Lecture 23: Code Optimization

[Adapted from notes by R. Bodik and G. Necula]

Public Service Announcement. "Hackers@Berkeley is hosting BEARHACK this Saturday at 11AM in the Wozniak Lounge. It's a 24 hour hackathon - there'll be tons of good food (Cheeseboard, sushi & boba!), activities (including massages!), and prizes (including an Occulus Rift).

Introduction to Code Optimization

Code optimization is the usual term, but is grossly misnamed, since code produced by "optimizers" is not optimal in any reasonable sense. *Program improvement* would be more appropriate.

Topics:

- Basic blocks
- Control-flow graphs (CFGs)
- Algebraic simplification
- Constant folding
- Static single-assignment form (SSA)
- Common-subexpression elimination (CSE)
- Copy propagation
- Dead-code elimination
- Peephole optimizations

Basic Blocks

- A *basic block* is a maximal sequence of instructions with:
  - no labels (except at the first instruction), and
  - no jumps (except in the last instruction)
- Idea:
  - Cannot jump into a basic block, except at the beginning.
  - Cannot jump within a basic block, except at end.
  - Therefore, each instruction in a basic block is executed after all the preceding instructions have been executed

Basic-Block Example

- Consider the basic block

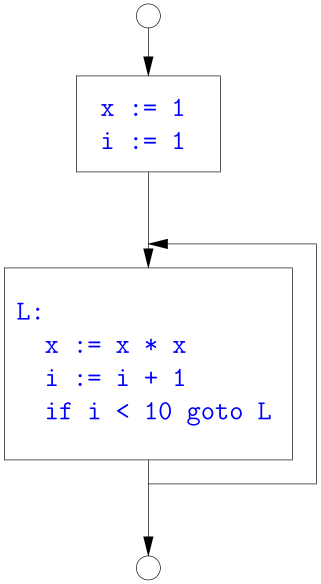
```
1. L1:
2. t := 2 * x
3. w := t + x
4. if w > 0 goto L2
```

- No way for (3) to be executed without (2) having been executed right before
- We can change (3) to `w := 3 * x`
- Can we eliminate (2) as well?

Control-Flow Graphs (CFGs)

- A control-flow graph is a directed graph with basic blocks as nodes
- There is an edge from block A to block B if the execution can flow from the last instruction in A to the first instruction in B:
  - The last instruction in A can be a jump to the label of B.
  - Or execution can fall through from the end of block A to the beginning of block B.

Control-Flow Graphs: Example



- The body of a method (or procedure) can be represented as a CFG
- There is one initial node
- All "return" nodes are terminal

Optimization Overview

- Optimization seeks to improve a program's utilization of some resource:
  - Execution time (most often)
  - Code size
  - Network messages sent
  - Battery power used, etc.
- Optimization should not depart from the programming language's semantics
- So if the semantics of a particular program is deterministic, optimization must not change the answer.
- On the other hand, some program behavior is undefined (e.g., what happens when an unchecked rule in C is violated), and in those cases, optimization may cause differences in results.

A Classification of Optimizations

- For languages like C and Java there are three granularities of optimizations
  1. *Local optimizations:* Apply to a basic block in isolation.
  2. *Global optimizations:* Apply to a control-flow graph (single function body) in isolation.
  3. *Inter-procedural optimizations:* Apply across function boundaries.
- Most compilers do (1), many do (2) and very few do (3)
- Problem is expense: (2) and (3) typically require superlinear time. Can usually handle that when limited to a single function, but gets problematic for larger program.
- In practice, generally *don't* implement fanciest known optimizations: some are hard to implement (esp., hard to get right), some require a lot of compilation time.
- The goal: maximum improvement with minimum cost.

Local Optimizations: Algebraic Simplification

- Some statements can be deleted

```
x := x + 0
x := x * 1
```

- Some statements can be simplified or converted to use faster operations:

Original	Simplified
`x := x * 0`	`x := 0`
`y := y ** 2`	`y := y * y`
`x := x * 8`	`x := x << 3`
`x := x * 15`	`t := x << 4; x := t - x`

(on some machines << is faster than *; but not on all!)

Local Optimization: Constant Folding

- Operations on constants can be computed at compile time.
- Example: `x := 2 + 2` becomes x := 4.
- Example: if 2 < 0 jump L becomes a no-op.
- When might constant folding be dangerous?

Global Optimization: Unreachable code elimination

- Basic blocks that are not reachable from the entry point of the CFG may be eliminated.
- Why would such basic blocks occur?
- Removing unreachable code makes the program smaller (sometimes also faster, due to instruction-cache effects, but this is probably not a terribly large effect.)

Single Assignment Form

- Some optimizations are simplified if each assignment is to a temporary that has not appeared already in the basic block.
- Intermediate code can be rewritten to be in *(static) single assignment (SSA) form:*

```
x := a + y          x := a + y
a := x              a1 := x
x := a * x          x1 := a1 * x
b := x + a          b := x1 + a1
```

where `x1` and `a1` are fresh temporaries.

Common SubExpression (CSE) Elimination in Basic Blocks

- A *common subexpression* is an expression that appears multiple times on a right-hand side in contexts where the operands have the same values in each case (so that the expression will yield the same value).
- Assume that the basic block on the left is in single assignment form.

```
x := y + z              x := y + z
            ...
            ...
w := y + z              w := x
```

- That is, if two assignments have the same right-hand side, we can replace the second instance of that right-hand side with the variable that was assigned the first instance.
- How did we use the assumption of single assignment here?

Copy Propagation

- If `w := x` appears in a block, can replace all subsequent uses of `w` with uses of `x`.
- Example:

```
b:=z+y        b:=z+y
a := b        a := b
x:=2*a        x:=2*b
```

- This does not make the program smaller or faster but might enable other optimizations. For example, if `a` is not used after this statement, we need not assign to it.
- Or consider:

```
b:=13         b:=13
x:=2*b        x:=2*13
```

  which immediately enables constant folding.
- Again, the optimization, as described, won't work unless the block is in single assignment form.

Another Example of Copy Propagation and Constant Folding

```
a := 5         a := 5         a := 5          a := 5           a := 5
x := 2 * a     x := 2 * 5     x := 10         x := 10          x := 10
y := x + 6     y := x + 6     y := 10 + 6     y := 16          y := 16
t := x * y     t := x * y     t := 10 * y     t := 10 * 16     t := 160
```


Dead Code Elimination

- If that statement `w := rhs` appears in a basic block and `w` does not appear anywhere else in the program, we say that the statement is *dead* and can be eliminated; it does not contribute to the program's result.
- Example: (`a` is not used anywhere else)

```
b := z + y       b := z + y     b := z + y
a := b           a := b
x := 2 * a       x := 2 * b     x := 2 * b
```

- How have I used SSA here?

Applying Local Optimizations

- As the examples show, each local optimization does very little by itself.
- Typically, optimizations interact: performing one optimization enables others.
- So typical optimizing compilers repeatedly perform optimizations until no improvement is possible, or it is no longer cost effective.

An Example: Initial Code

```
a := x ** 2
b := 3
c := x
d := c * c
e := b * 2
f := a + d
g := e * f
```


An Example II: Algebraic simplification

```
a := x * x
b := 3
c := x
d := c * c
e := b + b
f := a + d
g := e * f
```


An Example: Copy propagation

```
a := x * x
b := 3
c := x
d := x * x
e := 3 + 3
f := a + d
g := e * f
```

An Example: Constant folding

```
a := x * x
b := 3
c := x
d := x * x
e := 6
f := a + d
g := e * f
```


An Example: Common Subexpression Elimination

```
a := x * x
b := 3
c := x
d := a
e := 6
f := a + d
g := e * f
```


An Example: Copy propagation

```
a := x * x
b := 3
c := x
d := a
e := 6
f := a + a
g := 6 * f
```


An Example: Dead code elimination

```
a := x * x




f := a + a
g := 6 * f
```

This is the final form.

Peephole Optimizations on Assembly Code

- The optimizations presented before work on intermediate code.
- *Peephole optimization* is a technique for improving assembly code directly
  - The "*peephole*" is a short subsequence of (usually contiguous) instructions, either continguous, or linked together by the fact that they operate on certain registers that no intervening instructions modify.
  - The optimizer replaces the sequence with another equivalent, but (one hopes) better one.
  - Write peephole optimizations as replacement rules

    i1; ...; in ⇒ j1; ...; jm

    possibly plus additional constraints. The j's are the improved version of the i's.

Peephole optimization examples:

- We'll use the notation '@A' for pattern variables.
- Example:

  movl %@a %@b; L: movl %@b %@a ⇒ movl %@a %@b

  assuming `L` is not the target of a jump.
- Example:

  addl $@k1, %@a; movl @k2(%@a), %@b
  ⇒ movl @k1+@k2(%@a), %@b

  assuming `%@a` is "dead".
- Example (PDP11):

  mov #@I, @I(@ra) ⇒ mov (r7), @I(@ra)

  This is a real hack: we reuse the value `I` as both the immediate value and the offset from `ra`. On the PDP11, the program counter is r7.
- As for local optimizations, peephole optimizations need to be applied repeatedly to get maximum effect.

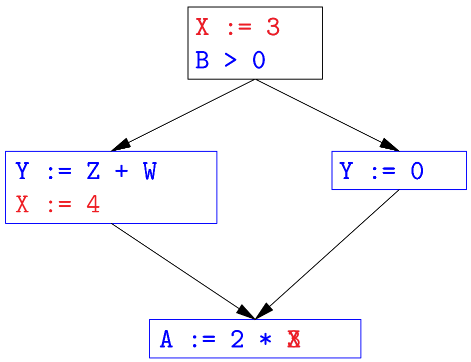
Problems:

- Serious problem: what to do with pointers? Problem is *aliasing:* two names for the same variable:
  - As a result, `*t` may change even if local variable `t` does not and we never assign to `*t`.
  - Affects language design: rules about overlapping parameters in Fortran, and the **restrict** keyword in C.
  - Arrays are a special case (address calculation): is `A[i]` the same as `A[j]`? Sometimes the compiler can tell, depending on what it knows about `i` and `j`.
- What about globals variables and calls?
  - Calls are not exactly jumps, because they (almost) always return.
  - Can modify global variables used by caller

Global Optimization

- *Global optimization* refers to program optimizations that encompass multiple basic blocks in a function.
- (I have used the term *galactic optimization* to refer to going beyond function boundaries, but it hasn't caught on; we call it just *interprocedural optimization.*)
- Since we can't use the usual assumptions about basic blocks, global optimization requires *global flow analysis* to see where values can come from and get used.
- The overall question is: When can local optimizations (from the last lecture) be applied across multiple basic blocks?

A Simple Example: Copy Propagation



- Without other assignments to X, it is valid to treat the red parts as if they were in the same basic block.
- But as soon as *one* other block on the path to the bottom block assigns to X, we can no longer do so.
- It is correct to apply copy propagation to a variable x from an assignment statement `A: x := ...` to a given use of x in statement B only if the last assignment to x in every path from to B is A.

Issues

- This correctness condition is not trivial to check
- "All paths" includes paths around loops and through branches of conditionals
- Checking the condition requires global analysis: an analysis of the entire control-flow graph for one method body.
- This is typical for optimizations that depend on some property P at a particular point in program execution.
- Indeed, property P is typically undecidable, so program optimization is all about making *conservative* (but not cowardly) approximations of P.

Undecidability of Program Properties

- Rice's "theorem:" Most interesting dynamic properties of a program are undecidable. E.g.,
  - Does the program halt on all (some) inputs? (Halting Problem)
  - Is the result of a function F always positive? (Consider

    ```
    def F(x):
        H(x)
        return 1
    ```

    Result is positive iff H halts.)
- Syntactic properties are typically decidable (e.g., "How many occurrences of x are there?").
- Theorem does not apply in absence of loops

Conservative Program Analyses

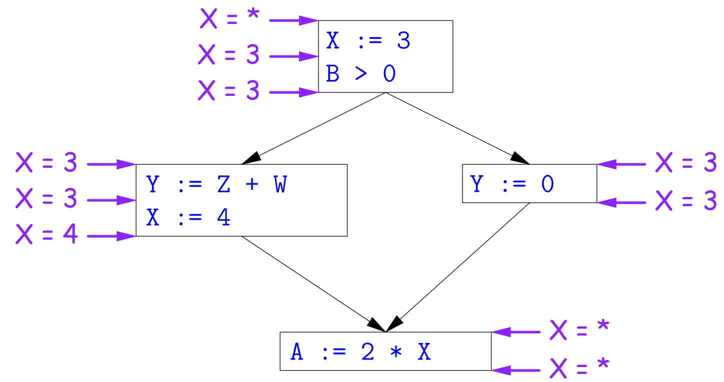
- If a certain optimization requires P to be true, then
  - If we know that P is definitely true, we can apply the optimization
  - If we don't know whether P is true, we simply don't do the optimization. Since optimizations are not supposed to change the meaning of a program, this is safe.
- In other words, in analyzing a program for properties like P, it is *always correct* (albeit non-optimal) to say "don't know."
- The trick is to say it as seldom as possible.
- *Global dataflow analysis* is a standard technique for solving problems with these characteristics.

Example: Global Constant Propagation

- *Global constant propagation* is just the restriction of copy propagation to constants.
- In this example, we'll consider doing it for a single variable (`X`).
- At every program point (i.e., before or after any instruction), we associate one of the following values with `X`

Value	Interpretation
#	(aka *bottom*) No value has reached here (yet)
c	(For c a constant) `X` definitely has the value c.
$*$	(aka *top*) Don't know what, if any, constant value `X` has.

Example of Result of Constant Propagation



Using Analysis Results

- Given global constant information, it is easy to perform the optimization:
  - If the point immediately before a statement using `x` tells us that `x = c`, then replace `x` with `c`.
  - Otherwise, leave it alone (the conservative option).
- But how do we compute these properties `x = ...`?

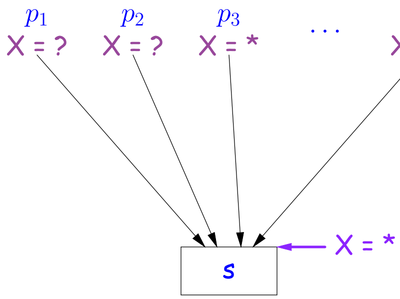
Transfer Functions

- **Basic Idea:** Express the analysis of a complicated program as a combination of simple rules relating the change in information between adjacent statements
- That is, we *"push"* or *transfer* information from one statement to the next.
- For each statement `s`, we end up with information about the value of `x` immediately before and after `s`:

  Cin(X,s) = value of `x` before `s`
  Cout(X,s) = value of `x` after `s`

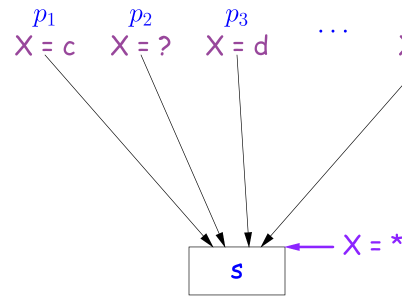
- Here, the "values of `x`" we use come from an *abstract domain*, containing the values we care about—#, $*$, k—values computed *statically* by our analysis.
- For the constant propagation problem, we'll compute Cout from Cin, and we'll get Cin from the Couts of predecessor statements, Cout(X, p_1),...,Cout(X,p_n).

Constant Propagation: Rule 1



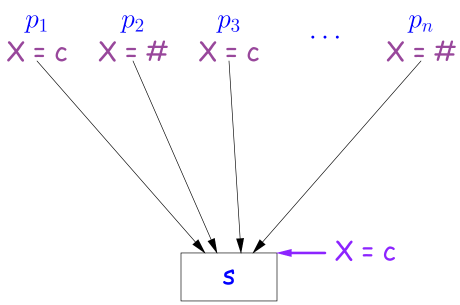
If Cout(X, p_i) = $*$ for some i, then Cin(X, s) = $*$

Constant Propagation: Rule 2



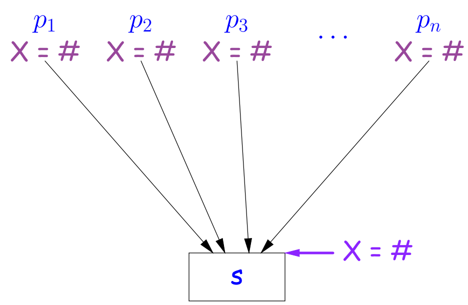
If Cout(X, p_i) = c and Cout(X, p_j) = d with constants $c \neq d$,
then Cin(X, s) = $*$

Constant Propagation: Rule 3



If Cout(X, p_i) = c for some i and
Cout(X, p_j) = c or Cout(X, p_j) = # for all j,
then Cin(X, s) = c

Constant Propagation: Rule 4

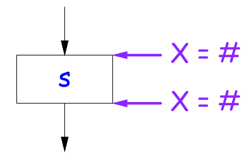


If Cout(X, p_j) = # for all j, then Cin(X, s) = #

Constant Propagation: Computing Cout

- Rules 1–4 relate the *out* of one statement to the *in* of the successor statements, thus propagating information *forward* across CFG edges.
- Now we need *local* rules relating the *in* and *out* of a single statement to propagate information across statements.

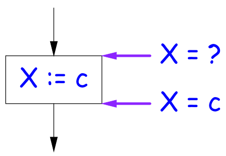
Constant Propagation: Rule 5



Cout(X, s) = # if Cin(X, s) = #

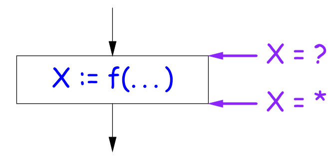
The value '#' means "so far, no value of X gets here, because the we don't (yet) know that this statement ever gets executed."

Constant Propagation: Rule 6



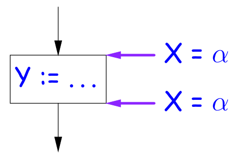
Cout(X, X := c) = c if c is a constant and ? is not #.

Constant Propagation: Rule 7



Cout(X, X := f(...)) = ∗ for any function call, if ? is not #.

Constant Propagation: Rule 8

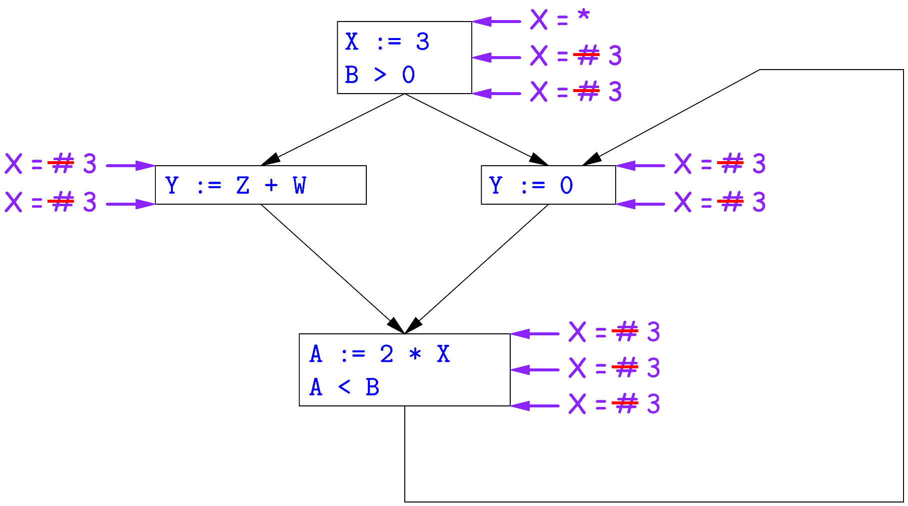


Cout(X, Y := . . .) = Cin(X, Y := . . .) if X and Y are different variables.

Propagation Algorithm

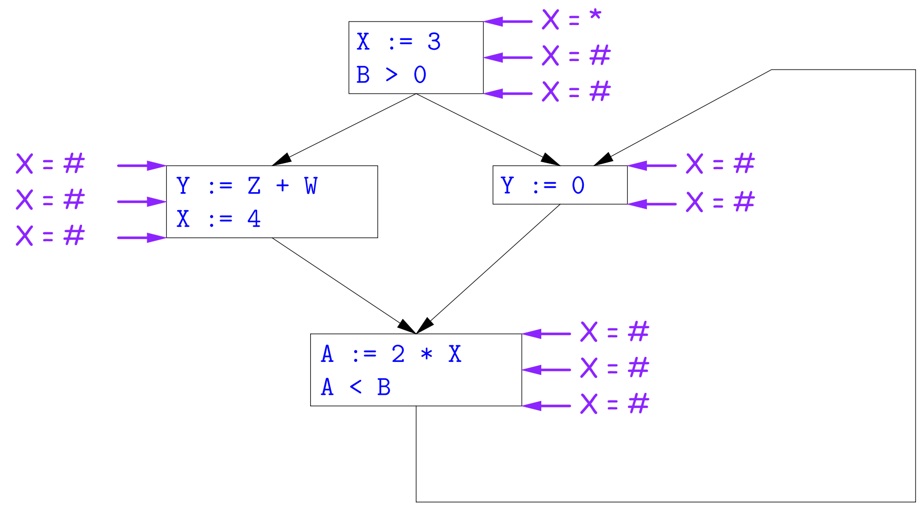
- To use these rules, we employ a standard technique: *iteration to a fixed point*:
- Mark all points in the program with *current approximations* of the variable(s) of interest (X in our examples).
- Set the initial approximations to X = * for the program entry point and X = # everywhere else.
- Repeatedly apply rules 1–8 every place they are applicable until nothing changes—until the program is at a *fixed point* with respect to all the transfer rules.
- We can be clever about this, keeping a list of all nodes any of whose predecessors' *Cout* values have changed since the last rule application.

An Example of the Algorithm

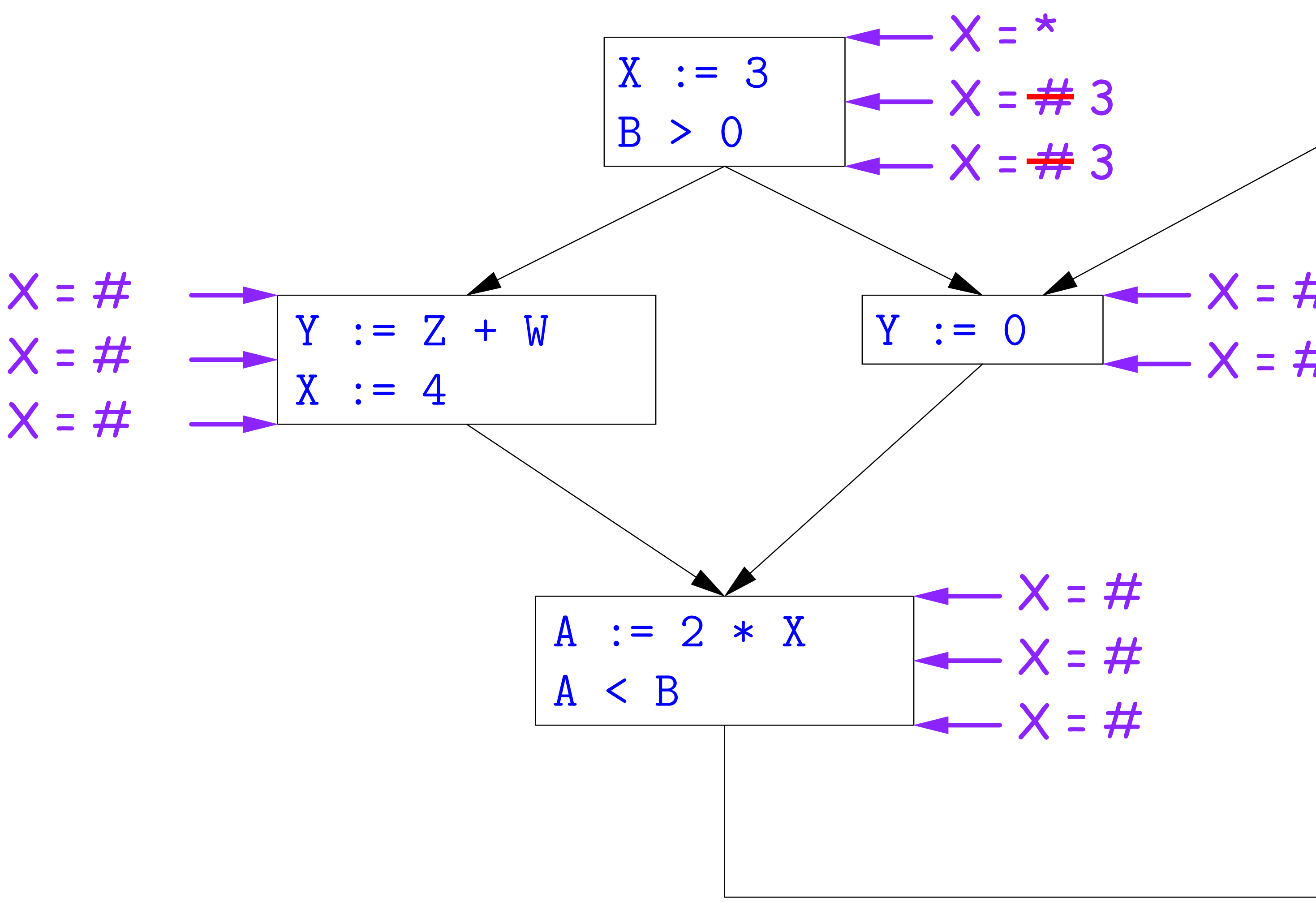


So we *can* replace X with 3 in the bottom block.

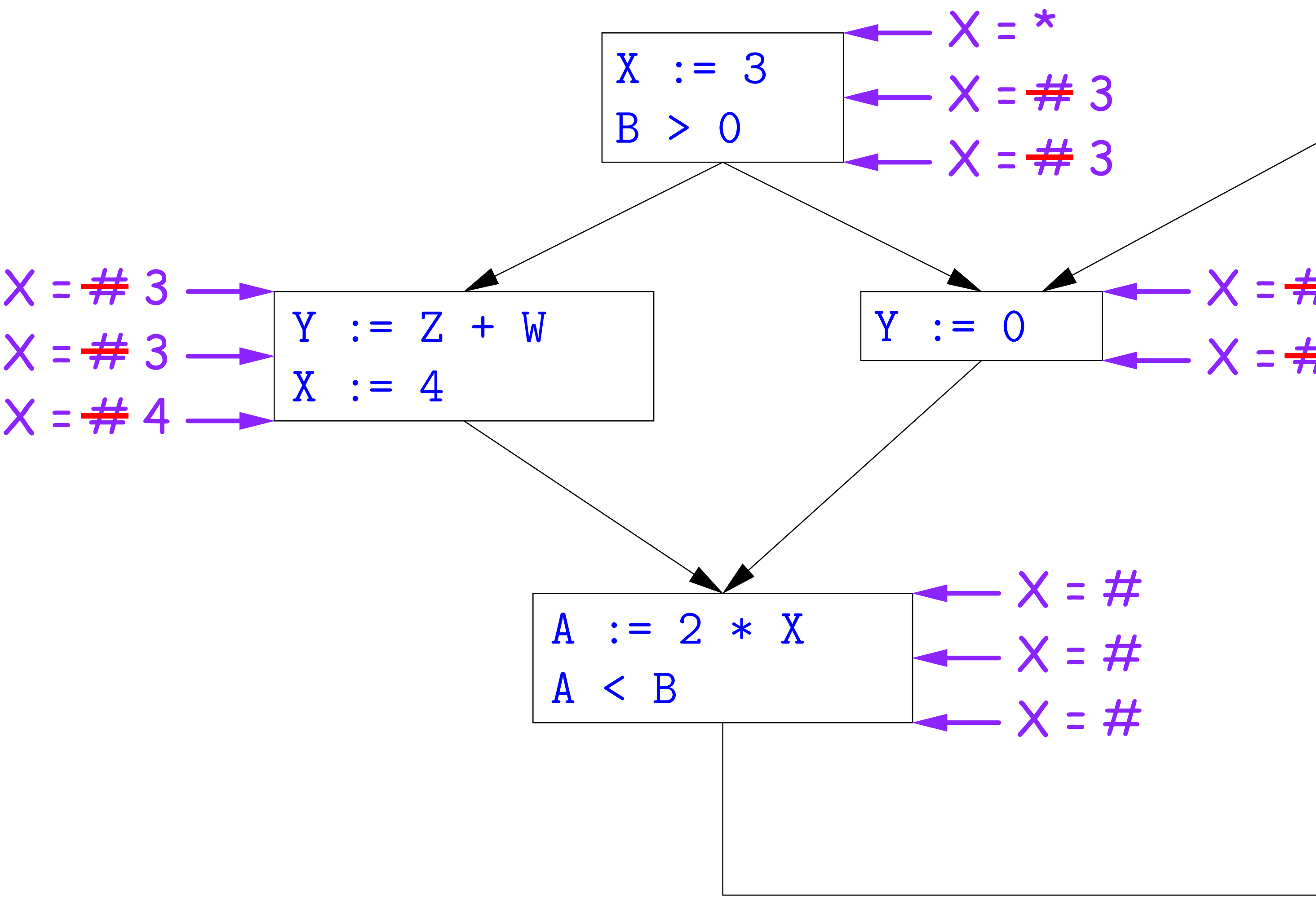
Another Example of the Propagation Algorithm



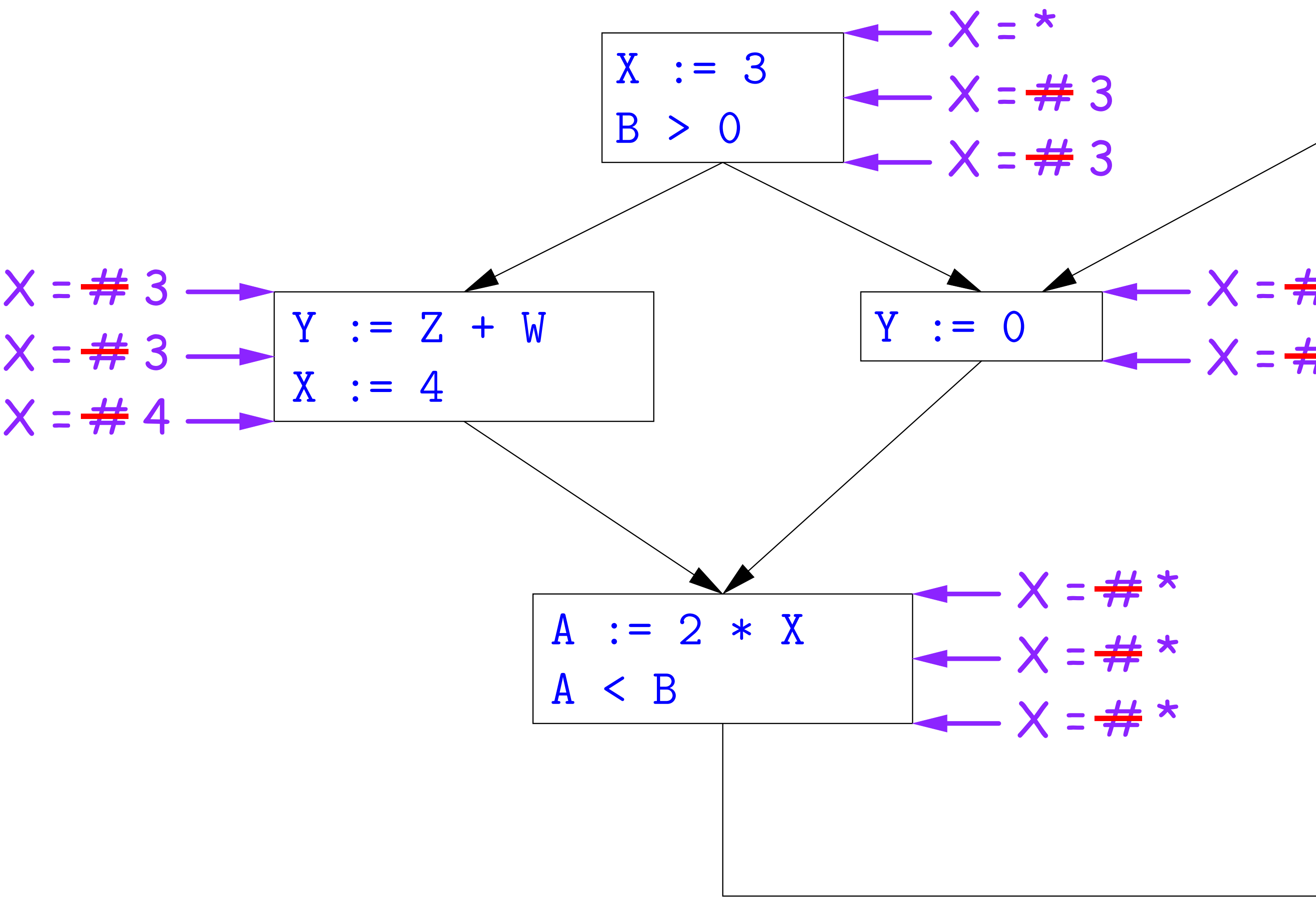
Another Example of the Propagation Algorithm



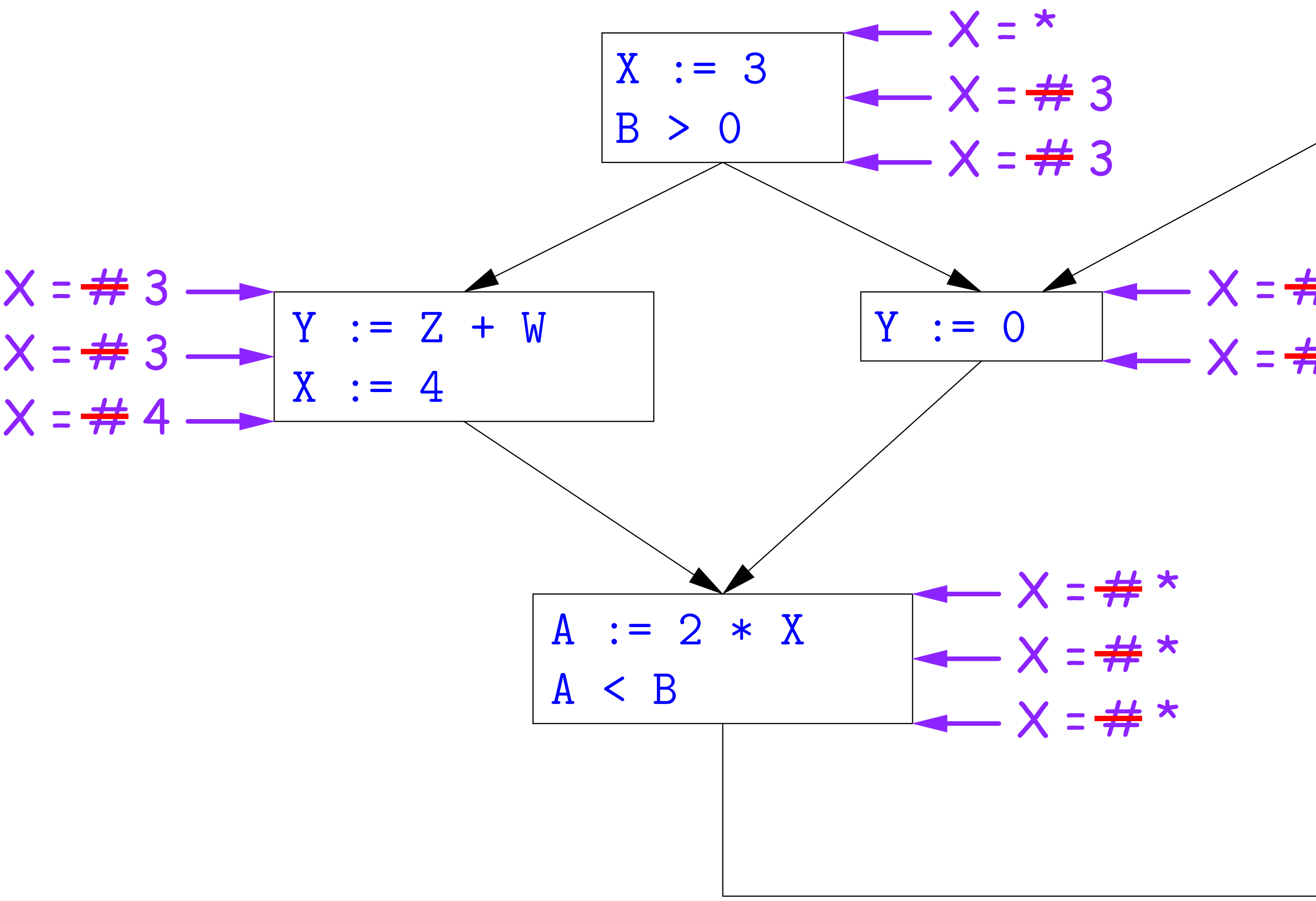
Another Example of the Propagation Algorithm



Another Example of the Propagation Algorith

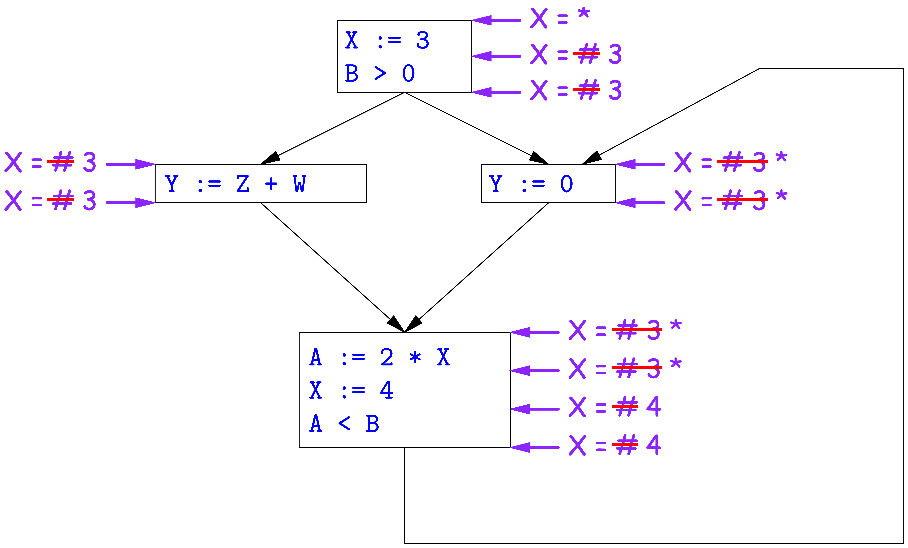


Another Example of the Propagation Algorith



Here, we *cannot* replace X in two of the basic blocks.

A Third Example



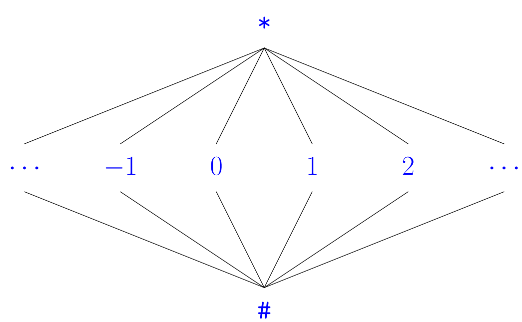
Likewise, we *cannot* replace X.

Comments

- The examples used a depth-first approach to considering possible places to apply the rules, starting from the entry point.
- In fact, the order in which one looks at statements is irrelevant. We could have changed the Cout values after the assignments to X first, for example.
- The # value is necessary to avoid deciding on a final value too soon. In effect, it allows us to tentatively propogate constant values through before finding out what happens in paths we haven't looked at yet.

Ordering the Abstract Domain

- We can simplify the presentation of the analysis by ordering the values $\# < c < *$.
- Or pictorially, with lower meaning less than,



- ...a mathematical structure known as a *lattice.*
- With this, our rule for computing Cin is simply a *least upper bound:*

  Cin(x, s) = lub { Cout(x, p) such that p is a predecessor of s }.

Termination

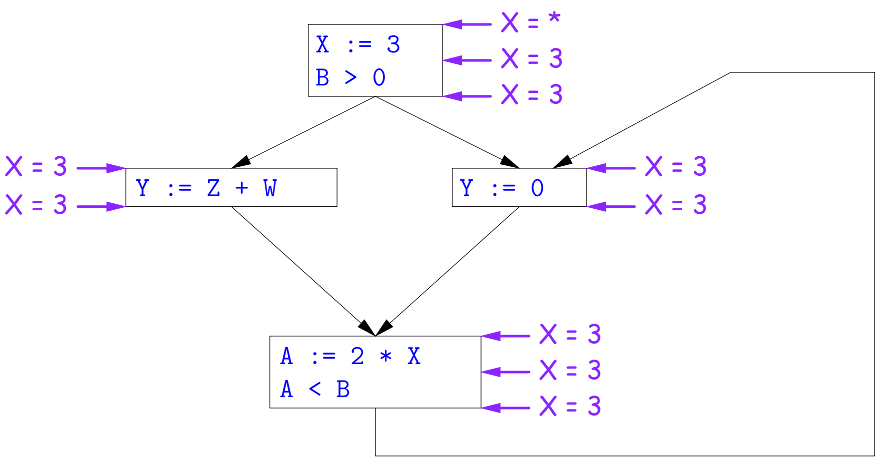
- Simply saying "repeat until nothing changes" doesn't guarantee that eventually nothing changes.
- But the use of lub explains why the algorithm terminates:
  - Values start as # and only increase
  - By the structure of the lattice, therefore, each value can only change twice.
- Thus the algorithm is linear in program size. The number of steps

  = $2\times$ Number of Cin and Cout values computed

  = $4\times$ Number of program statements.

Liveness Analysis

Once constants have been globally propagated, we would like to eliminate dead code



After constant propagation, `X := 3` is dead code (assuming this is the entire CFG)

Terminology: Live and Dead

- In the program

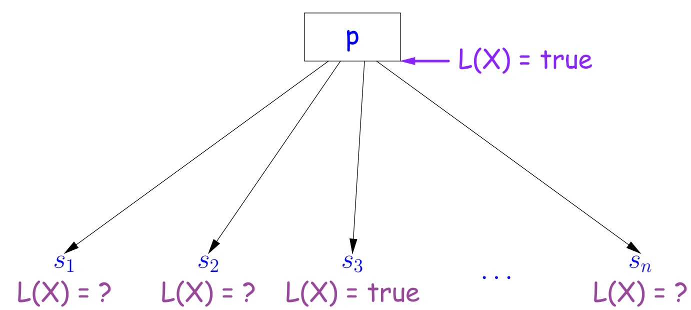
  `X := 3;  /*(1)*/  X = 4;  /*(2)*/  Y := X  /*(3)*/`

- the variable `X` is *dead* (never used) at point (1), *live* at point (2), and may or may not be live at point (3), depending on the rest of the program.
- More generally, a variable `x` is live at statement `s` if
  - There exists a statement `s'` that uses `x`;
  - There is a path from `s` to `s'`; and
  - That path has no intervening assignment to `x`
- A statement `x := ...` is dead code (and may be deleted) if `x` is dead after the assignment.

Computing Liveness

- We can express liveness as a function of information transferred between adjacent statements, just as in copy propagation
- Liveness is simpler than constant propagation, since it is a boolean property (true or false).
- That is, the lattice has two values, with `false<true`.
- It also differs in that liveness depends on what comes *after* a statement, not before—we propagate information *backwards* through the flow graph, from Lout (liveness information at the end of a statement) to Lin.

Liveness Rule 1

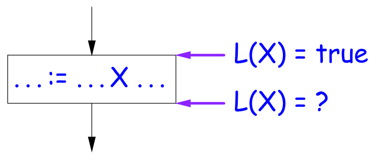


- So

  Lout(x, p) = lub { Lin(x, s) such that s is a predecessor of p }.

- Here, least upper bound (lub) is the same as "or".

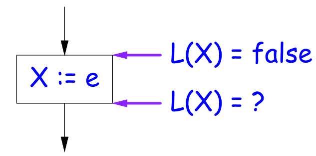
Liveness Rule 2



Lout(X, s) = true if s uses the previous value of X.

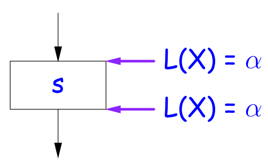
- The same rule applies to any other statement that uses the value of X, such as tests (e.g., X < 0).

Liveness Rule 3



Lout(X, X := e) = false if e does not use the previous value of X.

Liveness Rule 4

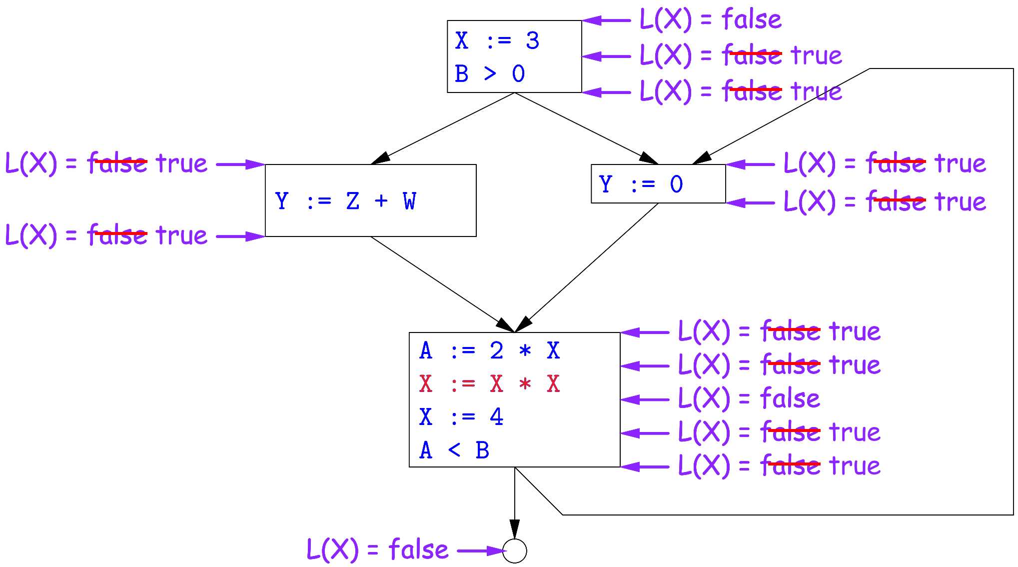


Lout(X, s) = Lin(X, s) if s does not mention X.

Propagation Algorithm for Liveness

- Initially, let all Lin and Lout values be false.
- Set Lout value at the program exit to true iff x is going to be used elsewhere (e.g., if it is global and we are analyzing only one procedure).
- As before, repeatedly pick s where one of 1–4 does not hold and update using the appropriate rule, until there are no more violations.
- When we're done, we can eliminate assignments to X if X is dead at the point after the assignment.

Example of Liveness Computation



Termination

- As before, a value can only change a bounded number of times: the bound being 1 in this case.
- Termination is guaranteed
- Once the analysis is computed, it is simple to eliminate dead code, but having done so, we must recompute the liveness information.

SSA and Global Analysis

- For local optimizations, the single static assignment (SSA) form was useful.
- But applying it to a full CFG is requires a trick.
- E.g., how do we avoid two assignments to the temporary holding x after this conditional?

```
if a > b:
    x = a
else:
    x = b
# where is x at this point?
```

- Answer: a small kludge known as ϕ "functions"
- Turn the previous example into this:

```
if a > b:
    x1 = a
else:
    x2 = b
x3 = φ(x1, x2)
```


ϕ Functions

- An artificial device to allow SSA notation in CFGs.
- In a basic block, each variable is associated with one definition,
- ϕ functions in effect associate each variable with a set of possible definitions.
- In general, one tries to introduce them in strategic places so as to minimize the total number of ϕs.
- Although this device increases number of assignments in IL, register allocation can remove many by assigning related IL registers to the same real register.
- Their use enables us to extend such optimizations as CSE elimination in basic blocks to *Global CSE Elimination*.
- With SSA form, easy to tell (conservatively) if two IL assignments compute the same value: just see if they have the same right-hand side. The same variables indicate the same values.

Summary

- We've seen two kinds of analysis:
  - Constant propagation is a *forward analysis:* information is pushed from inputs to outputs.
  - Liveness is a *backwards analysis*: information is pushed from outputs back towards inputs.
- But both make use of essentially the same algorithm.
- Numerous other analyses fall into these categories, and allow us to use a similar formulation:
  - An abstract domain (abstract relative to actual values);
  - Local rules relating information between consecutive program points around a single statement; and
  - Lattice operations like least upper bound (or *join*) or greatest lower bound (or *meet*) to relate inputs and outputs of adjoining statements.