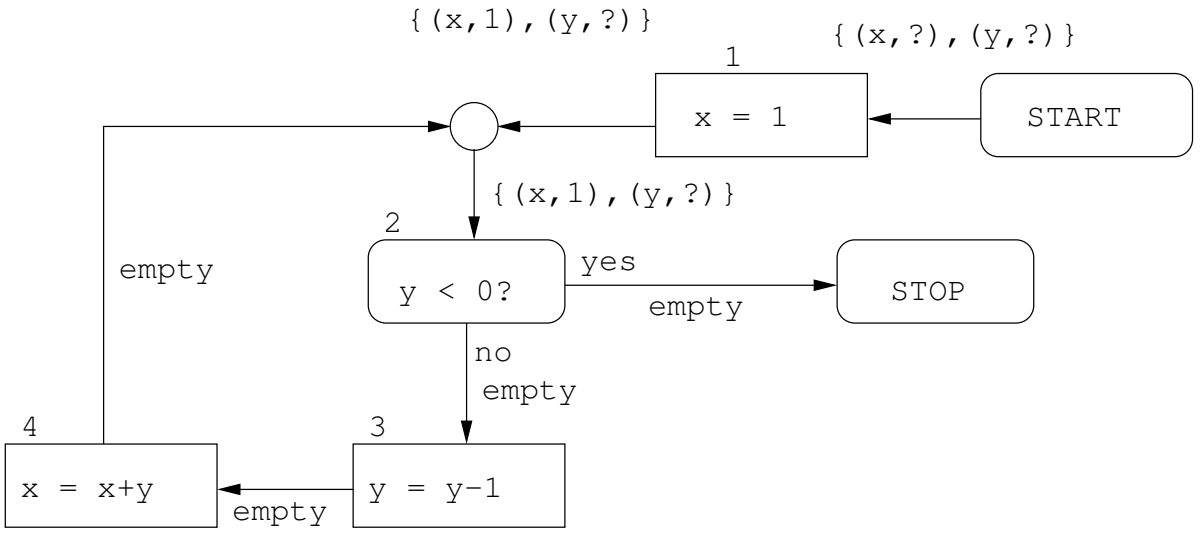
$$
\begin{array}{lcl}
S_{0,1} & = & \{(\mathtt{x}, ?), (\mathtt{y}, ?)\} \\
S_{1,1} & = & \{(\mathtt{x}, 1), (\mathtt{y}, ?)\} \\
S_{2,0} & = & \emptyset \\
S_{3,0} & = & \emptyset \\
S_{4,0} & = & \emptyset \\
S_{5,0} & = & \emptyset \\
S_{6,0} & = & \emptyset
\end{array}
$$



Iteration 3

Update using equation $S_2 = S_1 \cup S_6$:

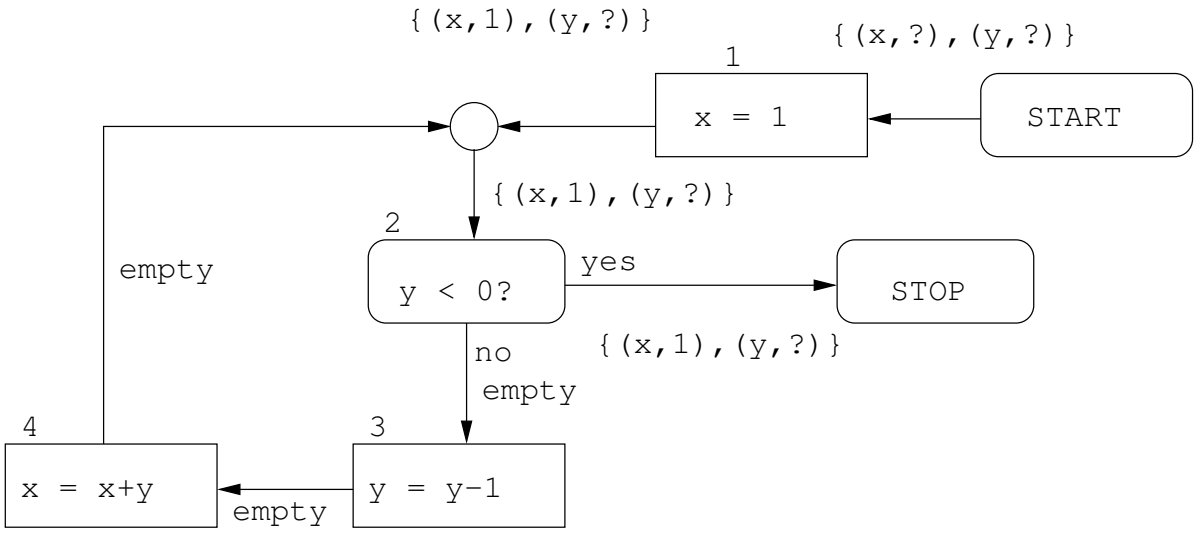
$$
\begin{array}{lcl}
S_{0,1} & = & \{(\mathtt{x}, ?), (\mathtt{y}, ?)\} \\
S_{1,1} & = & \{(\mathtt{x}, 1), (\mathtt{y}, ?)\} \\
S_{2,1} & = & \{(\mathtt{x}, 1), (\mathtt{y}, ?)\} \\
S_{3,0} & = & \emptyset \\
S_{4,0} & = & \emptyset \\
S_{5,0} & = & \emptyset \\
S_{6,0} & = & \emptyset
\end{array}
$$



Iteration 4

Update using equation $S_3 = S_2$:

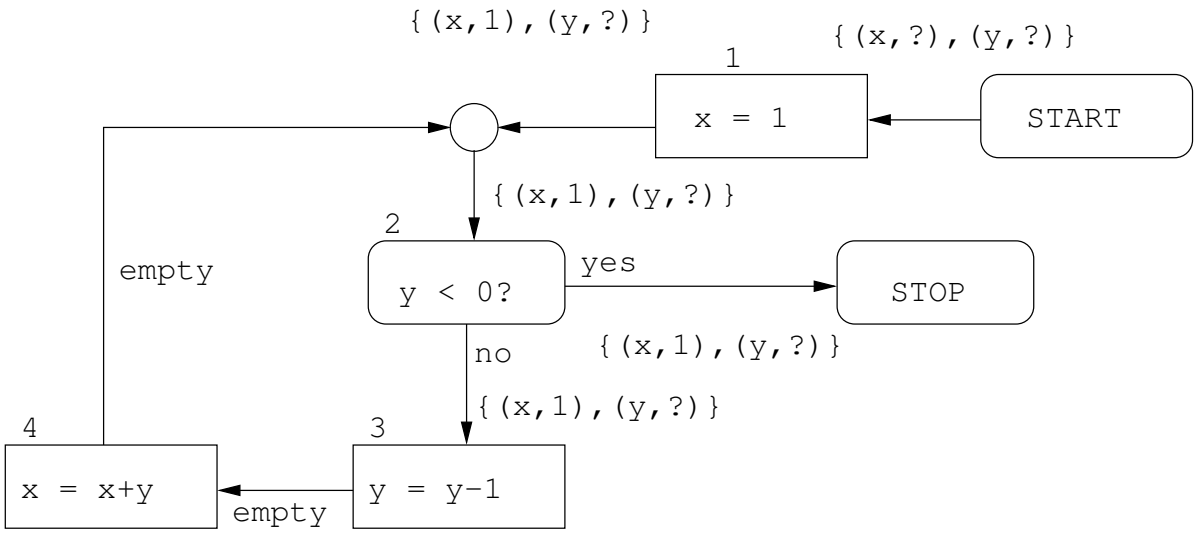
$$
\begin{array}{lcl}
S_{0,1} & = & \{(\mathtt{x}, ?), (\mathtt{y}, ?)\} \\
S_{1,1} & = & \{(\mathtt{x}, 1), (\mathtt{y}, ?)\} \\
S_{2,1} & = & \{(\mathtt{x}, 1), (\mathtt{y}, ?)\} \\
S_{3,1} & = & \{(\mathtt{x}, 1), (\mathtt{y}, ?)\} \\
S_{4,0} & = & \emptyset \\
S_{5,0} & = & \emptyset \\
S_{6,0} & = & \emptyset
\end{array}
$$



Iteration 5

Update using equation $S_4 = S_2$:

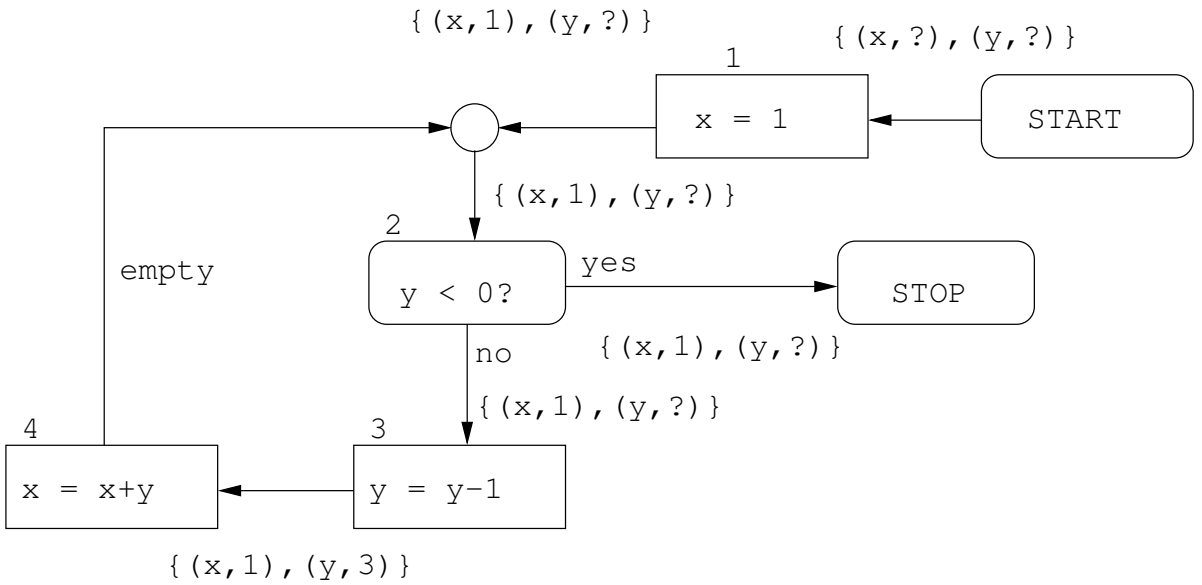
$$
\begin{array}{lcl}
S_{0,1} & = & \{(\mathtt{x},?),(\mathtt{y},?)\} \\
S_{1,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{2,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{3,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{4,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{5,0} & = & \emptyset \\
S_{6,0} & = & \emptyset
\end{array}
$$



Iteration 6

Update using equation $S_5 = f_3(S_4)$:

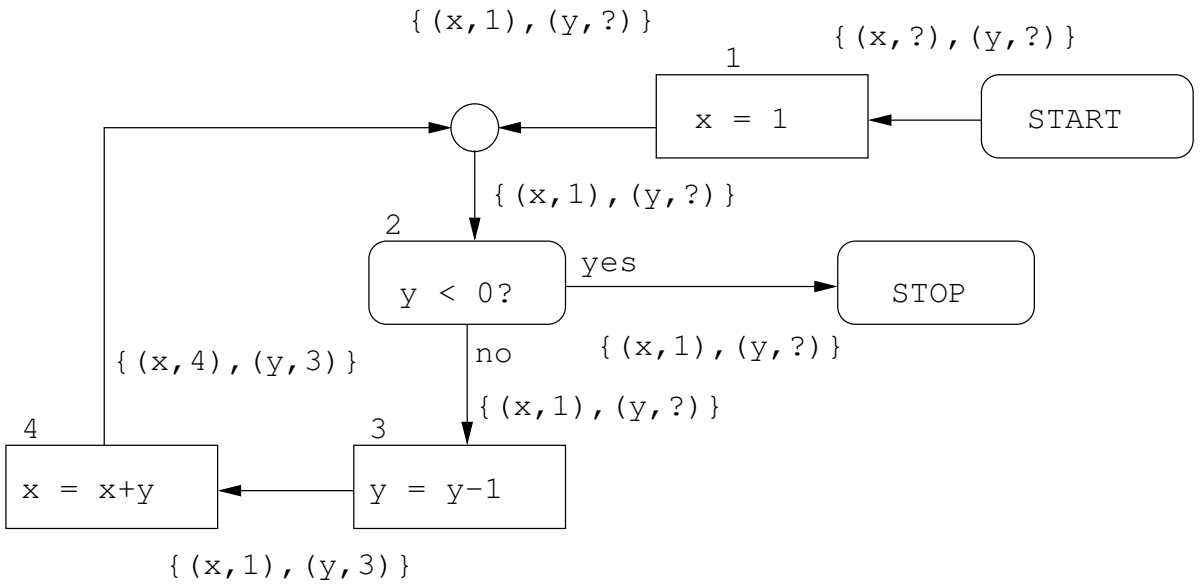
$$
\begin{array}{lcl}
S_{0,1} & = & \{(\mathtt{x},?),(\mathtt{y},?)\} \\
S_{1,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{2,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{3,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{4,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{5,1} & = & \{(\mathtt{x},1),(\mathtt{y},3)\} \\
S_{6,0} & = & \emptyset
\end{array}
$$



Iteration 7

Update using equation $S_6 = f_4(S_5)$:

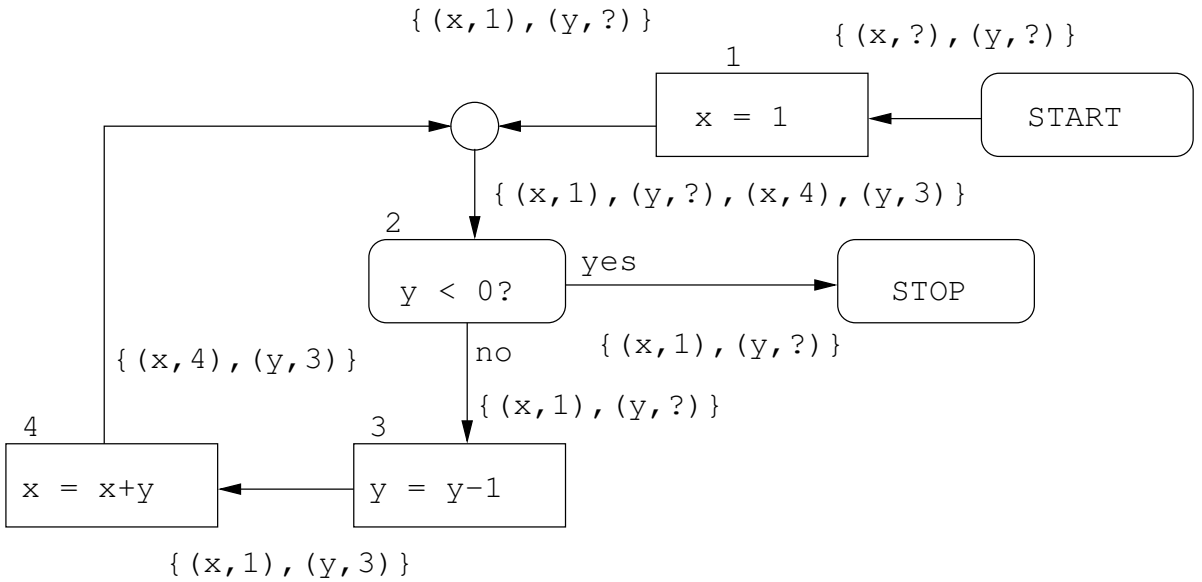
$$
\begin{array}{lcl}
S_{0,1} & = & \{(\mathtt{x},?),(\mathtt{y},?)\} \\
S_{1,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{2,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{3,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{4,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{5,1} & = & \{(\mathtt{x},1),(\mathtt{y},3)\} \\
S_{6,1} & = & \{(\mathtt{x},4),(\mathtt{y},3)\}
\end{array}
$$



Iteration 8

Update using equation $S_2 = S_1 \cup S_6$:

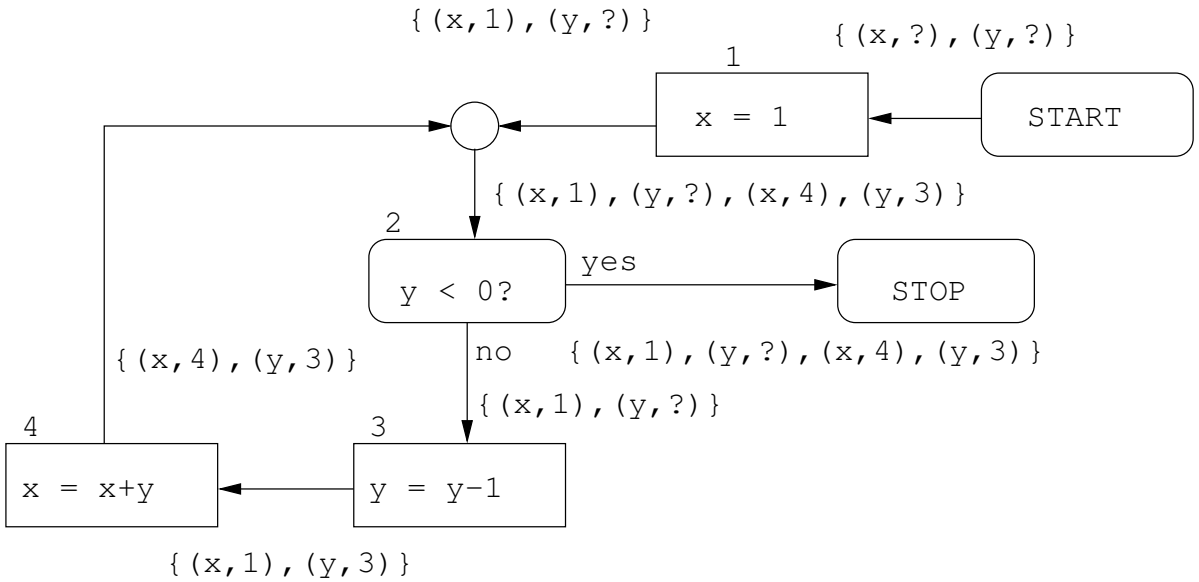
$$
\begin{array}{lcl}
S_{0,1} & = & \{(\mathtt{x},?),(\mathtt{y},?)\} \\
S_{1,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{2,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \ (\mathtt{x},4),(\mathtt{y},3)\} \\
S_{3,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{4,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{5,1} & = & \{(\mathtt{x},1),(\mathtt{y},3)\} \\
S_{6,1} & = & \{(\mathtt{x},4),(\mathtt{y},3)\}
\end{array}
$$



Iteration 9

Update using equation $S_3 = S_2$:

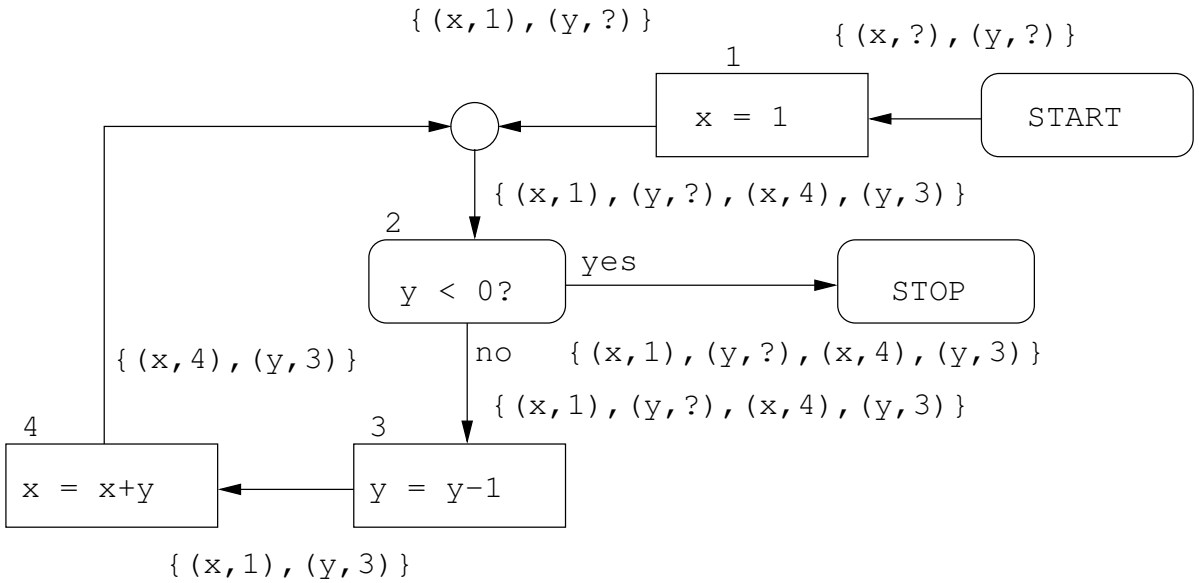
$$
\begin{array}{lcl}
S_{0,1} & = & \{(\mathtt{x},?),(\mathtt{y},?)\} \\
S_{1,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{2,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \quad (\mathtt{x},4),(\mathtt{y},3)\} \\
S_{3,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \quad (\mathtt{x},4),(\mathtt{y},3)\} \\
S_{4,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{5,1} & = & \{(\mathtt{x},1),(\mathtt{y},3)\} \\
S_{6,1} & = & \{(\mathtt{x},4),(\mathtt{y},3)\}
\end{array}
$$



Iteration 10

Update using equation $S_4 = S_2$:

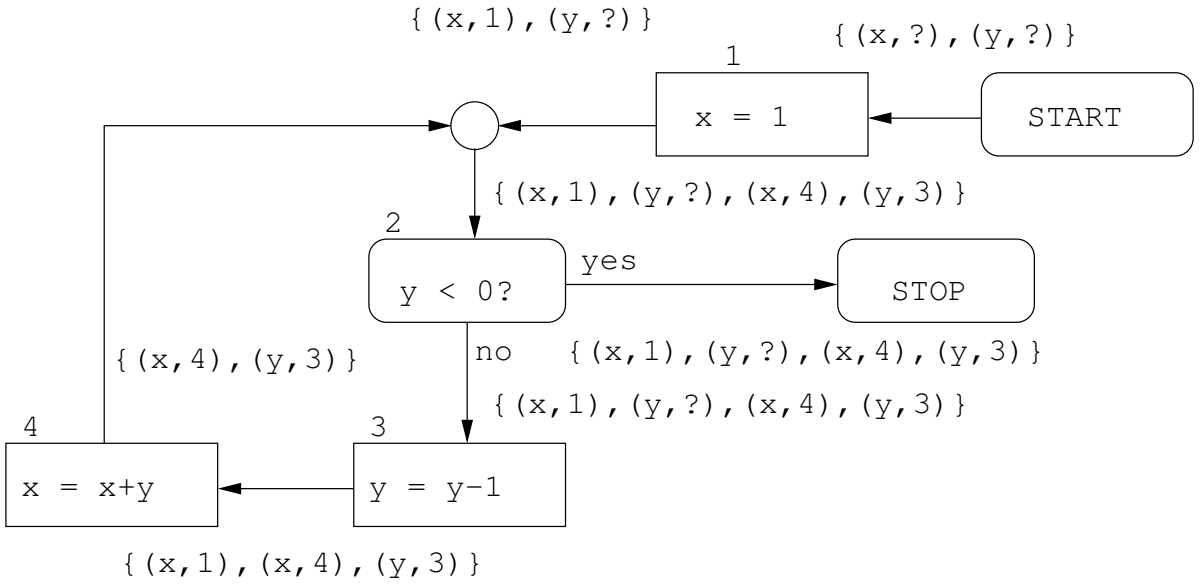
$$
\begin{array}{lcl}
S_{0,1} & = & \{(\mathtt{x},?),(\mathtt{y},?)\} \\
S_{1,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{2,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \ (\mathtt{x},4),(\mathtt{y},3)\} \\
S_{3,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \ (\mathtt{x},4),(\mathtt{y},3)\} \\
S_{4,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \ (\mathtt{x},4),(\mathtt{y},3)\} \\
S_{5,1} & = & \{(\mathtt{x},1),(\mathtt{y},3)\} \\
S_{6,1} & = & \{(\mathtt{x},4),(\mathtt{y},3)\}
\end{array}
$$



Iteration 11

Update using equation $S_5 = f_3(S_4)$:

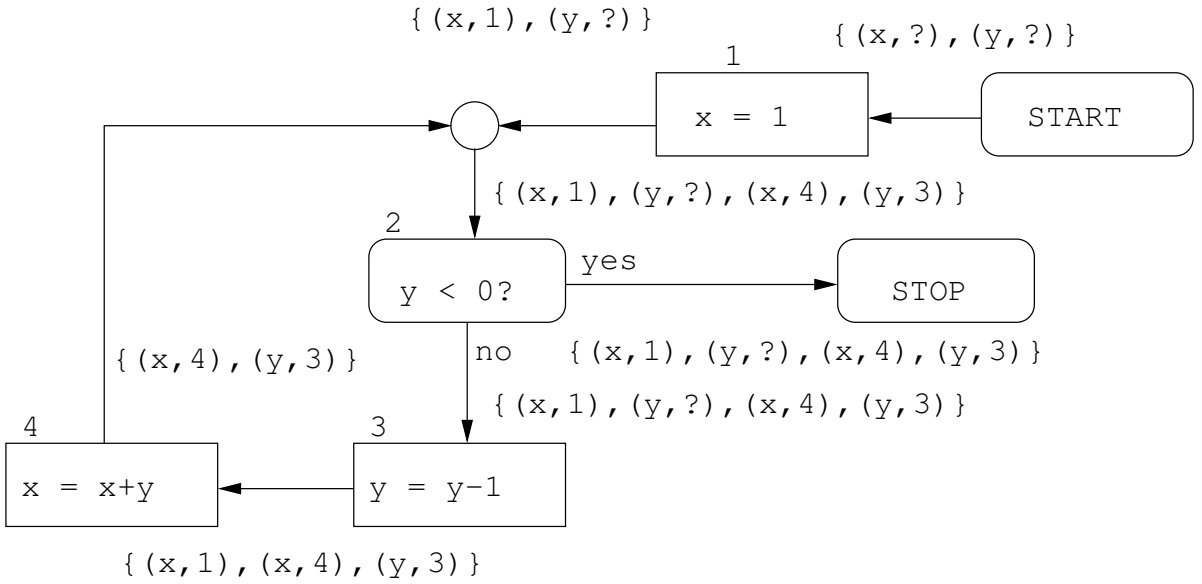
$$
\begin{array}{lcl}
S_{0,1} & = & \{(\mathtt{x},?),(\mathtt{y},?)\} \\
S_{1,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{2,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \ (\mathtt{x},4),(\mathtt{y},3)\} \\
S_{3,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \ ,(\mathtt{x},4),(\mathtt{y},3)\} \\
S_{4,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \ (\mathtt{x},4),(\mathtt{y},3)\} \\
S_{5,2} & = & \{(\mathtt{x},1),(\mathtt{x},4), \\
 & & \ (\mathtt{y},3)\} \\
S_{6,1} & = & \{(\mathtt{x},4),(\mathtt{y},3)\}
\end{array}
$$

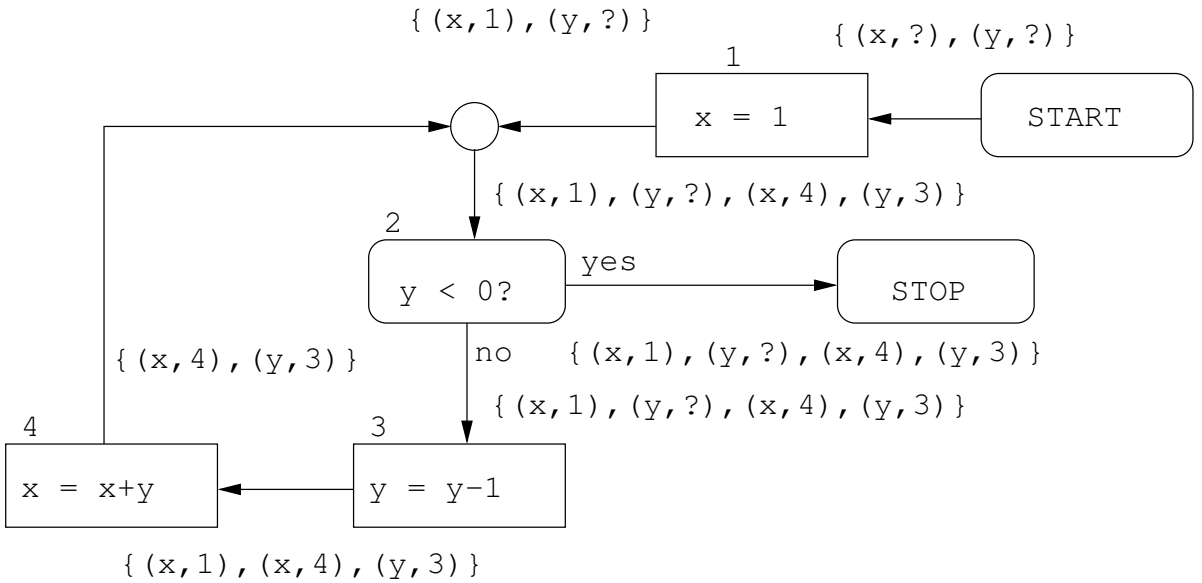


Iteration 12

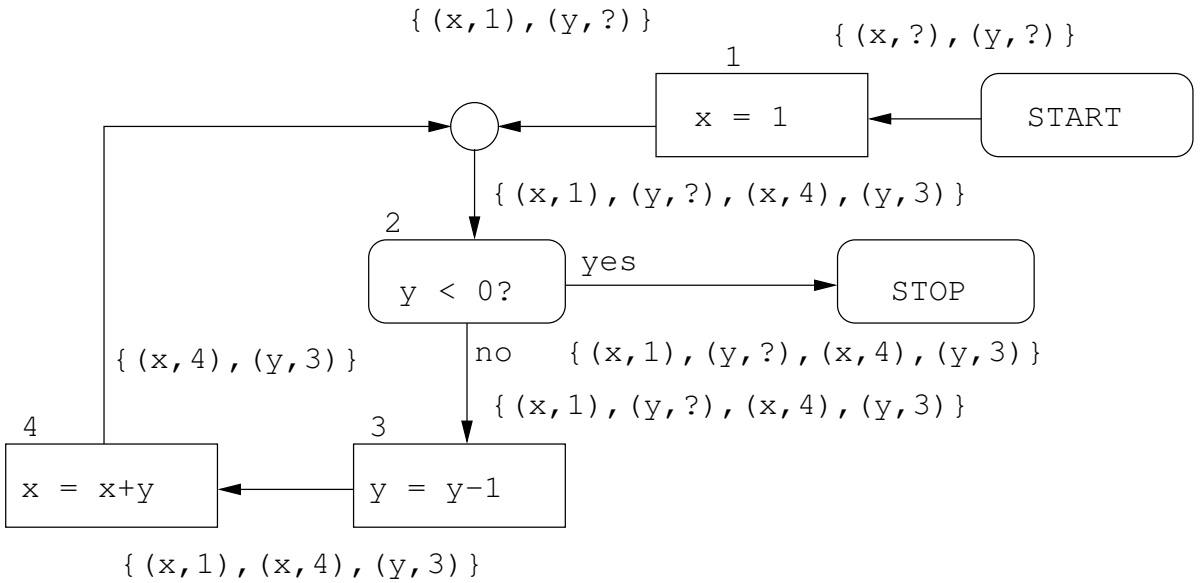
Update using equation $S_6 = f_4(S_5)$:

$$
\begin{array}{lcl}
S_{0,1} & = & \{(\mathtt{x},?),(\mathtt{y},?)\} \\
S_{1,1} & = & \{(\mathtt{x},1),(\mathtt{y},?)\} \\
S_{2,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \ (\mathtt{x},4),(\mathtt{y},3)\} \\
S_{3,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \ (\mathtt{x},4),(\mathtt{y},3)\} \\
S_{4,2} & = & \{(\mathtt{x},1),(\mathtt{y},?), \\
 & & \ (\mathtt{x},4),(\mathtt{y},3)\} \\
S_{5,2} & = & \{(\mathtt{x},1),(\mathtt{x},4), \\
 & & \ (\mathtt{y},3)\} \\
S_{6,2} & = & \{(\mathtt{x},4),(\mathtt{y},3)\}
\end{array}
$$





S_6 did not change. Thus, nothing more can change (no red sets left). We have reached a *fixed-point* – a solution to the system of equations! This is the result of the analysis



Statements 2 and 3 use $\mathtt{y}$. Presence of $(\mathtt{y},\mathtt{?})$ implies that y *possibly is not initialized there*, a potential bug

A Remark on the Underlying Math

None of the sets $S_0, \ldots, S_6$ ever shrunk during the iteration

This is due to a property of the transfer functions called *monotonicity*

This property guarantees that the fixed-point iteration terminates in a finite number of steps: since there is an upper limit how big each set can be, and since sets never shrink, sooner or later no set will change any more!

Starting with empty sets gives the *least possible* (most precise) solution

Underlying mathematical theory of *complete lattices*. *Tarski's Fixed-Point Theorem* guarantees that least solutions always exist and can be found in this way

Wrapping Up Lectures 1 – 4

Semantics-based static program analysis works by:

1. setting up equations according to program structure
2. solving the equations by fixed-point iteration

The solution tells something about the properties of the program

Many different properties can be derived, by setting up different kinds of equations

The analyses make *safe overapproximations* – reports on absence of errors can be trusted, but there may be false positives

Automated tools exist, and are increasingly being used