Data-flow Analysis: Theoretical Foundations - Part 1

Y.N. Srikant

Department of Computer Science
Indian Institute of Science
Bangalore 560 012

NPTEL Course on Compiler Design

Foundations of Data-flow Analysis

- Basic questions to be answered
  1. Under what situations is the iterative DFA algorithm correct?
  2. How precise is the solution produced by it?
  3. Will the algorithm converge?
  4. What is the meaning of a "solution"?
- The above questions can be answered accurately by a DFA framework
- Further, reusable components of the DFA algorithm can be identified once a framework is defined
- A DFA framework $(D, V, \wedge, F)$ consists of

D	:	A direction of the dataflow, either forward or backward
V	:	A domain of values
$\wedge$	:	A meet operator $(V, \wedge)$ form a semi-lattice
F	:	A family of transfer functions, $V \longrightarrow V$ F includes constant transfer functions for the ENTRY/EXIT nodes as well

Semi-Lattice

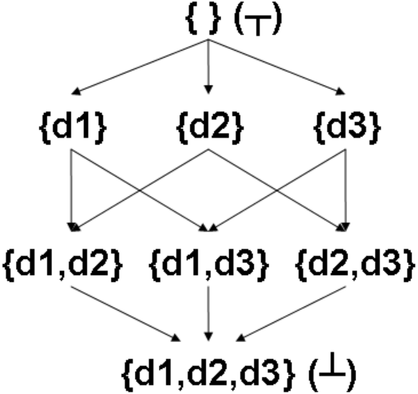
- A semi-lattice is a set V and a binary operator $\wedge$, such that the following properties hold
  1. V is closed under $\wedge$
  2. $\wedge$ is idempotent ($x \wedge x = x$), commutative ($x \wedge y = y \wedge x$), and associative ($x \wedge (y \wedge z) = (x \wedge y) \wedge z$)
  3. It has a *top* element, $\top$, such that $\forall\ x \in V,\ \top \wedge x = x$
  4. It may have a *bottom* element, $\bot$, such that $\forall x \in\ V,\ \bot \wedge x = \bot$
- The operator $\wedge$ defines a partial order $\leq$ on V, such that $x \leq y\ iff\ x \wedge y = x$
- Any two elements x and y in a semi-lattice have a greatest lower bound (*glb*), g, such that $g = x \wedge y,\ g \leq x,\ g \leq y$, and if $z \leq x,\ and\ z \leq y$, then $z \leq g$

Semi-Lattice of Reaching Definitions

- 3 definitions, {d1,d2,d3}
- V is the set of all subsets of {d1,d2,d3}
- $\wedge$ is $\cup$
- The diagram (next slide) shows the partial order relation induced by $\wedge$ (i.e., $\cup$)
- Partial order relation is $\supseteq$
- An arrow, $y \rightarrow x$ indicates $x \supseteq y$ ($x \leq y$)
- Each set in the diagram is a data-flow value
- Transitivity is implied in the diagram ($a \rightarrow b$ & $b \rightarrow c$ imples $a \rightarrow c$)
- An ascending chain: ($x_1 < x_2 < ... < x_n$)
- Height of a semi-lattice: largest number of '<' relations in any ascending chain
- Semi-lattices in our DF frameworks will be of finite height

Lattice Diagram of Reaching Definitions

$y \rightarrow x$ indicates $x \supseteq y$ ($x \leq y$)



Transfer Functions

$F: V \rightarrow V$ has the following properties

1. F has an identity function, $I(x) = x$, for all $x \in V$
2. F is closed under composition, *i.e.,* for $f, g \in F,\ f.g \in F$

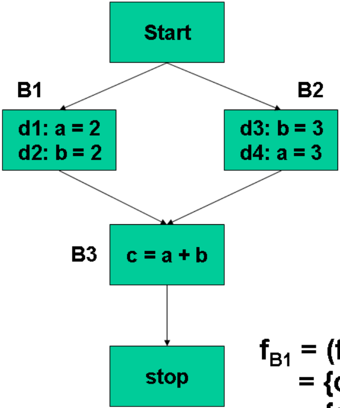
Example: Again considering the R-D problem

- Assume that each quadruple is in a separate basic block
- $OUT[B] = GEN[B] \cup (IN[B] - KILL[B])$
- In its general form, this becomes $f(x) = G \cup (x - K)$
- Let $f_1(x) = G_1 \cup (x - K_1)$ and $f_2(x) = G_2 \cup (x - K_2)$ be the transfer functions of two basic blocks $B1$ and $B2$
- Identity function exists here (when both G and K (GEN and $KILL$) are empty)

Transfer Functions

- If control flows from $B1$ to $B2$, then $f_2(f_1(x)) = G_2 \bigcup((G_1 \cup (x - K_1)) - K_2)$
- The right side above is algebraically equivalent to $(G_2 \cup (G_1 - K_2)) \bigcup(x - (K_1 \cup K_2))$
- If we let $K = K_1 \cup K_2$ and $G = G_2 \cup (G_1 - K_2)$, then $f_2(f_1(x))$ is of the same form as $f(x) = G \cup (x - K)$, and composition is proved to be true

Reaching Definitions Framework - Example



Transfer functions:

$f_{d1}(x) = \{d1\} \cup (x - \{d4\})$
$f_{d2}(x) = \{d2\} \cup (x - \{d3\})$
$f_{d3}(x) = \{d3\} \cup (x - \{d2\})$
$f_{d4}(x) = \{d4\} \cup (x - \{d1\})$
$f_{d5}(x) = \{d5\} \cup (x - \Phi)$

Transfer functions for start and stop blocks are identity functions

$$\begin{aligned}
f_{B1} &= (f_{d2}.f_{d1})(x) \\
&= \{d2\} \cup (\{d1\} \cup (x - \{d4\}) - \{d3\} \\
&= \{d1,d2\} \cup (x - \{d3,d4\}) \\
f_{B2} &= (f_{d4}.f_{d3})(x) \\
&= \{d3,d4\} \cup (x - \{d1,d2\}) \\
f_{B3} &= f_{d5} = \{d5\} \cup x
\end{aligned}$$

Monotone Frameworks

- A DF framework $(D, F, V, \wedge)$ is monotone, if $\forall x, y \in V,\ f \in F,\ x \leq y \Rightarrow f(x) \leq f(y)$, OR $f(x \wedge y) \leq f(x) \wedge f(y)$
- The reaching definitions lattice is monotone
- **Proof:** $\wedge$ is $\cup$. Therefore, we need to prove that
  $f(x \cup y) \supseteq f(x) \cup f(y)$
  $f(x \cup y) = G \cup (x \cup y - K)$
  $f(x) \cup f(y) = (G \cup (x - K)) \cup (G \cup (y - K))$
  $= G \cup (x - K) \cup (y - K)$
  $= G \cup (x \cup y) - K) = f(x \cup y)$
  Therefore, the Reaching Definitions framework is monotone

Distributive Frameworks

- A DF framework is distributive, if
  $\forall x, y \in V,\ f \in F,\ f(x \wedge y) = f(x) \wedge f(y)$
- Distributivity $\Rightarrow$ monotonicity, but not vice-versa

  proof: If $a = b,\ a \wedge b = a$, so, $a \leq b$ (by definition of $\leq$)
  From the definition of distributivity, we know that
  $f(x \wedge y) = f(x) \wedge f(y)$
  Substituting $f(x \wedge y)$ for a and $f(x) \wedge f(y)$ for b,
  in $a \leq b$, we get $f(x \wedge y) \leq f(x) \wedge f(y)$,
  which is the requirement of monotonocity
- The reaching definitions lattice is distributive

  Proof: We have already proved during the proof of
  monotonocity of the RD framework, that
  $f(x \cup y) = f(x) \cup f(y)$. This proves distributivity also

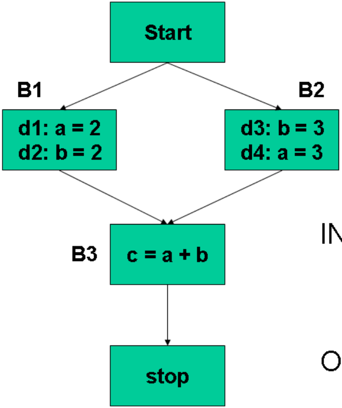
Iterative Algorithm for DFA (forward flow)

$\{OUT[B1] = v_{init};$
for each block $B \neq B1$ do $OUT[B] = \top;$
while (*changes to any OUT occur*) do
  for each block $B \neq B1$ do $\{$

$$IN[B] = \bigwedge_{P\ a\ predecessor\ of\ B} OUT[P];$$
$$OUT[B] = f_B(IN[B]);$$

  $\}$
$\}$

Reaching Definitions Framework - Example contd.



f_{B1} = {d1,d2} U (x – {d3,d4})

f_{B2} = {d3,d4} U (x – {d1,d2})

f_{B3} = {d5} U x

$IN[B] = \wedge_{P,\text{ a predecessor of B}}\ OUT[P]$

$= \cup_{P,\text{a predecessor of B}}\ OUT[P]$

$OUT[B] = f_B(IN[B])$

Needs 3 iterations to converge
IN[B1] = IN[B2] = Φ; OUT[B1] = {d1,d2}; OUT[B2] = {d3,d4}
IN[B3] = OUT[B1] U OUT[B2] = {d1,d2,d3,d4}
OUT[B3] = {d5} U IN[B3] = {d1,d2,d3,d4,d5}

Properties of the Iterative DFA Algorithm

- If the iterative algorithm converges, the result is a solution to the DF equations

  Proof: If the equations are not satisfied by the time the loop ends, atleast one of the *OUT* sets changes and we iterate again
- If the framework is monotone, then the solution found is the maximum fixpoint (MFP) of the DF equations
  An MFP solution is such that in any other solution, values of $IN[B]$ and $OUT[B]$ are $\leq$ the corresponding values of the MFP (i.e., less precise)

  Proof: We can show by induction that the values of $IN[B]$ and $OUT[B]$ only decrease (in the sense of $\leq$ relation) as the algorithm iterates

Properties of the Iterative DFA Algorithm (2)

- If the semi-lattice of the framework is monotone and is of finite height, then the algorithm is guaranteed to converge

  Proof: Dataflow values decrease with each iteration
  Max no. of iterations = height of the lattice $\times$ no. of nodes in the flow graph

Meaning of the Ideal Data-flow Solution

- Find all possible execution paths from the start node to the beginning of B
- (Assuming forward flow) Compute the data-flow value at the end of each path (using composition of transfer functions) and apply the $\wedge$ operator to these values to find their *glb*
- No execution of the program can produce a *smaller* value for that program point

$$IDEAL[B] \;=\; \bigwedge_{P,\ a\ possible\ execution\ path\ from\ start\ node\ to\ B} f_P(v_{init})$$

- Answers greater (in the sense of $\leq$) than IDEAL are incorrect (one or more execution paths have been ignored)
- Any value smaller than or equal to IDEAL is conservative, *i.e.,* safe (one or more infeasible paths have been included)
- Closer the value to IDEAL, more precise it is

Meaning of the Meet-Over-Paths Data-flow Solution

- Since finding all execution paths is an undecidable problem, we approximate this set to include all paths in the flow graph

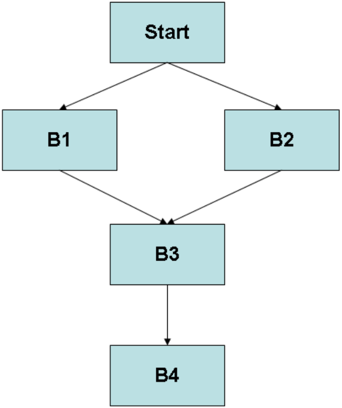
$$MOP[B] = \bigwedge_{P,\ a\ path\ from\ start\ node\ to\ B} f_P(v_{init})$$

- $MOP[B] \leq IDEAL[B]$, since we consider a superset of the set of execution paths

Meaning of the Maximum Fixpoint Data-flow Solution

- Finding all paths in a flow graph may still be impossible, if it has cycles
- The iterative algorithm does not try this
  - It visits all basic blocks, not necessarily in execution order
  - It applies the $\wedge$ operator at each join point in the flow graph
  - The solution obtained is the Maximum Fixpoint solution (MFP)
- If the framework is distributive, then the MOP and MFP solutions will be identical
- Otherwise, with just monotonicity, $MFP \leq MOP \leq IDEAL$, and the solution provided by the iterative algorithm is safe

Example to show $MFP \leq MOP$



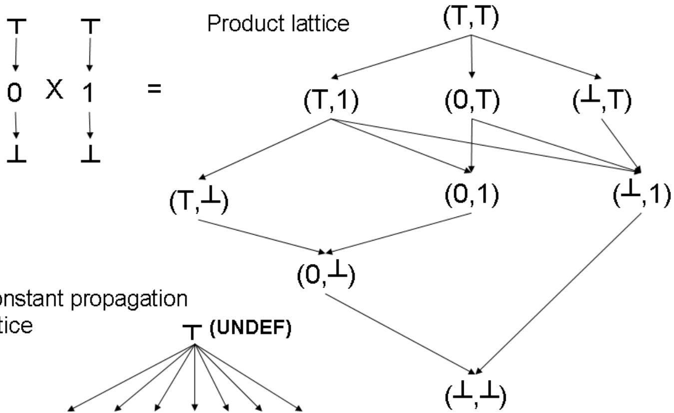
Example to show $MFP \leq MOP$ (2)

- There are two paths from Start to B4:
  $Start \rightarrow B1 \rightarrow B3 \rightarrow B4$ and $Start \rightarrow B2 \rightarrow B3 \rightarrow B4$
- $MOP[B4] = ((f_{B3} \cdot f_{B1}) \wedge (f_{B3} \cdot f_{B2}))(v_{init})$
- In the iterative algorithm, if we chose to visit the nodes in the order $(Start, B1, B2, B3, B4)$, then
  $IN[B4] = f_{B3}(f_{B1}(v_{init}) \wedge f_{B2}(v_{init}))$
- Note that the $\wedge$ operator is being applied differently here than in the *MOP* equation
- The two values above will be equal only if the framework is distributive
- With just monotonicity, we would have $IN[B4] \leq MOP[B4]$

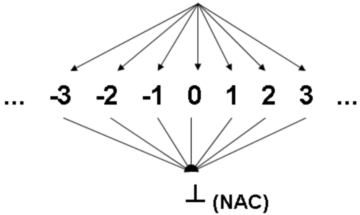
Constant Propagation Framework - Data-flow Values

- The lattice for a single variable in the CP framework is shown in the next slide
- An example of product of two lattices is in the next slide
- DF values in the RD framework can also be considered as
    - values in a product of lattices of definitions
    - one lattice for each definition, with ϕ as $\top$ and {d} as the only other element
- The lattice of the DF values in the CP framework
    - Product of the semi-lattices of the variables (one lattice for each variable)

Product of Two Lattices and Lattice of Constants



Constant propagation lattice



$|S_1 \times S_2| = |S_1| \times |S_2|$
$(a,b) \leq (c,d)$ iff $a \leq c$ & $b \leq d$

CP Framework - The $\wedge$ (meet) Operator

- In a product lattice, $(a_1, b_1) \leq (a_2, b_2)$ iff $a_1 \leq_A a_2$ and $b_1 \leq_B b_2$ assuming $a_1, a_2 \in A$ and $b_1, b_2 \in B$
- Each variable is associated with a map m
- $m(v)$ is the abstract value (as in the lattice) of the variable v in a map m
- Each element of the product lattice is a similar, but "larger" map m
    - which is defined for all variables, and
    - where $m(v)$ is the abstract value of the variable v
- Thus, $m \leq m'$ (in the product lattice), iff for all variables v, $m(v) \leq m'(v)$, OR, $m \wedge m' = m''$, if $m''(v) = m(v) \wedge m'(v)$, for all variables v

Transfer Functions for the CP Framework

- Assume one statement per basic block
- Transfer functions for basic blocks containing many statements may be obtained by composition
- $m(v)$ is the abstract value of the variable v in a map m.
- The set F of the framework contains transfer functions which accept maps and produce maps as outputs
- F contains an identity map
- Map for the *Start* block is $m_0(v) = UNDEF$, for all variables v
- This is reasonable since all variables are undefined before a program begins

Transfer Functions for the CP Framework

- Let f_s be the transfer function of the statement s
- If $m' = f_s(m)$, then f_s is defined as follows
  1. If s is not an assignment, f_s is the identity function
  2. If s is an assignment to a variable x, then $m'(v) = m(v)$, for all $v \neq x$, provided, one of the following conditions holds
     (a) If the RHS of s is a constant c, then $m'(x) = c$
     (b) If the RHS is of the form $y + z$, then

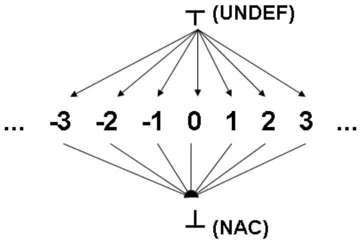
$$\begin{aligned} m'(x) &= m(y) + m(z), \text{ if } m(y) \text{ and } m(z) \text{ are constants} \\ &= NAC, \text{ if either } m(y) \text{ or } m(z) \text{ is } NAC \\ &= UNDEF, \text{ otherwise} \end{aligned}$$

     (c) If the RHS is any other expression, then $m'(x) = NAC$

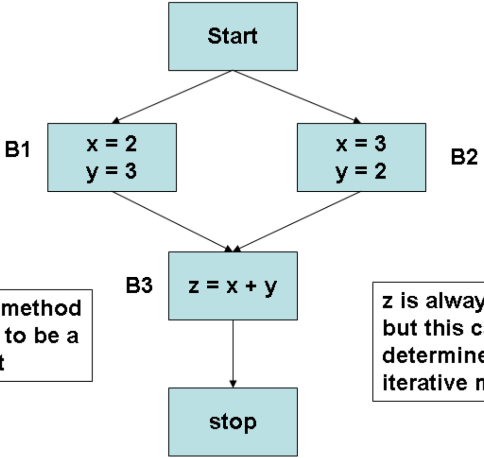
Monotonicity of the CP Framework

It must be noted that the transfer function ($m' = f_s(m)$) always produces a "lower" or same level value in the CP lattice, whenever there is a change in inputs

$m(y)$	$m(z)$	$m'(x)$
UNDEF	UNDEF	UNDEF
	c_2	UNDEF
	NAC	NAC
c_1	UNDEF	UNDEF
	c_2	$c_1 + c_2$
	NAC	NAC
NAC	UNDEF	NAC
	c_2	NAC
	NAC	NAC



Non-distributivity of the CP Framework



The iterative method determines z to be a non-constant

z is always a constant but this cannot be determined by the iterative method

Non-distributivity of the CF Framework - Example

- If f_1, f_2, f_3 are transfer functions of $B1, B2, B3$ (resp.), then $f_3(f_1(m_0) \wedge f_2(m_0)) < f_3(f_1(m_0)) \wedge f_3(f_2(m_0))$ as shown in the table, and therefore the CF framework is non-distributive

m	$m(x)$	$m(y)$	$m(z)$
m_0	UNDEF	UNDEF	UNDEF
$f_1(m_0)$	2	3	UNDEF
$f_2(m_0)$	3	2	UNDEF
$f_1(m_0) \wedge f_2(m_0)$	NAC	NAC	UNDEF
$f_3(f_1(m_0) \wedge f_2(m_0))$	NAC	NAC	NAC
$f_3(f_1(m_0))$	2	3	5
$f_3(f_2(m_0))$	3	2	5
$f_3(f_1(m_0)) \wedge f_3(f_2(m_0))$	NAC	NAC	5