Need for dataflow analysis

- Most optimizations require program analysis to determine safety
- This lecture: dataflow analysis
- Standard program analysis framework

Dataflow analyses

- **Live variable analysis** — register allocation, dead-code elimination
- **Reaching definitions**: what points in program does each variable definition reach? — copy, constant propagation
- **Available expressions**: which expressions computed earlier still have same value? — common sub-expression elimination

IR for data-flow analysis

- Tree IR: good for instruction selection, not so good for dataflow analysis
- Can flatten tree representation into simple nodes (a, b, c temps, labels L)

Quadruples

- Quadruple sequence is control flow graph (flowgraph)
- Nodes in graph: quadruples (not assembly statements)
- Edges in graph: ways to transfer control between quadruples (including fall-through)
- For node \( n \), use[\( n \)] is variables used, def[\( n \)] is variables defined (assigned)
- Can generate directly from AST

IR optimization

- Canonical IR
- Convert basic blocks to tree form
- Analyze, optimize
- Quadruples

AST

Canonical IR

Instruction selection

Flatten

Register allocation

Assembly code

Analyse, optimize

Quadruples
Converting to quadruples

- Conversion is tree simplification that aggressively adds new temporaries

\[
\begin{align*}
\text{MOVE} & \quad a \quad b \quad c \quad a \\
\text{MOVE} & \quad t \quad a \quad b \\
a = b + (c \times a) & \quad t = c \times a \\
a = b + t & \quad \text{MOVE} \quad a \quad +(b, \times(c,a))
\end{align*}
\]

Converting back to tree

- Convert quadruples to simple trees
- Look for temporaries in statement sequence used and defined only once
- Move definition just before use
- Glue tree, eliminating temporary

\[
\begin{align*}
t = c \times a & \quad \text{MOVE} \quad t \quad \times(c,a) \\
\cdots & \quad \text{MOVE} \quad a \quad +(b, \times(c,a)) \\
a = b + t & \quad \text{MOVE} \quad a \quad +(b,t)
\end{align*}
\]

- Requires dataflow analyses to do right (reaching definitions, available expressions)

Def & Use

\[
\begin{array}{ccc}
 n & \text{def}[n] & \text{use}[n] \\
 a = b \text{ OP } c & a & b, c \\
a = [a] = b & a & b \\
\text{if a goto L1} \text{ else goto L2} & & \\
a = f(\ldots) & a & -- \\
f(\ldots) & & -- \\
\end{array}
\]

Live variable analysis

- Useful even for IR: dead code elimination
- Output: in[n] and out[n] associated with every node n in flowgraph
- Constraints:
  \[
  \text{in}[n] \supseteq \text{use}[n] \quad \text{in}[n] \cup \text{def}[n] \supseteq \text{out}[n] \quad \text{out}[n] \supseteq \text{in}[n'] \text{ for all successors } n' \text{ of } n
  \]
- Dataflow equations:
  \[
  \text{in}[n] = \text{use}[n] \cup (\text{out}[n] - \text{def}[n]) \quad \text{out}[n] = \bigcup_{n'} \text{in}[n']
  \]

Reaching definitions analysis

- Question: what uses in program does a given variable definition reach?
- Used for constant propagation, copy propagation
  - If only one definition reaches a particular use, can replace use by definition
  - Copy propagation requires that copied value still has same value – use available expressions
- Input: flowgraph
- Output: in[n], out[n] is set of nodes defining some variable such that def may reach beginning, end of n

Reaching definitions

\[
\begin{array}{c}
W \\
X \\
Y \\
Z
\end{array}
\begin{array}{c}
b = a + 2 \\
c = b \times b \\
b = c + 1 \\
\text{return } b \times a
\end{array}
\begin{array}{c}
\text{out}: W \\
\text{in}: W \\
\text{out}: X, W \\
\text{in}: X, W \\
\text{out}: X, Y \\
\text{in}: X, Y
\end{array}
\]
**Define:** 
defs(x) is the set of nodes defining var \( x \)

**Define:** 
gen[n], kill[n]

\[
a = b \oplus c
\]

\[
a = [b]
\]

\[
[a] = b
\]

\[
\text{if a goto L1 else goto L2}
\]

\[
a = f(\ldots)
\]

\[
(a) \rightarrow \text{defs} - \{ n \}
\]

\[
[a] = b \{ \}
\]

\[
goto L \{ \}
\]

\[
\text{if a goto L1 else goto L2} \{ \}
\]

\[
f(\ldots) \{ \}
\]

\[
\text{out}[n] \supseteq \text{gen}[n]
\]

“A definition made by \( n \) at least reaches \( n \)'s output”

\[
in[n'] \supseteq \text{out}[n] \text{ (if } n' \text{ is successor of } n \)
\]

“Definitions reach node \( n' \) if they exit any predecessor \( n \)”

\[
\text{out}[n] \cup \text{kill}[n] = \text{in}[n]
\]

“A definition that reaches the input either reaches the output or is killed”

**Data-flow equations**

\[
\text{in}[n'] = \bigcup \{ n \in \text{prev}(n') \} \text{ out}[n]
\]

\[
\text{out}[n] = \text{gen}[n] \cup (\text{in}[n] - \text{kill}[n])
\]

- Algorithm: init \( \text{in}[n], \text{out}[n] \) with empty sets, apply equations as assignments until no progress (usual representation: bit vector)
- Eventually all equations satisfied
- Will terminate because \( \text{in}[n], \text{out}[n] \) can only grow, can be no larger than set of all defs
- Finds minimal solution to constraint eqns: accurate

**Def-use chains**

- Reaching definitions tells which nodes a def can reach
- If node uses same variable, definition affects node (conservatively)
- Def-use (du-) chain: def node + all nodes with affected uses
- Use-def (ud-) chain: use node + all nodes with defs that might affect use

**Webs**

- du-chain, ud-chain intersect if share some use or definition
- web: maximal set of intersecting du, ud-chains – disjoint set union algorithm with path compression: computable in nearly linear time
- Same variable may comprise multiple non-interacting webs: permits more optimization
Webs

- Register allocation by webs avoids false conflicts

```c
int i;
for (i = 0; i<n; i++) {... }  // no use/def pairs!
... for (i = 0; i<n; i++) {... }
```

- Two different webs: can allocate i to two different registers

Register allocation

1. use reaching definitions to compute all related uses and defs
2. compute disjoint webs, rename all temporaries to their web names
3. run register allocation as before: fewer interfering temporaries

Forward vs. Backward

- Liveness: backward analysis
  
  $$in[n] = use[n] \cup (out[n] - def[n])$$
  $$out[n] = \bigcup_{n' \in \text{succ}[n]} in[n']$$

- Reaching definitions: forward analysis
  
  $$out[n] = gen[n] \cup (in[n] - kill[n])$$
  $$in[n'] = \bigcup_{n \in \text{prev}[n']} out[n]$$

Dataflow analysis

- Many dataflow analyses characterized simply by
  
  - forward vs. backward analysis
  - gen[n]
  - kill[n]
  
  - Use of intersection vs. union when combining data from several nodes (operator $\cap$)
    
    $$out[n] = gen[n] \cup (in[n] - kill[n])$$
    $$in[n'] = \bigcap_{n \in \text{prev}[n']} out[n]$$

Available expressions

- Idea: want to perform common subexpression elimination

```
  a = x+1  a = x+1
  ... b = x+1  b = a
```

- Transformation is safe if original x+1 is an available expression (still computes same value)

Dataflow values

- Let $in[n]$, $out[n]$ be sets of nodes whose computed expression is available at $n$
  
  $$n \quad gen[n] \quad kill[n]$$
  $$a=b \quad OP \quad c \quad (n) - kill[n] \quad \text{uses}(a)$$
  $$a=[b] \quad (n) - kill[n] \quad \text{uses}(a)$$
  $$[a]=b \quad \{} \quad \text{uses}([x]) \quad \text{for all x that may be equal to a}$$
  $$a=f(b_1,...,b_n) \quad \{} \quad \text{uses}([x]) \quad \text{for all x}$$
  $$\text{other} \quad \{} \quad \{}$$
Constraints

\[ \text{out}[n] \supseteq \text{gen}[n] \]
“An expression made available by \( n \) at least reaches \( n \)'s output”

\[ \text{in}[n'] \subseteq \text{out}[n] \] (if \( n' \) is succ. of \( n \))
“An expression is available at \( n' \) only if it is available at every predecessor \( n \)”

\[ \text{out}[n] \cup \text{kill}[n] \supseteq \text{in}[n] \]
“An expression available on input is either available on output or killed”

Dataflow equations

\[ \text{out}[n] \supseteq \text{gen}[n] \]

\[ \text{in}[n'] \subseteq \text{out}[n] \] (if \( n' \) is succ. of \( n \))

\[ \text{out}[n] \cup \text{kill}[n] \supseteq \text{in}[n] \]

Equations for iterative solution:

\[ \text{out}[n] = \text{gen}[n] \cup (\text{in}[n] - \text{kill}[n]) \]

\[ \text{in}[n'] = \bigcap_{n \in \text{pred}[n']} \text{out}[n] \]

\[ \bigcap = \bigcap \] Starting condition:

\[ \text{in}[n] \] is set of all nodes
\[ \text{in}[\text{start}] = \emptyset \]

Summary

• Tree IR makes dataflow more difficult
• Saw reaching definitions, available expressions analyses
• How to use reaching definitions for better register allocations via webs
• Next time: a theory to explain why iterative solving works