General Data Flow Frameworks

Uday Khedker
(www.cse.iitb.ac.in/~uday)

Department of Computer Science and Engineering,
Indian Institute of Technology, Bombay

September 2017

Copyright

These slides constitute the lecture notes for CS618 Program Analysis course at IIT Bombay and have been made available as teaching material accompanying the book:

  (Indian edition published by Ane Books in 2013)

Apart from the above book, some slides are based on the material from the following book


These slides are being made available under GNU FDL v1.2 or later purely for academic or research use.

Outline

- Modelling General Flows
- Constant Propagation
- Strongly Live Variables Analysis (after mid-sem)
- Pointer Analyses (after mid-sem)
- Heap Reference Analysis (after mid-sem)
Part 2

Precise Modelling of General Flows

Complexity of Constant Propagation?

\[
\begin{align*}
1 &: a = b + 1 \\
2 &: a = b + 1 \\
3 &: b = c + 1 \\
4 &: c = d + 1 \\
5 &: d = 2
\end{align*}
\]

Iteration #1

\[
\begin{align*}
1 &: a = b + 1 \\
2 &: a = b + 1 \\
3 &: b = c + 1 \\
4 &: c = d + 1 \\
5 &: d = 2
\end{align*}
\]

Iteration #2

\[
\begin{align*}
1 &: a = 5 \\
2 &: a = 5 \\
3 &: b = 3 \\
4 &: c = 3 \\
5 &: d = 2
\end{align*}
\]

Iteration #3

\[
\begin{align*}
1 &: a = 5 \\
2 &: a = 5 \\
3 &: b = 3 \\
4 &: c = 3 \\
5 &: d = 2
\end{align*}
\]

Iteration #4

Loop Closures of Flow Functions

- For static analysis we need to summarize the value at \( p_2 \) by a value which is safe after any iteration.

\[
\begin{align*}
f^*(x) &= x \cap f(x) \cap f^2(x) \cap \ldots
\end{align*}
\]

- \( f^* \) is called the loop closure of \( f \).

Loop Closure Boundedness

- Boundedness of \( f \) requires the existence of some \( k \) such that

\[
f^*(x) = x \cap f(x) \cap f^2(x) \cap \ldots \cap f^{k-1}(x)
\]

- This follows from the descending chain condition

- For efficiency, we need a constant \( k \) that is independent of the size of the lattice
Loop Closures in Bit Vector Frameworks

- Flow functions in bit vector frameworks have constant Gen and Kill

\[
\begin{align*}
F^*(x) &= x \sqcap f(x) \sqcap f^2(x) \sqcap f^3(x) \sqcap \ldots \\
F^2(x) &= f(Gen \cup (x - Kill)) \\
&= Gen \cup ((Gen \cup (x - Kill)) - Kill) \\
&= Gen \cup ((Gen - Kill) \cup (x - Kill)) \\
&= Gen \cup (Gen - Kill) \cup (x - Kill) \\
F^*(x) &= x \sqcap f(x)
\end{align*}
\]

- Loop Closures of Bit Vector Frameworks are 2-bounded.

- Intuition: Since Gen and Kill are constant, same things are generated or killed in every application of \( f \).
  Multiple applications of \( f \) are not required unless the input value changes.

Larger Values of Loop Closure Bounds

- Fast Frameworks \( \equiv \) 2-bounded frameworks (eg. bit vector frameworks)
  Both these conditions must be satisfied
  - Separability
    Data flow values of different entities are independent
  - Constant or Identity Flow Functions
    Flow functions for an entity are either constant or identity

- Non-fast frameworks
  At least one of the above conditions is violated

Separability

\( f : L \rightarrow L \) is \( \langle \hat{h}_1, \hat{h}_2, \ldots, \hat{h}_m \rangle \) where \( \hat{h}_i \) computes the value of \( \hat{x}_i \)

<table>
<thead>
<tr>
<th>Separable</th>
<th>Non-Separable</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \langle \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m \rangle )</td>
<td>( \langle \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m \rangle )</td>
</tr>
<tr>
<td>( f )</td>
<td>( f )</td>
</tr>
<tr>
<td>( \langle \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m \rangle )</td>
<td>( \langle \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m \rangle )</td>
</tr>
</tbody>
</table>

Example: All bit vector frameworks Example: Constant Propagation

Example: All bit vector frameworks Example: Constant Propagation
Separability of Bit Vector Frameworks

- \( \hat{L} \) is \{0, 1\}, \( L \) is \{0, 1\}^m
- \( \wedge \) is either boolean AND or boolean OR
- \( \top \) and \( \bot \) are 0 or 1 depending on \( \wedge \).
- \( \hat{h} \) is a bit function and could be one of the following:

<table>
<thead>
<tr>
<th>Raise</th>
<th>Lower</th>
<th>Propagate</th>
<th>Negate</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \top )</td>
<td>( \top )</td>
<td>( \top )</td>
<td>( \bot )</td>
</tr>
<tr>
<td>( \bot )</td>
<td>( \bot )</td>
<td>( \bot )</td>
<td>( \bot )</td>
</tr>
</tbody>
</table>

Part 3

Constant Propagation

Example of Constant Propagation

Larger Values of Loop Closure Bounds

Composite flow function for the loop is:

\[
f((v_a, v_b, v_c, v_d)) = (v_b + 1, v_c + 1, v_d + 1, 2)
\]

\( f \) is not 2-bounded because:

- \( f((\top, \top, \top, \top)) = (\top, \top, \top, 2) \)
- \( f^2((\top, \top, \top, \top)) = (\top, \top, 3, 2) \)
- \( f^3((\top, \top, \top, \top)) = (\top, 4, 3, 2) \)
- \( f^4((\top, \top, \top, \top)) = (5, 4, 3, 2) \)

Example of Constant Propagation

MoP: \( \langle \top, \top, \top \rangle \)  \( \langle 1, 2, 3, \top \rangle \)

MFP: \( \langle \top, \top, \top \rangle \)  \( \langle 1, 2, 3, \top \rangle \)

\( n_1 \)
- \( a = 1 \)
- \( b = 2 \)
- \( c = a + b \)

\( n_2 \)
- \( c = a + b \)
- \( d = a * b \)

\( n_3 \)
- \( d = c - 1 \)
- \( a = 2 \)
- \( b = 1 \)
- \( c = a + b \)
Component Lattice for Integer Constant Propagation

- Overall lattice $L$ is a set of mappings $\forall \text{Var} \rightarrow \hat{L}$:
  \[ L = \{ \varnothing, \bot, \top \} \]

- $\cap$ and $\cap^*$ get defined by $\subseteq$ and $\hat{\cap}$:
  \[ \begin{align*}
  (x, v_1) \subseteq (y, v_2) & \iff x = y \land v_1 \hat{\subseteq} v_2 \\
  \text{OR} \quad x \mapsto v_1 \subseteq y \mapsto v_2 & \iff x = y \land v_1 \hat{\subseteq} v_2
  \end{align*} \]

- Partial order is restricted to data flow values of the same variable
  Data flow values of different variables are incomparable.

Notations for Mappings as Data Flow Values

- Accessing and manipulating a mapping $X \subseteq A \rightarrow B$:
  \[ X(a) \text{ denotes the image of } a \in A \]
  \[ X(a) \in B \]
  \[ X[a \mapsto v] \text{ changes the image of } a \text{ in } X \text{ to } v \]
  \[ X[a \mapsto v] = (X - \{(a, u) \mid u \in B\}) \cup \{(a, v)\} \]

Defining Data Flow Equations for Constant Propagation

- $\text{In}_n$ is defined as:
  \[ \text{In}_n = \begin{cases} 
  \{ (y, ud) \mid y \in \var{\text{Var}} \} & n = \text{Start} \\
  \prod_{p \in \text{pred}(n)} \text{Out}_p & \text{otherwise}
  \end{cases} \]

- $\text{Out}_n$ is defined as:
  \[ \text{Out}_n = f_n(\text{In}_n) \]

- $\text{eval}(e, X)$ is defined as:
  \[ \begin{cases} 
  nc & a \in \text{Opd}(e) \land \var{\text{Var}}, X(a) = nc \\
  ud & a \in \text{Opd}(e) \land \var{\text{Var}}, X(a) = ud \\
  -X(a) & e \text{ is } a \\
  X(a) \oplus X(b) & e \text{ is } a \oplus b
  \end{cases} \]
### Example Program for Constant Propagation

```
n1 input (e);

n2 a = 7; b = 2; f = e;
   if (f > 0) 
      a = 2;
      if (f ≥ e + 2) 
         b = c + 1;
         if (b ≥ 7) 
            f = f + 1;

n3 c = d * a;

n4 d = a + b;

n5 e = a + b;

n6 d = a + 1;
   f = f + 1
```

### Result of Constant Propagation

<table>
<thead>
<tr>
<th>Iteration #1</th>
<th>Changes in iteration #2</th>
<th>Changes in iteration #3</th>
<th>Changes in iteration #4</th>
</tr>
</thead>
<tbody>
<tr>
<td>In</td>
<td>T, T, T, T, T</td>
<td>T, T, T, T, T</td>
<td>T, T, T, T, T</td>
</tr>
<tr>
<td>Out</td>
<td>T, T, T, T, T</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In</td>
<td>7, 7, 7, 7, 7</td>
<td>2, 2, 2, 2, 2</td>
<td>1, 1, 6, 3, 1</td>
</tr>
<tr>
<td>Out</td>
<td>7, 7, 7, 7, 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In</td>
<td>7, 7, 7, 7, 7</td>
<td>2, 2, 2, 2, 2</td>
<td>2, 2, 6, 3, 1</td>
</tr>
<tr>
<td>Out</td>
<td>7, 7, 7, 7, 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In</td>
<td>2, 2, 2, 2, 2</td>
<td>2, 2, 6, 3, 1</td>
<td>2, 2, 6, 3, 1</td>
</tr>
<tr>
<td>Out</td>
<td>2, 2, 2, 2, 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In</td>
<td>2, 2, 2, 2, 2</td>
<td>2, 2, 6, 3, 1</td>
<td>2, 2, 6, 3, 1</td>
</tr>
<tr>
<td>Out</td>
<td>2, 2, 2, 2, 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In</td>
<td>2, 2, 2, 2, 2</td>
<td>2, 2, 6, 3, 1</td>
<td>2, 2, 6, 3, 1</td>
</tr>
<tr>
<td>Out</td>
<td>2, 2, 2, 2, 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In</td>
<td>2, 2, 2, 2, 2</td>
<td>2, 2, 6, 3, 1</td>
<td>2, 2, 6, 3, 1</td>
</tr>
<tr>
<td>Out</td>
<td>2, 2, 2, 2, 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In</td>
<td>2, 2, 2, 2, 2</td>
<td>2, 2, 6, 3, 1</td>
<td>2, 2, 6, 3, 1</td>
</tr>
<tr>
<td>Out</td>
<td>2, 2, 2, 2, 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In</td>
<td>2, 2, 2, 2, 2</td>
<td>2, 2, 6, 3, 1</td>
<td>2, 2, 6, 3, 1</td>
</tr>
<tr>
<td>Out</td>
<td>2, 2, 2, 2, 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>In</td>
<td>2, 2, 2, 2, 2</td>
<td>2, 2, 6, 3, 1</td>
<td>2, 2, 6, 3, 1</td>
</tr>
<tr>
<td>Out</td>
<td>2, 2, 2, 2, 2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Result of Constant Propagation

```
n1 input (e);

n2 a = 7; b = 2; f = e;
   if (f > 0) 
      a = 2;
      if (f ≥ e + 2) 
         b = c + 1;
         if (b ≥ 7) 
            f = f + 1;

n3 c = d * a;

n4 d = a + b;

n5 e = a + b;

n6 d = a + 1;
   f = f + 1
```

### Example Program for Constant Propagation

For readability, we have combined many statements in a single block. However, constant propagation requires every basic block to contain a single statement because of the presence of dependent parts in flow functions.
Monotonicity of Constant Propagation

Proof obligation: $X_1 \sqsubseteq X_2 \Rightarrow f_n(X_1) \sqsubseteq f_n(X_2)$

where,

\[
\begin{align*}
X[y \mapsto c] & \quad n \text{ is } y = c, y \in \text{Var}, c \in \text{Const} \\
X[y \mapsto nc] & \quad n \text{ is } \text{input}(y), y \in \text{var} \\
X[y \mapsto X(z)] & \quad n \text{ is } y = z, y \in \text{Var}, z \in \text{Var} \\
X[y \mapsto \text{eval}(e, X)] & \quad n \text{ is } y = e, y \in \text{Var}, e \in \text{Expr} \\
X & \quad \text{otherwise}
\end{align*}
\]

\[f_n(X) = \begin{cases}
X[y \mapsto c] & \quad n \text{ is } y = c, y \in \text{Var}, c \in \text{Const} \\
X[y \mapsto nc] & \quad n \text{ is } \text{input}(y), y \in \text{var} \\
X[y \mapsto X(z)] & \quad n \text{ is } y = z, y \in \text{Var}, z \in \text{Var} \\
X[y \mapsto \text{eval}(e, X)] & \quad n \text{ is } y = e, y \in \text{Var}, e \in \text{Expr} \\
X & \quad \text{otherwise}
\end{cases}\]

- The proof obligation trivially follows for cases C1, C2, C3, and C5
- For case C4, it requires showing $X_1 \sqsubseteq X_2 \Rightarrow \text{eval}(e, X_1) \sqsubseteq \text{eval}(e, X_2)$
- which follows from the definition of $\text{eval}(e, X)$

Why is Constant Propagation Non-Distributive?

Possible combinations due to merging

- Correct combination.

Non-Distributivity of Constant Propagation

- $x = (1, 2, 3, ?)$ (Along $\text{Out}_{n_1} \rightarrow \text{In}_{n_2}$)
- $y = (2, 1, 3, 2)$ (Along $\text{Out}_{n_1} \rightarrow \text{In}_{n_2}$)
- Function application before merging
  \[
  f(x) \cap f(y) = f((1, 2, 3, ?)) \cap f((2, 1, 3, 2)) = (1, 2, 3, 2) \cap (2, 1, 3, 2) = (\top, \top, 3, 2)
  \]
- Function application after merging
  \[
  f(x \cap y) = f((1, 2, 3, ?) \cap (2, 1, 3, 2)) = f((\top, \top, 3, 2)) = (\top, \top, \top, \top)
  \]
- $f(x \cap y) \sqsubseteq f(x) \cap f(y)$

Why is Constant Propagation Non-Distributive?

Possible combinations due to merging

- Correct combination.
Why is Constant Propagation Non-Distributive?

Possible combinations due to merging

- Wrong combination.
- Mutually exclusive information.
- No execution path along which this information holds.

Tutorial Problem on Constant Propagation

How many iterations do we need?

Every back edge occurs only once in the ifp from $n_3$ to $n_1$ that goes via $n_5$, $n_7$, and $n_9$.

- $5 + 1$ iterations for computing data flow values (+1 iteration to detect convergence)
Tutorial Problem on Constant Propagation

And now how many iterations do we need?

Back edge $n_{10} \rightarrow n_1$ needs to be traversed once each for back edges $n_9 \rightarrow n_8$, $n_7 \rightarrow n_6$, $n_5 \rightarrow n_4$, and $n_3 \rightarrow n_2$ (in that order).

$\Rightarrow$ 8 + 1 iterations.

Boundedness of Constant Propagation

Summary flow function:
(data flow value at node 7)

$$f(\langle v_a, v_b, v_c \rangle) = \langle 1 \cap (v_b + 1), (v_c + 1), (v_a + 1) \rangle$$

$$f^0(T) = \langle \top, \top, \top \rangle$$
$$f^1(T) = \langle 1, \top, \top \rangle$$
$$f^2(T) = \langle 1, \top, 2 \rangle$$
$$f^3(T) = \langle 1, 3, 2 \rangle$$
$$f^4(T) = \langle 1, 3, 2 \rangle$$
$$f^5(T) = \langle 1, 3, \bot \rangle$$
$$f^6(T) = \langle 1, 1, \bot \rangle$$
$$f^7(T) = \langle 1, 1, \bot \rangle$$

$$f^*(T) = \prod_{i=0}^6 f^i(T)$$
**Conditional Constant Propagation**

An execution trace of the program when the value read for variable e is some number \( x \leq 0 \) (otherwise the loop will not be entered)

\[
\begin{align*}
&\text{n}_1 \quad \text{input (e)}; \\
&\text{n}_2 \quad a = 7; b = 2; f = e; \\
&\quad \text{if (} f > 0 \text{)} \\
&\quad \text{true} \\
&\quad a = 2; \\
&\quad \text{if (} f \geq e + 2 \text{)} \\
&\quad \text{true} \\
&\quad b = c + 1; \\
&\quad \text{if (} b \geq 7 \text{)} \\
&\quad \text{true} \\
&\quad f = f + 1; \\
&\quad \text{true} \\
&\quad d = a + b; \\
&\quad \text{true} \\
&\quad e = a + b; \\
&\text{n}_{10} \quad e = a + b;
\end{align*}
\]

Boundedness of Constant Propagation

The moral of the story:

- The data flow value of every variable could change twice
- In the worst case, only one change may happen in every step of a function application
- Maximum number of steps: \( 2 \times |\text{Var}| \)
- Boundedness parameter \( k \) is \( (2 \times |\text{Var}|) + 1 \)

\[
\begin{align*}
&\text{n}_1 \quad \text{input (e)}; \\
&\text{n}_2 \quad a = 7; b = 2; f = e; \\
&\quad \text{if (} f > 0 \text{)} \\
&\quad \text{false} \\
&\quad a = 2; \\
&\quad \text{if (} f \geq e + 2 \text{)} \\
&\quad \text{false} \\
&\quad b = c + 1; \\
&\quad \text{if (} b \geq 7 \text{)} \\
&\quad \text{false} \\
&\quad f = f + 1; \\
&\quad \text{false} \\
&\quad c = d + a; \\
&\quad \text{false} \\
&\quad d = a + b; \\
&\quad \text{false} \\
&\quad e = a + b; \\
&\text{n}_{10} \quad e = a + b;
\end{align*}
\]
Conditional Constant Propagation

Data Flow Equations for Conditional Constant Propagation

\[
\begin{align*}
\text{In}_n &= \begin{cases}
\langle \text{reachable}, B_1 \rangle & \text{n is Start} \\
\bigcap_{p \in \text{pred}(n)} g_{p \rightarrow n}(\text{Out}_p) & \text{otherwise}
\end{cases} \\
\text{Out}_n &= \begin{cases}
\langle \text{reachable}, f_n(X) \rangle & \text{In}_n = \langle \text{reachable}, X \rangle \\
\langle \text{notReachable}, T \rangle & \text{otherwise}
\end{cases}
\end{align*}
\]

\[
g_{m \rightarrow n}(s, X) = \begin{cases}
\langle s, X \rangle & \text{label}(m \rightarrow n) \in \text{evalCond}(m, X) \\
\langle \text{notReachable}, T \rangle & \text{otherwise}
\end{cases}
\]

- \text{label}(m \rightarrow n) is \text{T} or \text{F} if edge \text{m} \rightarrow \text{n} is a conditional branch
- Otherwise \text{label}(m \rightarrow n) is \text{T}
- \text{evalCond}(m, X) evaluates the condition in block \text{m} using the data flow values in \text{X}

Lattice for Conditional Constant Propagation

\[
\text{notReachable} \quad \times \quad \hat{L} \quad \times \quad \hat{L} \quad \times \quad \cdots \quad \times \quad \hat{L}
\]

- Let \langle s, X \rangle denote an augmented data flow value where \text{s} \in \{\text{reachable, notReachable}\} and \text{X} \in L.
- If we can maintain the invariant \langle s = \text{notReachable} \Rightarrow X = T \rangle, then the meet can be defined as

\[
\langle s_1, X_1 \rangle \cap \langle s_2, X_2 \rangle = \langle s_1 \cap s_2, X_1 \cap X_2 \rangle
\]

Compile Time Evaluation of Conditions using the Data Flow Values

\[
\text{evalCond}(m, X)
\]

| \{T, F\} | Block \text{m} does not have a condition, or some variable in the condition is \perp in \text{X} |
| {} | No variable in the condition in block \text{m} is \perp in \text{X}, but some variable is \top in \text{X} |
| \{T\} | The condition in block \text{m} evaluates to \text{T} with the data flow values in \text{X} |
| \{F\} | The condition in block \text{m} evaluates to \text{F} with the data flow values in \text{X} |
### Conditional Constant Propagation

<table>
<thead>
<tr>
<th>Iteration #1</th>
<th>Changes in Iteration #2</th>
<th>Changes in Iteration #3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$I_{in}$</td>
<td>$R \langle T, T, T, T, I \rangle$</td>
<td></td>
</tr>
<tr>
<td>$O_{in}$</td>
<td>$R \langle T, T, T, T, I \rangle$</td>
<td></td>
</tr>
<tr>
<td>$I_{out}$</td>
<td>$R \langle T, T, T, T, I \rangle$</td>
<td></td>
</tr>
<tr>
<td>$O_{out}$</td>
<td>$R \langle T, T, T, T, I \rangle$</td>
<td></td>
</tr>
<tr>
<td>$I_{m}$</td>
<td>$R \langle T, T, T, T, I \rangle$</td>
<td>$R \langle 1, 2, T, 3, I \rangle$</td>
</tr>
<tr>
<td>$O_{m}$</td>
<td>$R \langle T, T, T, T, I \rangle$</td>
<td>$R \langle 2, 2, T, 3, I \rangle$</td>
</tr>
<tr>
<td>$I_{m}$</td>
<td>$R \langle T, T, T, T, I \rangle$</td>
<td>$R \langle 2, 2, T, 3, I \rangle$</td>
</tr>
<tr>
<td>$O_{m}$</td>
<td>$R \langle T, T, T, T, I \rangle$</td>
<td>$R \langle 2, 2, T, 3, I \rangle$</td>
</tr>
</tbody>
</table>

### Strongly Live Variables Analysis

- A variable is strongly live if
  - it is used in a statement other than assignment statement, or (same as simple liveness)
  - it is used in an assignment statement defining a variable that is strongly live (different from simple liveness)
- Killing: An assignment statement, an input statement, or BI (this is same as killing in simple liveness)
- Generation: A direct use or a use for defining values that are strongly live (this is different from generation in simple liveness)
Live Variables Analysis: Simple and Strong Liveness

- A variable is live at a program point if its current value is likely to be used later.
- We want to compute the smallest set of variables that are live.
- Simple liveness considers every use of a variable as useful.
- Strong liveness checks the liveness of the result before declaring the operands to be live.
- Strong liveness is more precise than simple liveness.

Distributivity of Strongly Live Variables Analysis (1)

We need to prove that
\[ \forall X_1, X_2 \in L, \quad f_n(X_1 \cup X_2) = f_n(X_1) \cup f_n(X_2) \]

- Intuitively,
  - The value does not depend on the argument \( X \)
  - Incomparable results cannot be produced
    (A fixed set of variable are excluded or included)

- Formally,
  - We prove it for \( \text{input}(y), \text{use}(y), y = e, \) and empty statements independently.

Data Flow Equations for Strongly Live Variables Analysis

- \( \text{ln}_n = f_n(\text{Out}_n) \)
- \( \text{Out}_n = \begin{cases} \text{Bl} \cup \sum_{s \in \text{succ}(n)} \text{ln}_s & \text{n is End} \\ \text{ln}_n & \text{otherwise} \end{cases} \)

where,
\[ f_n(x) = \begin{cases} (X - \{y\}) \cup (\text{Opd}(e) \cap \text{Var}) & \text{n is input}(y) \\ X - \{y\} & \text{n is use}(y) \\ X & \text{otherwise} \end{cases} \]
Distributivity of Strongly Live Variables Analysis (2)

- For `input(y)` statement:
  \[ f_n(X_1 \cup X_2) = (X_1 \cup X_2) - \{y\} = (X_1 - \{y\}) \cup (X_2 - \{y\}) = f_n(X_1) \cup f_n(X_2) \]
- For `use(y)` statement:
  \[ f_n(X_1 \cup X_2) = (X_1 \cup X_2) \cup \{y\} = (X_1 \cup \{y\}) \cup (X_2 \cup \{y\}) = f_n(X_1) \cup f_n(X_2) \]
- For empty statement:
  \[ f_n(X_1 \cup X_2) = X_1 \cup X_2 = f_n(X_1) \cup f_n(X_2) \]

Distributivity of Strongly Live Variables Analysis (3)

For \( y = e \) statement: Let \( Y = Opd(e) \cap \text{Var} \). There are three cases:

1. \( y \in X_1 \), \( y \in X_2 \):
   \[ f_n(X_1 \cup X_2) = ((X_1 \cup X_2) - \{y\}) \cup Y = (X_1 - \{y\}) \cup (X_2 - \{y\}) \cup Y = ((X_1 - \{y\}) \cup Y) \cup ((X_2 - \{y\}) \cup Y) = f_n(X_1) \cup f_n(X_2) \]

2. \( y \in X_1 \), \( y \notin X_2 \):
   \[ f_n(X_1 \cup X_2) = ((X_1 \cup X_2) - \{y\}) \cup Y = ((X_1 - \{y\}) \cup Y) \cup (X_2) = f_n(X_1) \cup f_n(X_2) \]
   \( y \notin X_2 \Rightarrow f_n(X_2) \) is identity

3. \( y \notin X_1 \), \( y \notin X_2 \):
   \[ f_n(X_1 \cup X_2) = X_1 \cup X_2 = f_n(X_1) \cup f_n(X_2) \]

Result of Strongly Live Variables Analysis

<table>
<thead>
<tr>
<th>Node</th>
<th>Iteration #1</th>
<th>Iteration #2</th>
<th>Iteration #3</th>
<th>Iteration #4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( Out_n )</td>
<td>( In_n )</td>
<td>( Out_n )</td>
<td>( In_n )</td>
</tr>
<tr>
<td>( n_6 )</td>
<td>( \emptyset )</td>
<td>( \emptyset )</td>
<td>( \emptyset )</td>
<td>( \emptyset )</td>
</tr>
<tr>
<td>( n_5 )</td>
<td>( \emptyset )</td>
<td>( \emptyset )</td>
<td>{a}</td>
<td>{a}</td>
</tr>
<tr>
<td>( n_4 )</td>
<td>( \emptyset )</td>
<td>( \emptyset )</td>
<td>{a}</td>
<td>{a}</td>
</tr>
<tr>
<td>( n_3 )</td>
<td>( \emptyset )</td>
<td>( \emptyset )</td>
<td>{a}</td>
<td>{b}</td>
</tr>
<tr>
<td>( n_2 )</td>
<td>{a}</td>
<td>{a, b}</td>
<td>{a, b}</td>
<td>{a, b, c}</td>
</tr>
<tr>
<td>( n_1 )</td>
<td>{a}</td>
<td>{a, b}</td>
<td>{b}</td>
<td>{a, b, c}</td>
</tr>
</tbody>
</table>
Tutorial Problem: Strongly May-Must Liveness Analysis?

- Instead of viewing liveness information as
  - a map $\forall \text{Var} \rightarrow \{0, 1\}$ with the lattice $\{0, 1\}$,
  view it as
  - a map $\forall \text{Var} \rightarrow \hat{L}$ where $\hat{L}$ is the May-Must Lattice
- Write the data flow equations
- Prove that the flow functions are distributive

Part 5

Pointer Analyses

An Outline of Pointer Analysis Coverage

- The larger perspective
- Comparing Points-to and Alias information
- Flow Insensitive Points-to Analysis
- Flow Sensitive Points-to Analysis
- Pointer Analyses: An Engineer’s Landscape
- Liveness Based Points-to Analysis
- Generalizations to Heap, Arrays, Pointer Arithmetic, and Unions

Code Optimization In Presence of Pointers

Program

1. $q = p$;
2. while (...) { do {
3. $q = q \rightarrow \text{next}$;
4. }while (...) \}
5. $p \rightarrow \text{data} = r1$;
6. print ($q \rightarrow \text{data}$);
7. $p \rightarrow \text{data} = r2$;

Memory graph at statement 5

- Is $p \rightarrow \text{data}$ live at the exit of line 5? Can we delete line 5?
- We cannot delete line 5 if $p$ and $q$ can be possibly aliased (while loop or do-while loop with a circular list)
- We can delete line 5 if $p$ and $q$ are definitely not aliased (do-while loop without a circular list)
### Code Optimization In Presence of Pointers

\[
\begin{align*}
  a &= 5 \\
  x &= \&a \\
  b &= \ast x \\
  a &= 5 \\
  x &= \&a \\
  b &= \ast x \\
  a &= 5 \\
  x &= \&a \\
  b &= 5
\end{align*}
\]

- Original Program
- Constant Propagation without aliasing
- Constant Propagation with aliasing

### The World of Pointer Analysis

![Diagram of Pointer Analysis](image)

- Alias Analysis
- Pointer Analysis
- Alias analysis of reference parameters, fields of unions, array indices
- Alias analysis of data pointers
- Points-to analysis of data and function pointers

### Pointer Analysis Musings

- Pointer analysis collects information about indirect accesses in programs
  - Enables precise data analysis
  - Enable precise interprocedural control flow analysis
- Needs to scale to large programs
- Pointer Analysis Musings
  - Which Pointer Analysis should I Use?
    - Michael Hind and Anthony Pioli. ISTAA 2000
  - Pointer Analysis: Haven’t we solved this problem yet?
    - Michael Hind PASTE 2001
  - 2017 😞

### The Mathematics of Pointer Analysis

In the most general situation

- Alias analysis is undecidable.
- Flow insensitive alias analysis is NP-hard
  - Horwitz [TOPLAS 1997]
- Points-to analysis is undecidable
  - Chakravarty [POPL 2003]

Adjust your expectations suitably to avoid disappointments!
The Engineering of Pointer Analysis

So what should we expect? To quote Hind [PASTE 2001]

- “Fortunately many approximations exist”
- “Unfortunately too many approximations exist!”

*Engineering of pointer analysis is much more dominant than its science*

--

Pointer Analysis: Engineering or Science?

- Engineering view.
  - Build quick approximations
  - The tyranny of (exclusive) OR!
    - Precision OR Efficiency?
- Science view.
  - Build clean abstractions
  - Can we harness the Genius of AND?
    - Precision AND Efficiency?

- A distinction between approximation and abstraction is subjective
  
  Our working definition
    - Abstractions focus on precision and conciseness of modelling
    - Approximations focus on efficiency and scalability

--

An Outline of Pointer Analysis Coverage

- The larger perspective
- Comparing Points-to and Alias information
- Flow Insensitive Points-to Analysis
- Flow Sensitive Points-to Analysis
- Pointer Analyses: An Engineer’s Landscape
- Liveness Based Points-to Analysis
- Generalizations to Heap, Arrays, Pointer Arithmetic, and Unions

--

Alias Information Vs. Points-to Information

1. $x = &a$
   - “$x$ Points-to $a$” denoted $x \rightarrow a$
   - Neither Symmetric Nor Reflexive

2. $b = x$
   - “$x$ and $b$ are Aliases” denoted $x \triangleq b$
   - Symmetric and Reflexive

- What about transitivity?
  - Points-to: No.
  - Alias: Depends.
Comparing Points-to and Alias Relations (1)

<table>
<thead>
<tr>
<th>Statement</th>
<th>Memory</th>
<th>Points-to</th>
<th>Aliases</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = &amp; y$</td>
<td>Before (assume)</td>
<td>$x \leftarrow y$</td>
<td>Existing</td>
</tr>
<tr>
<td></td>
<td>After</td>
<td>$x \leftrightarrow y$</td>
<td>New</td>
</tr>
<tr>
<td>$x = y$</td>
<td>Before (assume)</td>
<td>$x \leftarrow z$</td>
<td>Existing</td>
</tr>
<tr>
<td></td>
<td>After</td>
<td>$x \leftrightarrow y \leftarrow z$</td>
<td>New</td>
</tr>
</tbody>
</table>

- Indirect aliases. Substitute a name by its aliases for transitivity
- Derived aliases. Apply indirection operator to aliases (ignored here)

$\; x \equiv y \Rightarrow \star x \equiv \star y$

Comparing Points-to and Alias Relations (2)

<table>
<thead>
<tr>
<th>Statement</th>
<th>Memory</th>
<th>Points-to</th>
<th>Aliases</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\star x = y$</td>
<td>Before (assume)</td>
<td>$\star x \leftarrow \star y \leftarrow u$</td>
<td>Existing</td>
</tr>
<tr>
<td></td>
<td>After</td>
<td>$\star x \leftrightarrow \star y \leftarrow u$</td>
<td>New Indirect</td>
</tr>
<tr>
<td>$x = \star y$</td>
<td>Before (assume)</td>
<td>$x \leftarrow z \rightarrow u$</td>
<td>Existing</td>
</tr>
<tr>
<td></td>
<td>After</td>
<td>$x \leftrightarrow y \leftarrow z \rightarrow u$</td>
<td>New</td>
</tr>
</tbody>
</table>

The resulting memories look similar but are different. In the first case we have $u \rightarrow z$ whereas in the second case the arrow direction is opposite (i.e. $z \rightarrow u$).

Comparing Points-to and Alias Relations (3)

- Points-to information records edges in the memory graph
  - aliases of the kind $x \equiv \& y$
  - $x$ holds the address of $y$
  - other aliases can be discovered by composing edges
  - since addresses are explicated, it can represent only those memory locations that can be named at compile time

  More compact but less general

- Alias information records paths in the memory graph
  - paths incident on the same node
  - $x$ and $y$ hold the same address (and the address is left implicit)
  - since addresses are implicit, it can represent unnamed memory locations too
  - if we have $x \equiv y$ then $\star x \equiv \star y$ is redundant and is not recorded

  More general and more complex

An Outline of Pointer Analysis Coverage

- The larger perspective
- Comparing Points-to and Alias information
- Flow Insensitive Points-to Analysis
- Flow Sensitive Points-to Analysis
- Pointer Analyses: An Engineer’s Landscape
- Liveness Based Points-to Analysis
- Generalizations to Heap, Arrays, Pointer Arithmetic, and Unions
Flow Sensitive Vs. Flow Insensitive Pointer Analysis

- Flow insensitive pointer analysis
  - Inclusion based: Andersen’s approach
  - Equality based: Steensgaard’s approach
- Flow sensitive pointer analysis
  - May points-to analysis
  - Must points-to analysis

Flow Insensitivity in Data Flow Analysis

- Assumption: Statements can be executed in any order.
- Instead of computing point-specific data flow information, summary data flow information is computed. The summary information is required to be a safe approximation of point-specific information for each point.
- Kill\(_n\)(X) component is ignored.
  If statement \(n\) kills data flow information, there is an alternate path that excludes \(n\).

The control flow graph is a complete graph (except for the Start and End nodes)

Examples of Flow Insensitive Analyses

- Type checking/inferencing
  (What about interpreted languages?)
- Address taken analysis
  Which variables have their addresses taken?
- Side effects analysis
  Does a procedure modify a global variable? Reference Parameter?

Function composition is replaced by function confluence
Flow Insensitivity in Data Flow Analysis

Assuming Gen_n(X) has dependent parts and Kill_n(X) is ignored

\[ \text{Start} \rightarrow f_0 \rightarrow f_1 \rightarrow f_2 \rightarrow f_3 \rightarrow \cdots \rightarrow f_i \rightarrow \cdots \rightarrow f_m \rightarrow \text{End} \]

Allows arbitrary compositions of flow functions in any order
\[ \Rightarrow \text{Flow insensitivity} \]

In practice, dependent constraints are collected in a global repository in one pass and then are solved independently

Examples of dependent parts in Gen
- Pointer analysis for statements
  - \( x = y \), \( x = \ast y \), \( \ast x = y \)

Notation for Andersen’s and Steensgaard’s Points-to Analysis

- \( P_x \) denotes the set of pointees of pointer variable \( x \)
- \( \text{Unify}(x, y) \) unifies locations \( x \) and \( y \)
  - \( x \) and \( y \) are treated as equivalent locations
  - the pointees of the unified locations are also unified transitively
- \( \text{UnifyPTS}(x, y) \) unifies the pointees of \( x \) and \( y \)
  - \( x \) and \( y \) themselves are not unified

<table>
<thead>
<tr>
<th>Statement</th>
<th>Andersen’s Points-to Sets</th>
<th>Steensgaard’s Points-to Sets</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = &amp;y )</td>
<td>( P_x \supset { y } )</td>
<td>( P_x \supset { y } )</td>
</tr>
<tr>
<td>( x = y )</td>
<td>( P_x \supset P_y )</td>
<td>( \text{UnifyPTS}(x, y) )</td>
</tr>
<tr>
<td>( x = \ast y )</td>
<td>( P_x \supset P_y ) ( \forall z \in P_y )</td>
<td>( \forall z \in P_y ), ( \text{UnifyPTS}(x, z) )</td>
</tr>
<tr>
<td>( \ast x = y )</td>
<td>( P_x \supset P_y ) ( \forall z \in P_x )</td>
<td>( \forall z \in P_x ), ( \text{UnifyPTS}(y, z) )</td>
</tr>
</tbody>
</table>

Andersen’s view
- \( x \) points to \( y \)
- Include \( y \) in the points-to set of \( x \)

Steensgaard’s view
- Equivalence between: All pointees of \( x \)
- Unify \( y \) and pointees of \( x \)
Andersen’s and Steensgaard’s Points-to Analysis

Statement | Andersen’s Points-to Sets | Steensgaard’s Points-to Sets
---|---|---
\(x = \&y\) | \(P_x \supseteq \{y\}\) | \(P_x \supseteq \{y\}\) Unify(\(y, z\)) for some \(z \in P_x\)
\(x = y\) | \(P_x \supseteq P_y\) | UnifyPTS(\(x, y\))
\(x = \*y\) | \(P_x \supseteq P_y, \forall z \in P_y\) | \(\forall z \in P_y, \text{UnifyPTS}(x, z)\)
\(*x = y\) | \(P_x \supseteq P_y, \forall z \in P_x\) | \(\forall z \in P_x, \text{UnifyPTS}(y, z)\)

Andersen’s view
- \(x\) points to pointees of \(y\)
- Include the pointees of \(y\) in the points-to set of \(x\)

Steensgaard’s view
- Equivalence between: Pointees of \(x\) and pointees of \(y\)
- Unify points-to sets of \(x\) and \(y\)

Inclusion | Equality
Inclusion Based (aka Andersen’s) Points-to Analysis: Example 1

Program
1. a = &b
2. c = a
3. a = &d
4. a = &e
5. b = a

Node
1. a
2. c
3. a
4. a
5. b

Constraint
1. \( P_a \supseteq \{ b \} \)
2. \( P_c \supseteq P_a \)
3. \( P_a \supseteq \{ d \} \)
4. \( P_a \supseteq \{ e \} \)
5. \( P_b \supseteq P_a \)

Points-to Graph

- Since \( P_a \) has changed, \( P_c \) needs to be processed again

Equality Based (aka Steensgaard’s) Points-to Analysis: Example 1

Program
1. a = &b
2. c = a
3. a = &d
4. a = &e
5. b = a

Node
1. a
2. c
3. a
4. a
5. b

Constraint
1. \( P_a \supseteq \{ b \} \)
2. \( P_c \supseteq P_a \)
3. \( P_a \supseteq \{ d \} \)
4. \( P_a \supseteq \{ e \} \)
5. \( P_b \supseteq P_a \)

Points-to Graph

- Observe that \( P_c \) is processed for the third time
- Order of processing the sets influences efficiency significantly
- A plethora of heuristics have been proposed

Actually:
- c does not point to any location in block 1
- a does not point b in block 5
  (the method ignores the kill due to 3 and 4)
- b does not point to itself at any time
Equality Based (aka Steensgaard’s) Points-to Analysis: Example 1

Program

1. \( a = \& b \)
2. \( c = a \)
3. \( a = \& d \)
4. \( a = \& e \)
5. \( b = a \)

Node | Constraint
--- | ---
1 | \( P_a \supseteq \{ b \} \)
2 | \( \text{Unify}(x, d), x \in P_a \)
3 | \( P_a \supseteq \{ d \} \)
4 | \( \text{Unify}(x, e), x \in P_a \)
5 | \( \text{UnifyPTS}(b, a) \)

The full blown up points-to-graph has far more edges than in the graph created by Andersen’s method
Far more efficient but far less precise

Comparing Equality and Inclusion Based Analyses (2)

- Andersen’s algorithm is cubic in number of pointers
- Steensgaard’s algorithm is nearly linear in number of pointers
  - How can it be more efficient by an orders of magnitude?

Efficiency of Equality Based Approach

<table>
<thead>
<tr>
<th>Program</th>
<th>Andersen’s approach</th>
<th>Steensgaard’s approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>( a = &amp; b )</td>
<td>( a \rightarrow b \rightarrow d )</td>
<td>( a \rightarrow b \rightarrow d )</td>
</tr>
<tr>
<td>( a = &amp; c )</td>
<td>( a \rightarrow c \rightarrow d )</td>
<td>( a \rightarrow c \rightarrow d )</td>
</tr>
<tr>
<td>( b = &amp; d )</td>
<td>( b \rightarrow d )</td>
<td>( b \rightarrow d )</td>
</tr>
<tr>
<td>( b = &amp; c )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Andersen’s inclusion based wisdom:
  - Add edges and let the number of successors increase
- Steensgaard’s equality based wisdom:
  - Merge multiple successors and maintain a single successor of any node
  - Since a larger number of pointers treated are alike and fewer distinctions are maintained, we get much smaller points-to graphs
  - Efficient \( \text{Union-Find} \) algorithms to merge intersecting subsets

Inclusion Based (aka Andersen’s) Points-to Analysis: Example 2

Constraints on Points-to Sets

\[
\begin{align*}
P_x & \supseteq \{ y \} \\
P_y & \supseteq \{ z \} \\
P_z & \supseteq \{ u \} \\
\forall w \in P_z, P_w & \supseteq P_y \\
P_z & \supseteq P_y \\
P_y & \supseteq \{ x \}
\end{align*}
\]

Points-to Graph

\[
\begin{align*}
\text{Constraints} & \\
& x = \& y \\
y & = \& z \\
z & = \& u \\
& \text{use } u \\
& \text{use } x
\end{align*}
\]
Equality Based (aka Steensgaard’s) Points-to Analysis:
Example 2

- Treat all pointees of a pointer as “equivalent” locations
- Transitive closure
  Pointees of all equivalent locations become equivalent

Steensgaard’s Points-to Graph

Tutorial Problem for Flow Insensitive Pointer Analysis (1)

Program
p = &q
r = &s
t = &p
u = p
\[ \ast t = r \]

Inclusion based

Equality based

Tutorial Problem for Flow Insensitive Pointer Analysis (2)

Compute flow insensitive points-to information using inclusion based method as well as equality based method

```
if (...) 
  p = &x;
else
  p = &y;
  x = &a;
  y = &b;
  *p = &c;
  *y = &a;
```

Tutorial Problem for Flow Insensitive Pointer Analysis (3)

Compute flow insensitive points-to information using inclusion based method as well as equality based method

```
if (...) 
  p = &x;
else
  p = &y;
  x = &a;
  y = &b;
  *p = &c;
  *y = &a;
```
An Outline of Pointer Analysis Coverage

- The larger perspective
- Comparing Points-to and Alias information
- Flow Insensitive Points-to Analysis
- Flow Sensitive Points-to Analysis
- Pointer Analyses: An Engineer’s Landscape
- Liveness Based Points-to Analysis
- Generalizations to Heap, Arrays, Pointer Arithmetic, and Unions

Must Points-to Information

1. $x = \& a$
2. $x = \& b$
3. $x = \& b$
4. $x = \& b$

May Points-to Information

1. $x = \& a$
2. $x = \& b$
3. $x = \& b$
4. $x = \& b$

Must Alias Information

1. $x = \& a$
2. $b = x$
3. $y = b$
4. $x \equiv b$ and $b \equiv y \Rightarrow x \equiv y$
May Alias Information

\[ x = \& a \]
\[ b = \& z \]
\[ b = x \]
\[ y = b \]
\[ z = \& x \]
\[ z = \& y \]

\[ x \equiv b \text{ and } b \not\equiv y \not\Rightarrow x \equiv y \]

Strong and Weak Updates

\[ x = \& a \]
\[ y = \& b \]
\[ w = \& c \]
\[ z = \& x \]
\[ z = \& y \]

Weak update: Modification of \( x \) or \( y \) due to \( *z \) in block 5
Strong update: Modification of \( c \) due to \( *w \) in block 5

How is this concept related to May/Must nature of information?

What About Heap Data?

- Compile time entities, abstract entities, or summarized entities
- Three options:
  - Represent all heap locations by a single abstract heap location
  - Represent all heap locations of a particular type by a single abstract heap location
  - Represent all heap locations allocated at a given memory allocation site by a single abstract heap location
- Summarization: Usually based on the length of pointer expression
- Initially, we will restrict ourselves to stack and static data
  We will later introduce heap using the allocation site based abstraction

Lattice for May Points-to Analysis

Let \( P \subseteq \text{Var} \) be the set of pointers. Assume \( \text{Var} = \{p, q\} \) and \( P = \{p\} \)

<table>
<thead>
<tr>
<th>Product View</th>
<th>Mapping view</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \emptyset )</td>
<td>( {(p, \emptyset)} )</td>
</tr>
<tr>
<td>( {(p, p)} )</td>
<td>( {(p, {p})} )</td>
</tr>
<tr>
<td>( {(p, p), (p, q)} )</td>
<td>( {(p, {p, q})} )</td>
</tr>
</tbody>
</table>

Data flow values \( \subseteq P \times \text{Var} \)
Lattice = \( (2^{P \times \text{Var}}, \supseteq) \)
Points-to graph as a list of directed edges

Data flow values \( \in P \rightarrow 2^\text{Var} \)
Lattice = \( (P \rightarrow 2^\text{Var}, \subseteq_{\text{map}}) \)
Points-to graph as a list of adjacency lists
Lattice for Must Points-to Analysis

Let $P \subseteq \mathbb{Var}$ be the set of pointers. Assume $\mathbb{Var} = \{p, q, r\}$ and $P = \{p\}$

Mapping View

<table>
<thead>
<tr>
<th>Component Lattice</th>
<th>Set View</th>
</tr>
</thead>
<tbody>
<tr>
<td>${(p, \top)}$</td>
<td>${(p, p), (p, q), (p, r)}$</td>
</tr>
<tr>
<td>${(p, p)}$</td>
<td>${(p, p)}$</td>
</tr>
<tr>
<td>${(p, q)}$</td>
<td>${(p, q)}$</td>
</tr>
<tr>
<td>${(p, r)}$</td>
<td>${(p, r)}$</td>
</tr>
<tr>
<td>${(p, \bot)}$</td>
<td>$\emptyset$</td>
</tr>
</tbody>
</table>

Data flow values $= P \rightarrow \mathbb{Var} \cup \{\top, \bot\}$

Lattice $= (2^{P \rightarrow \mathbb{Var} \cup \{\top, \bot\}}, \subseteq_{map})$

Restricted subset of $P \times \mathbb{Var}$

$\cap$ can be used for $\cap$

A pointer can point to at most one location

Lattice for Combined May-Must Points-to Analysis (1)

- Consider the following abbreviation of the May-Must lattice $\hat{L}$
  
  $$\begin{array}{c}
  \text{Unknown} \\
  \text{No} \quad \text{Must} \\
  \text{May}
  \end{array}$$

  abbreviated as

  $$\begin{array}{c}
  \text{un} \\
  \text{no} \quad \text{mt} \\
  \text{my}
  \end{array}$$

- For $\mathbb{Var} = \{p, q\}, P = \{p\}$, the May-Must points-to lattice is the product $P \times \mathbb{Var} \times \hat{L}$

  - Some elements are prohibited because of the semantics of Must
  - If we have $(p, p, mt)$ in a data flow value $X \in P \times \mathbb{Var} \times \hat{L}$, then
    - we cannot have $(p, q, un)$, $(p, q, mt)$, or $(p, q, my)$ in $X$
    - we can only have $(p, q, no)$ in $X$

Lattice for Combined May-Must Points-to Analysis (2)

For $\mathbb{Var} = \{p, q\}, P = \{p\}$, the May-Must points-to lattice is
Lattice for Combined May-Must Points-to Analysis (2)

For $\forall a = \{p, q\}$, $P = \{p\}$, the May-Must points-to lattice is:

- $\{(p,p,un), (p,q,un)\}$
- $\{(p,p,no), (p,q,no)\}$
- $\{(p,p,mt), (p,q,mt)\}$
- $\{(p,p,my), (p,q,my)\}$
- $\{(p,p,my), (p,q,my)\}$

May and Must Analysis for Killing Points-to Information (1)

May Points-to Analysis

- $(a, b)$ should be in $MayIn_a$
- Holds along path 1-3-4
- Block 4 should not kill $(a, b)$
- Possible if pointee set of $c$ is $\emptyset$ (Use $MustIn_a$)
- However, $MayIn_a$ contains $(c, a)$

Must Points-to Analysis

- $(a, b)$ should not be in $MustIn_a$
- Does not hold along path 1-2-4
- Block 4 should kill $(a, b)$
- Possible if pointee set of $c$ is $\{a\}$ (Use $MayIn_a$)
- However, $MustIn_a$ contains $(a, b)$

For killing points-to information through indirection,

- **Must** points-to analysis should identify pointees of $c$ using $MayIn_a$
- **May** points-to analysis should identify pointees of $c$ using $MustIn_a$

Lattice for Combined May-Must Points-to Analysis (2)

For $\forall a = \{p, q\}$, $P = \{p\}$, the May-Must points-to lattice is:

- $\{(p,p,un), (p,q,un)\}$
- $\{(p,p,no), (p,q,no)\}$
- $\{(p,p,mt), (p,q,mt)\}$
- $\{(p,p,my), (p,q,my)\}$
- $\{(p,p,my), (p,q,my)\}$

May and Must Analysis for Killing Points-to Information (2)

- May Points-to analysis should remove a May points-to pair
  - Only if it must be removed along all paths
- Kill should remove only strong updates
  - Should use Must Points-to information
- Must Points-to analysis should remove a Must points-to pair
  - If it can be removed along any path
- Kill should remove all weak updates
  - Should use May Points-to information
Discovering Must Points-to Information from May Points-to Information

Let $P \subseteq \text{Var}$ be the set of pointer variables

May-points-to information: $A = \langle 2^P \times \text{Var}, \supseteq \rangle$

Standard algebraic operations on points-to relations

Given relation $R \subseteq P \times \text{Var}$ and $X \subseteq P$,

- Relation application $R[X] = \{v \mid u \in X \wedge (u, v) \in R\}$
  (Find out the pointees of the pointers contained in $X$)

- Relation restriction $(R \mid_X) = \{(u, v) \in R \mid u \in X\}$
  (Restrict the relation only to the pointers contained in $X$ by removing points-to information of other pointers)

Points-to Analysis Data Flow Equations

Let

$\text{Var} = \{a, b, c, d, e, f, g, ?\}$
$P = \{a, b, c, d, e\}$
$R = \{(a, b), (a, c), (b, d), (c, e), (c, g), (d, a), (e, ?)\}$
$X = \{a, c\}$

Then,

$R[X] = \{v \mid u \in X \wedge (u, v) \in R\}$
  $= \{b, c, e, g\}$

$R \mid_X = \{(u, v) \in R \mid u \in X\}$
  $= \{(a, b), (a, c), (c, e), (c, g)\}$
**Points-to Analysis Data Flow Equations**

\[ \text{Ain}_n = \begin{cases} \forall \text{Var} \times \{?\} & n \text{ is Start}_p \\ \bigcup_{p \in \text{pred}(n)} \text{Aout}_p & \text{otherwise} \end{cases} \]

\[ \text{Aout}_n = (\text{Ain}_n - (\text{Kill}_n \times \forall \text{Var})) \cup (\text{Def}_n \times \text{Pointee}_n) \]

- \text{Ain/Aout}: sets of \text{mAy} points-to pairs
- \text{Kill}_n, \text{Def}_n, \text{and Pointee}_n are defined in terms of \text{Ain}_n

Pointees (i.e., locations whose addresses are stored)

Pointers that are defined (i.e., pointers in which addresses are stored)

Pointees whose points-to relations should be removed

**Extractor Functions for Points-to Analysis**

Values defined in terms of \text{Ain}_n (denoted \text{A})

<table>
<thead>
<tr>
<th>Def$_n$</th>
<th>Kill$_n$</th>
<th>Pointee$_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>use $x$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$x = &amp;a$</td>
<td>${x}$</td>
<td>${a}$</td>
</tr>
<tr>
<td>$x = y$</td>
<td>${x}$</td>
<td>${y}$</td>
</tr>
<tr>
<td>$x = *y$</td>
<td>${x}$</td>
<td>$A(y)$</td>
</tr>
<tr>
<td>$*x = y$</td>
<td>$A{x} \cap P$</td>
<td>$\text{Must}(A{x} \cap P)$</td>
</tr>
<tr>
<td>other</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
</tbody>
</table>

Pointees of those pointees of $y$ in \text{Ain}_n which are pointers

Pointees of $y$ in \text{Ain}_n are the targets of defined pointers
Extractor Functions for Points-to Analysis

Values defined in terms of $A_{in}$ (denoted $A$)

<table>
<thead>
<tr>
<th></th>
<th>Def$_n$</th>
<th>Kill$_n$</th>
<th>Pointee$_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>use $x$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$x = &amp;a$</td>
<td>${x}$</td>
<td>${x}$</td>
<td>${a}$</td>
</tr>
<tr>
<td>$x = y$</td>
<td>${x}$</td>
<td>${x}$</td>
<td>$A(y)$</td>
</tr>
<tr>
<td>$x = *y$</td>
<td>${x}$</td>
<td>${x}$</td>
<td>$A(A(y) \cap P)$</td>
</tr>
<tr>
<td>$*x = y$</td>
<td>$A{x} \cap P$</td>
<td>Must($A{x} \cap P$)</td>
<td>$A(y)$</td>
</tr>
<tr>
<td>other</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
</tbody>
</table>

Pointees of $x$ in $A_{in}$ receive new addresses

Extractor Functions for Points-to Analysis

Values defined in terms of $A_{in}$

<table>
<thead>
<tr>
<th></th>
<th>Def$_n$</th>
<th>Kill$_n$</th>
<th>Pointee$_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>use $x$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$x = &amp;a$</td>
<td>${x}$</td>
<td>${x}$</td>
<td>${a}$</td>
</tr>
<tr>
<td>$x = y$</td>
<td>${x}$</td>
<td>${x}$</td>
<td>$A(y)$</td>
</tr>
<tr>
<td>$x = *y$</td>
<td>${x}$</td>
<td>${x}$</td>
<td>$A(A(y) \cap P)$</td>
</tr>
<tr>
<td>$*x = y$</td>
<td>$A{x} \cap P$</td>
<td>Must($A{x} \cap P$)</td>
<td>$A(y)$</td>
</tr>
<tr>
<td>other</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
</tbody>
</table>

Strong update using must-points-to information computed from $A_{in}$

$\text{Must}(R) = \bigcup_{z \in P} \times \begin{cases} \{w\} & R\{z\} = \{w\} \land w \neq \emptyset \\ \emptyset & \text{otherwise} \end{cases}$

$z$ has a single pointee $w$ in must-points-to relation

$\text{Must}(R) = \bigcup_{z \in P} \times \begin{cases} \{w\} & R\{z\} = \{w\} \land w \neq \emptyset \\ \emptyset & \text{otherwise} \end{cases}$

$z$ has no pointee in must-points-to relation
Extractor Functions for Points-to Analysis

Values defined in terms of $A_{in}$ (denoted $A$)

<table>
<thead>
<tr>
<th>$def_{in}$</th>
<th>$kill_{in}$</th>
<th>$pointee_{in}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>use $x$</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
<tr>
<td>$x = &amp;a$</td>
<td>${x}$</td>
<td>${a}$</td>
</tr>
<tr>
<td>$x = y$</td>
<td>${x}$</td>
<td>$A(y)$</td>
</tr>
<tr>
<td>$x = *y$</td>
<td>${x}$</td>
<td>$A(y) \cap P$</td>
</tr>
<tr>
<td>$*x = y$</td>
<td>$A{x} \cap P$</td>
<td>$M must(A) {x} \cap P$</td>
</tr>
<tr>
<td>other</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
</tbody>
</table>

$$Must(R) = \bigcup_{p} \{z\} \times \begin{cases} \{w\} & R(z) = \{w\} \land w \neq ? \\ \emptyset & \text{otherwise} \end{cases}$$

Pointees of $y$ in $A_{in}$ are the targets of defined pointers.

An Example of Flow Sensitive May Points-to Analysis

Assume that the program is type correct.

Weak Update $n_4$ $\ast u = \& x$

Strong Update $n_5$ $\ast y = \& y$

Compute May and Must points-to information

```
if ( . . . )
    p = &x;
else
    p = &y;
    x = &a;
    y = &b;
    *p = &c;
    *y = &a;
```

May Points-to

```
May Points-to
n1
n2 x = &z
n3 y = &w
n4 *x = y
```

Must Points-to

```
Must Points-to
n1
n2 b = &c
c = &d
n3 b = &e
e = &d
n4 a = *b
```

$z \rightarrow w$ is spurious

$a \rightarrow d$ is missing
An Outline of Pointer Analysis Coverage

- The larger perspective
- Comparing Points-to and Alias information
- Flow Insensitive Points-to Analysis
- Flow Sensitive Points-to Analysis
- Pointer Analyses: An Engineer’s Landscape
- Liveness Based Points-to Analysis
- Generalizations to Heap, Arrays, Pointer Arithmetic, and Unions

An Example of Flow Insensitive May Points-to Analysis

For simplicity, we ignore the BI with “?”

An Example of Flow Sensitive May Points-to Analysis

Context Sensitivity in Interprocedural Analysis
We will revisit this concept and study it in details in the fourth module (interprocedural data flow analysis) of the course.
**An Outline of Pointer Analysis Coverage**

- The larger perspective
- Comparing Points-to and Alias information
- Flow Insensitive Points-to Analysis
- Flow Sensitive Points-to Analysis
- Pointer Analyses: An Engineer’s Landscape
- Liveness Based Points-to Analysis
- Generalizations to Heap, Arrays, Pointer Arithmetic, and Unions

**Our Motivating Example for FCPA**

For simplicity, we ignore the BI with “?”

```
\( n_1 \)
```

```
\( x = \& y \\
y = \& z \\
z = \& u \\
x \Rightarrow y \Rightarrow z \Rightarrow u \\
x \Rightarrow y \Rightarrow z \Rightarrow u \\
x \Rightarrow y \Rightarrow z \Rightarrow u \\
x \Rightarrow y \Rightarrow z \Rightarrow u \\
```

```
\( n_2 \)
```

```
\( *z = y \\
z = y \\
x \Rightarrow y \Rightarrow z \Rightarrow u \\
x \Rightarrow y \Rightarrow z \Rightarrow u \\
```

```
\( n_3 \)
```

```
\( y = \& x \\
use \ u \\
use \ x \\
x \Rightarrow y \Rightarrow z \Rightarrow u \\
x \Rightarrow y \Rightarrow z \Rightarrow u \\
```

```
\( n_4 \)
```

```
\( x \Rightarrow y \Rightarrow z \Rightarrow u \\
```

Is All This Information Useful?

For simplicity, we ignore the BI with "?"

\[
\begin{align*}
  x &= &y \\
  y &= &z \\
  z &= &u
\end{align*}
\]

\[
\begin{array}{c}
  n_1 \\
  n_2 \\
  n_3 \\
  n_4
\end{array}
\]

\[
\begin{align*}
  n_1 &: \\
  n_2 &: z = y \\
  n_3 &: z = y \\
  n_4 &: y = &x \\
      &use &u \\
      &use &x
\end{align*}
\]

The L and P of LFCPA

Mutual dependence of liveness and points-to information
- Define points-to information only for live pointers
- For pointer indirections, define liveness information using points-to information

The F and C of LFCPA

- Use call strings method for full flow and context sensitivity
- Use value contexts for efficient interprocedural analysis
- \[Khedker-Karkare-CC-2008, Padhye-Khedker-SOAP-2013\]

Use of Strong Liveness

- Simple liveness considers every use of a variable as useful
- Strong liveness checks the liveness of the result before declaring the operands to be live
- Strong liveness is more precise than simple liveness
### Extractor Functions for LFCPA

<table>
<thead>
<tr>
<th>Def&lt;sub&gt;n&lt;/sub&gt;</th>
<th>Kill&lt;sub&gt;n&lt;/sub&gt;</th>
<th>Pointee&lt;sub&gt;n&lt;/sub&gt;</th>
<th>Ref&lt;sub&gt;n&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>use x</td>
<td>∅</td>
<td>∅</td>
<td>∅</td>
</tr>
<tr>
<td>x = &amp;a</td>
<td>{x}</td>
<td>{x}</td>
<td>{x}</td>
</tr>
<tr>
<td>x = y</td>
<td>{x}</td>
<td>{x}</td>
<td>{a}</td>
</tr>
<tr>
<td>x = *y</td>
<td>{x}</td>
<td>{x}</td>
<td>A(y)</td>
</tr>
<tr>
<td>*x = y</td>
<td>A(x) ∩ P</td>
<td>Must(A(x) ∩ P)</td>
<td>A(y)</td>
</tr>
<tr>
<td>other</td>
<td>0</td>
<td>0</td>
<td>{x}</td>
</tr>
</tbody>
</table>

- Lin/Lout: set of Live pointers, Ain/Aout: sets of mAy points-to pairs
- Ref<sub>n</sub>, Kill<sub>n</sub>, Def<sub>n</sub>, and Pointee<sub>n</sub> are defined in terms of Ain<sub>n</sub>

---

### Extractor Functions for LFCPA

<table>
<thead>
<tr>
<th>Def&lt;sub&gt;n&lt;/sub&gt;</th>
<th>Kill&lt;sub&gt;n&lt;/sub&gt;</th>
<th>Pointee&lt;sub&gt;n&lt;/sub&gt;</th>
<th>Ref&lt;sub&gt;n&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>use x</td>
<td>0</td>
<td>0</td>
<td>∅</td>
</tr>
<tr>
<td>x = &amp;a</td>
<td>{x}</td>
<td>{x}</td>
<td>{a}</td>
</tr>
<tr>
<td>x = y</td>
<td>{x}</td>
<td>{x}</td>
<td>A(y)</td>
</tr>
<tr>
<td>x = *y</td>
<td>{x}</td>
<td>{x}</td>
<td>A(A(y) ∩ P)</td>
</tr>
<tr>
<td>*x = y</td>
<td>A(x) ∩ P</td>
<td>Must(A(x) ∩ P)</td>
<td>A(y)</td>
</tr>
<tr>
<td>other</td>
<td>0</td>
<td>0</td>
<td>{x}</td>
</tr>
</tbody>
</table>

- Pointers that become live
- Defined pointers must be live at the exit for the read pointers to become live
- Some pointers are unconditionally live

---

### Extractor Functions for LFCPA

<table>
<thead>
<tr>
<th>Def&lt;sub&gt;n&lt;/sub&gt;</th>
<th>Kill&lt;sub&gt;n&lt;/sub&gt;</th>
<th>Pointee&lt;sub&gt;n&lt;/sub&gt;</th>
<th>Ref&lt;sub&gt;n&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>use x</td>
<td>0</td>
<td>0</td>
<td>∅</td>
</tr>
<tr>
<td>x = &amp;a</td>
<td>{x}</td>
<td>{x}</td>
<td>{a}</td>
</tr>
<tr>
<td>x = y</td>
<td>{x}</td>
<td>{x}</td>
<td>A(y)</td>
</tr>
<tr>
<td>x = *y</td>
<td>{x}</td>
<td>{x}</td>
<td>A(A(y) ∩ P)</td>
</tr>
<tr>
<td>*x = y</td>
<td>A(x) ∩ P</td>
<td>Must(A(x) ∩ P)</td>
<td>A(y)</td>
</tr>
<tr>
<td>other</td>
<td>0</td>
<td>0</td>
<td>{x}</td>
</tr>
</tbody>
</table>

- x is unconditionally live
- y is live if defined pointers are live
### Extractor Functions for LFCPA

<table>
<thead>
<tr>
<th>Def(_n)</th>
<th>Kill(_n)</th>
<th>Pointee(_n)</th>
<th>Ref(_n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{use } x)</td>
<td>(\emptyset)</td>
<td>(\emptyset)</td>
<td>({x})</td>
</tr>
<tr>
<td>(x = &amp;a)</td>
<td>({x})</td>
<td>({a})</td>
<td>(\emptyset)</td>
</tr>
<tr>
<td>(x = y)</td>
<td>({x})</td>
<td>(A(y))</td>
<td>({y})</td>
</tr>
<tr>
<td>(x = *y)</td>
<td>({x})</td>
<td>(A(A(y) \cap P))</td>
<td>(\emptyset)</td>
</tr>
<tr>
<td>(*x = y)</td>
<td>(A(x) \cap P)</td>
<td>(\text{Must}(A){x} \cap P)</td>
<td>(A(y))</td>
</tr>
<tr>
<td>other</td>
<td>(\emptyset)</td>
<td>(\emptyset)</td>
<td>(\emptyset)</td>
</tr>
</tbody>
</table>

- \(y\) and its pointees in \(Ain\) are live if defined pointers are live.

### Deriving Must Points-to for LFCPA

For \(*x = y\), unless the pointees of \(x\) are known:
- points-to propagation should be blocked
- liveness propagation should be blocked

to ensure monotonicity:

\[
\text{Must}(R) = \bigcup_{x \in P} \{x\} \times \begin{cases} \forall \text{Var} \ R\{x\} = \emptyset \lor R\{x\} = \{?\} \\ \{y\} R\{x\} = \{y\} \land y \neq ? \\ \emptyset \text{ otherwise} \end{cases}
\]

### LFCPA Data Flow Equations

\[
Lout_n = \begin{cases} \emptyset & n \text{ is End}\_p \\ \bigcup_{s \in \text{succ}(n)} Lin_s & \text{otherwise} \end{cases}
\]

\[
Lin_n = \left( Lout_n - \text{Kill}\_n \right) \cup \text{Ref}\_n
\]

\[
Ain_n = \begin{cases} \bigcup_{p, c \in \text{pred}(n)} Aout_p \left| \left| Lin_n \right| \right. & n \text{ is Start}\_p \\ \emptyset & \text{otherwise} \end{cases}
\]

\[
Aout_n = \left( \left( Ain_n - \text{Kill}\_n \times \text{Var} \right) \cup \left( \text{Def}\_n \times \text{Pointee}\_n \right) \right) \big| Lout_n
\]

- \(\text{Lin/Lout}\): set of Live pointers
- \(\text{Ain/Aout}\): definitions remain unchanged except for restriction to liveness
Motivating Example Revisited

- For convenience, we show complete sweeps of liveness and points-to analysis repeatedly
- This is not required by the computation
- The data flow equations define a single fixed point computation
First Round of Liveness Analysis and Points-to Analysis

Strong liveness: y is not made live because z is not live.

Second Round of Liveness Analysis and Points-to Analysis

Strong liveness: y is not made live because z is not live.

Liveness Analysis

Points-to Analysis
Second Round of Liveness Analysis and Points-to Analysis

Point-to Analysis

\[ \{u, x\} \quad \{u, x\} \quad \{u\} \quad \{u, x\} \]

Liveness

\[ n_1 \]

\[ x = \& y \]
\[ y = \& z \]
\[ z = \& u \]

Points-to

\[ n_2 \]

\[ *z = y \]
\[ z = y \]

\[ n_3 \]

\[ n_4 \]

\[ n_3 \]

\[ y = \& x \]

use \( u \) use \( x \)

Points-to Analysis

\[ \{u, x, y, z\} \]
\[ \{u, x, y, z\} \]
\[ \{u\} \]
\[ \{u\} \]

LFCPA Implementation

- LTO framework of GCC 4.6.0
- Naive prototype implementation
  (Points-to sets implemented using linked lists)
- Implemented FCPA without liveness for comparison
- Comparison with GCC’s flow and context insensitive method
- SPEC 2006 benchmarks

Analysis Time

<table>
<thead>
<tr>
<th>Program</th>
<th>kLoC</th>
<th>Call Sites</th>
<th>L-FCPA Liveness</th>
<th>FCPA</th>
<th>GPTA</th>
</tr>
</thead>
<tbody>
<tr>
<td>lbm</td>
<td>0.9</td>
<td>33</td>
<td>0.55</td>
<td>0.52</td>
<td>1.9</td>
</tr>
<tr>
<td>mcf</td>
<td>1.6</td>
<td>29</td>
<td>1.04</td>
<td>0.62</td>
<td>9.5</td>
</tr>
<tr>
<td>libquantum</td>
<td>2.6</td>
<td>258</td>
<td>2.0</td>
<td>1.8</td>
<td>5.6</td>
</tr>
<tr>
<td>bzip2</td>
<td>3.7</td>
<td>233</td>
<td>4.5</td>
<td>4.8</td>
<td>28.1</td>
</tr>
<tr>
<td>parser</td>
<td>7.7</td>
<td>1123</td>
<td>1.2 \times 10^4</td>
<td>145.6</td>
<td>4.3 \times 10^3</td>
</tr>
<tr>
<td>sjeng</td>
<td>10.5</td>
<td>678</td>
<td>858.2</td>
<td>99.0</td>
<td>3.2 \times 10^4</td>
</tr>
<tr>
<td>hmmer</td>
<td>20.6</td>
<td>1292</td>
<td>90.0</td>
<td>62.9</td>
<td>2.9 \times 10^3</td>
</tr>
<tr>
<td>h264ref</td>
<td>36.0</td>
<td>1992</td>
<td>2.2 \times 10^9</td>
<td>2.0 \times 10^9</td>
<td>?</td>
</tr>
</tbody>
</table>

Unique Points-to Pairs

<table>
<thead>
<tr>
<th>Program</th>
<th>kLoC</th>
<th>Call Sites</th>
<th>Unique points-to pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>lbm</td>
<td>0.9</td>
<td>33</td>
<td>12 507 1911</td>
</tr>
<tr>
<td>mcf</td>
<td>1.6</td>
<td>29</td>
<td>41 367 2159</td>
</tr>
<tr>
<td>libquantum</td>
<td>2.6</td>
<td>258</td>
<td>49 119 2701</td>
</tr>
<tr>
<td>bzip2</td>
<td>3.7</td>
<td>233</td>
<td>60 210 8.8 \times 10^4</td>
</tr>
<tr>
<td>parser</td>
<td>7.7</td>
<td>1123</td>
<td>531 4196 1.9 \times 10^4</td>
</tr>
<tr>
<td>sjeng</td>
<td>10.5</td>
<td>678</td>
<td>267 818 1.1 \times 10^4</td>
</tr>
<tr>
<td>hmmer</td>
<td>20.6</td>
<td>1292</td>
<td>232 5805 1.9 \times 10^6</td>
</tr>
<tr>
<td>h264ref</td>
<td>36.0</td>
<td>1992</td>
<td>1683 ? 1.6 \times 10^7</td>
</tr>
</tbody>
</table>
Points-to Information is Small and Sparse

LFCPA Observations

- Usable pointer information is very small and sparse
- Data flow propagation in real programs seems to involve only a small subset of all possible data flow values
- Earlier approaches reported inefficiency and non-scalability because they computed far more information than the actual usable information

LFCPA Conclusions

- Building quick approximations and compromising on precision may not be necessary for efficiency
- Building clean abstractions to separate the necessary information from redundant information is much more significant
  Our experience of points-to analysis shows that
  - Use of liveness reduced the pointer information . . .
  - which reduced the number of contexts required . . .
  - which reduced the liveness and pointer information . . .
- Approximations should come after building abstractions rather than before

LFCPA Lessons: The Larger Perspective

- Do not compute what you don’t need!
- Who defines what is needed? Client

- exhaustive computation
- computation restricted to usable information
- incremental computation
- demand driven computation

What should be computed?

Minimum Computation

Maximum Computation

When should it be computed?

Early Computation

Late Computation
LFCPA Lessons: The Larger Perspective

What should be computed?
Maximum Computation
Minimum Computation

When should it be computed?
Early Computation
Late Computation

Do not compute what you don’t need!

Who defines what is needed?
Algorithm, Data Structure

Avoid computing some values because
• they have been computed before, or
• they can just be “adjusted”, or
• they are equivalent to some other values
E.g. Value based termination of call strings, Work list based methods, BDDs

Definition of Analysis

No One!
**LFCPA Lessons: The Larger Perspective**

- exhaustive computation
- computation restricted to usable information
- incremental computation
- demand driven computation

Maximum Computation → What should be computed?

Early Computation → When should it be computed?

Minimum Computation

Late Computation

Do not compute what you don’t need!

Who defines what is needed?

These seem orthogonal and may be used together

**Tutorial Problems for FCPA and LFCPA**

- Perform may points-to analysis by deriving must info using “?” in BI
- Perform liveness based points-to analysis

```
1 b = &a
2 c = b
3 a = &b
4 a = &c
5 a = *a
6 +b = c
7 use c
```

```
1 y = &z
2 z = &w
3 x = &u
4 x = &v
5 t = *y
6 +x = t
7 use u
```

**An Outline of Pointer Analysis Coverage**

- The larger perspective
- Comparing Points-to and Alias information
- Flow Insensitive Points-to Analysis
- Flow Sensitive Points-to Analysis
- Pointer Analyses: An Engineer’s Landscape
- Liveness Based Points-to Analysis
- Generalizations to Heap, Arrays, Pointer Arithmetic, and Unions

**Original LFCPA Formulation**

- Data flow equations
  - Lin/Lout, Ain/Aout
- Extractors for statements
  - Def, Kill, Ref, Pointee
- Lattices
  - $2^P \times \mathbb{V}_{ar}, 2^P$
- Named locations
  - Variables $\mathbb{V}_{ar}$, Pointers $P$
Formulating Generalizations in LFCPA

Data flow equations
\[ \text{Lin}/\text{Lout, Ain}/\text{Aout} \]

Extractors for statements
\[ \text{Def, Kill, Ref, Pointee} \]

Extractors for pointer expressions
\[ \text{lval, rval, deref, ref} \]

Lattices
\[ 2^S \times T, 2^S \]

Named locations

Variables \( \text{Var} \), Pointers \( \text{P} \), Allocation Sites \( \text{H} \), Fields \( F, pF, npF \), Offsets \( C \)

Generalization for Heap and Structures

- Grammar.
  \[ \alpha ::= \text{malloc} | \& \beta | \beta \]
  \[ \beta ::= x | \beta.f | \beta \rightarrow f | \ast \beta \]

  where \( \alpha \) is a pointer expression, \( x \) is a variable, and \( f \) is a field

- Memory model: Named memory locations. No numeric addresses

  \[ S = \text{P} \cup \text{H} \cup S_p \]
  \[ T = \text{Var} \cup \text{H} \cup S_m \cup \{?\} \]
  \[ S_p = R \times npF^* \times pF \]
  \[ S_m = R \times npF^* \times (pF \cup npF) \]

Named Locations for Pointer Expressions

typedef struct B
{
    ...
    struct B *f;
} sB;

typedef struct A
{
    ...
    struct B g;
} sA;

1. \( a = (sA*)\text{malloc} (\text{sizeof}(sA))\);
2. \( y = \&a\rightarrow g; \)
3. \( b.f = y; \)
4. \( x = \&b; \)
5. \( y.f = \&x; \)
6. return \( x\rightarrow f\rightarrow f; \)

L- and R-values of Pointer Expressions

\[ \text{lval}(\alpha, A) = \begin{cases} 
\{\sigma\} & (\alpha \equiv \sigma) \land (\sigma \in \text{Var}) \\
\{\sigma.f \mid \sigma \in \text{lval}(\beta, A)\} & \alpha \equiv \beta.f \\
\{\sigma.f \mid \sigma \in \text{rval}(\beta, A), \sigma \not= ?\} & \alpha \equiv \beta \rightarrow f \\
\{\sigma \mid \sigma \in \text{rval}(\beta, A), \sigma \not= ?\} & \alpha \equiv \ast \beta \\
\emptyset & \text{otherwise}
\end{cases} \]

\[ \text{rval}(\alpha, A) = \begin{cases} 
\text{lval}(\beta, A) & \alpha \equiv \& \beta \\
\{\sigma_i\} & \alpha \equiv \text{malloc} \land \sigma_i = \text{get\_heap\_loc}() \\
A(\text{lval}(\alpha, A) \cap S) & \text{otherwise}
\end{cases} \]
Defining Extractor Functions

- Pointer assignment statement \( lhs_n = rhs_n \)
\[
\text{Def}_n = \text{lval}(lhs_n, Ain_n) \\
\text{Kill}_n = \text{lval}(lhs_n, \text{Must}(Ain_n)) \\
\text{Ref}_n = \begin{cases} 
\text{deref}(lhs_n, Ain_n) & \text{Def}_n \cap \text{Lout}_n = \emptyset \\
\text{deref}(lhs_n, Ain_n) \cup \text{ref}(rhs_n, Ain_n) & \text{otherwise} 
\end{cases}
\]
\[
\text{Pointee}_n = \text{rval}(rhs_n, Ain_n)
\]
- Use \( \alpha \) statement
\[
\text{Def}_n = \text{Kill}_n = \text{Pointee}_n = \emptyset \\
\text{Ref}_n = \text{ref}(\alpha, Ain_n)
\]
- Any other statement
\[
\text{Def}_n = \text{Kill}_n = \text{Ref}_n = \text{Pointee}_n = \emptyset
\]

Extensions for Handling Arrays and Pointer Arithmetic

- Grammar.
\[
\alpha := \text{malloc} \mid \& \beta \mid \beta \mid \& \beta + e \\
\beta := x \mid \beta.f \mid \beta \rightarrow f \mid * \beta \mid \beta[e] \mid \beta + e
\]
- Memory model: Named memory locations. No numeric addresses
  - No address calculation
  - R-values of index expressions retained for each dimension
    If \( \text{rval}(x) = 10 \), then \( \text{lval}(a.f[5][2 + x].g) = a.f.5.12.g \)
  - Sizes of the array elements ignored
\[
S = P \cup H \cup G_p \quad \text{(source locations)} \\
T = \text{Var} \cup H \cup G_m \cup \{?\} \quad \text{(target locations)} \\
G_p = R \times (C \cup \text{npF})^* \times (C \cup \text{pF}) \quad \text{(pointers in aggregates)} \\
G_m = R \times (C \cup \text{npF})^* \times (C \cup \text{pF} \cup \text{npF}) \quad \text{(locations in aggregates)}
\]

Extending L-Value Computation to Arrays and Pointer Arithmetic

- Pointer arithmetic does not have an l-value
- For handling arrays
  - if \( e \) cannot be evaluated at compile time, \( \text{eval}(e) = \bot \text{eval} \) (i.e. array accesses in that dimension are treated as index-insensitive)
\[
\text{lval}(\alpha, A) = \begin{cases} 
\{\sigma\} & (\alpha \equiv \sigma) \land (\sigma \in \text{Var}) \\
\{\sigma.f \mid \sigma \in \text{lval}(\beta, A)\} & \alpha \equiv \beta.f \\
\{\sigma.f \mid \sigma \in \text{rval}(\beta, A), \sigma \neq ?\} & \alpha \equiv \beta \rightarrow f \\
\{\sigma \mid \sigma \in \text{rval}(\beta, A), \sigma \neq ?\} & \alpha \equiv \star \beta \\
\{\sigma.\text{eval} \mid \sigma \in \text{lval}(\beta, A)\} & \alpha \equiv \beta[e] \\
\emptyset & \text{otherwise}
\end{cases}
\]

Extending R-Value Computation to Arrays and Pointer Arithmetic

For handling pointer arithmetic

- If the r-value of the pointer is an array location, add \( \text{eval} \) to the offset
- Otherwise, over-approximate the pointees to all possible locations
\[
\text{rval}(\alpha, A) = \begin{cases} 
\text{lval}(\beta, A) & \alpha \equiv \& \beta \\
\{\alpha_i\} & \alpha \equiv \text{malloc} \land \alpha_i = \text{get_heap_loc}() \\
T & (\alpha \equiv \beta + e) \land \\
(\exists \sigma \in \text{rval}(\beta, A), \sigma \neq \sigma'.c, \sigma' \in T, c \in C) & \text{otherwise}
\end{cases}
\]
\[
\bigcup \{\sigma.(c + \text{eval})\} & (\alpha \equiv \beta + e) \land \\
(\sigma.c \in \text{rval}(\beta, A)) \land (c \in C) & \text{otherwise}
\]
\]
Motivating Example for Heap Liveness Analysis

If the while loop is not executed even once.

1. \( w = x \) \( // x \) points to \( m_a \)
2. while \( (x.data < max) \)
3. \( x = x.rptr \)
4. \( y = x.lptr \)
5. \( z = \text{New class of } z \)
6. \( y = y.lptr \)
7. \( z.sum = x.data + y.data \)

Motivating Example for Heap Liveness Analysis

If the while loop is executed once.

1. \( w = x \) \( // x \) points to \( m_a \)
2. while \( (x.data < max) \)
3. \( x = x.rptr \)
4. \( y = x.lptr \)
5. \( z = \text{New class of } z \)
6. \( y = y.lptr \)
7. \( z.sum = x.data + y.data \)

Motivating Example for Heap Liveness Analysis

If the while loop is executed twice.

1. \( w = x \) \( // x \) points to \( m_a \)
2. while \( (x.data < max) \)
3. \( x = x.rptr \)
4. \( y = x.lptr \)
5. \( z = \text{New class of } z \)
6. \( y = y.lptr \)
7. \( z.sum = x.data + y.data \)
The Moral of the Story

- Mappings between access expressions and l-values keep changing
- This is a rule for heap data
  For stack and static data, it is an exception!
- Static analysis of programs has made significant progress for stack and static data.

What about heap data?
  - Given two access expressions at a program point, do they have the same l-value?
  - Given the same access expression at two program points, does it have the same l-value?

---

**Our Solution**

```plaintext
1 w = x
   w = null
2 while (x.data < max)
   { x.lptr = null
   x.rptr = x.lptr.rptr = null
   x.lptr.lptr.lptr = null
   x.lptr.lptr.rptr = null
3 x = x.rptr
4 y = x.lptr
   x.lptr = y.rptr = null
   y.lptr.lptr = y.lptr.rptr = null
5 z = New class of z
   z.lptr = z.rptr = null
6 y = y.lptr
   y.lptr = y.rptr = null
7 z.sum = x.data + y.data
   x = y = z = null
```

---

**Our Solution**

While loop is not executed even once

```plaintext
1 w = x
   w = null
2 while (x.data < max)
   { x.lptr = null
   x.rptr = x.lptr.rptr = null
   x.lptr.lptr.lptr = null
   x.lptr.lptr.rptr = null
3 x = x.rptr
4 y = x.lptr
   x.lptr = y.rptr = null
   y.lptr.lptr = y.lptr.rptr = null
5 z = New class of z
   z.lptr = z.rptr = null
6 y = y.lptr
   y.lptr = y.rptr = null
7 z.sum = x.data + y.data
   x = y = z = null
```

---

**Our Solution**

While loop is executed once

```plaintext
1 w = x
   w = null
2 while (x.data < max)
   { x.lptr = null
   x.rptr = x.lptr.rptr = null
   x.lptr.lptr.lptr = null
   x.lptr.lptr.rptr = null
3 x = x.rptr
4 y = x.lptr
   x.lptr = y.rptr = null
   y.lptr.lptr = y.lptr.rptr = null
5 z = New class of z
   z.lptr = z.rptr = null
6 y = y.lptr
   y.lptr = y.rptr = null
7 z.sum = x.data + y.data
   x = y = z = null
```
Our Solution

While loop is executed twice

y = z = null
1 w = x
w = null
2 while (x.data < max)
   { x.lptr = null
3   x.rptr = x.lptr.rptr = null
   x.lptr.rptr = null
   x.lptr.lptr = null
   x = x.rptr
4   y = x.lptr
   x.lptr = y.rptr = null
   y.lptr.rptr = null
   y.lptr.lptr = null
   y.lptr = y
   y = y.lptr
   y.lptr = y.rptr = null
   y.lptr.lptr.lptr = null
   y.lptr.lptr.rptr = null
   z = New class
5   z.lptr = z.rptr = null
   z.sum = x.data + y.data
6   y = y.lptr
   y.lptr = y.rptr = null
   z = New class
7   z.sum = x.data + y.data
   x = y = z = null

HeapStack

Some Observations

- The memory address that x holds when the execution reaches a given program point is not an invariant of program execution
- Whether we dereference lptr out of x or rptr out of x at a given program point is an invariant of program execution
- A static analysis can discover only some invariants
**Some Observations**

1. \( w = x \)
2. while \( (x.data < max) \)
   
   ```
   \{ x.lptr = null \\
   x.rptr = x.lptr.rptr = null \\
   x.lptr.lptr = null \\
   \}
   ```
3. \( x = x.rptr \)
4. \( y = x.lptr \)
5. \( z = \text{New class of } z \)
6. \( y = y.lptr \)
7. \( z.sum = x.data + y.data \)
8. \( x = y = z = \text{null} \)

**New access expressions are created. Can they cause exceptions?**

**An Overview of Heap Reference Analysis**

- A reference (called a *link*) can be represented by an *access path*.
  - Eg. \( x \rightarrow \text{lptr} \rightarrow \text{rptr} \)
- A link may be accessed in multiple ways
- Setting links to null
  - *Alias Analysis*. Identify all possible ways of accessing a link
  - *Liveness Analysis*. For each program point, identify "dead" links (i.e. links which are not accessed after that program point)
  - *Availability and Anticipability Analyses*. Dead links should be reachable for making null assignment.
  - *Code Transformation*. Set "dead" links to null

**Assumptions**

- For simplicity of exposition
  - Java model of heap access
    - Root variables are on stack and represent references to memory in heap.
    - Root variables cannot be pointed to by any reference.
  - Simple extensions for C++
    - Root variables can be pointed to by other pointers.
    - Pointer arithmetic is not handled.
Key Idea #1: Access Paths Denote Links

- Root variables: x, y, z
- Field names: rptr, lptr
- Access path: x -> rptr -> lptr
  Semantically, sequence of “links”
- Frontier: name of the last link
- Live access path: If the link corresponding to its frontier is used in future

What Makes a Link Live?

Assuming that a statement must be executed, if nullifying a link read in the statement can change the semantics of the program, then the link is live.

Reading a link for accessing the contents of the corresponding target object:

<table>
<thead>
<tr>
<th>Example</th>
<th>Objects read</th>
<th>Live access paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>y = x.rptr</td>
<td>x, O₁, x, x.rptr</td>
<td></td>
</tr>
</tbody>
</table>

Reading a link for copying the contents of the corresponding target object:

<table>
<thead>
<tr>
<th>Example</th>
<th>Objects read</th>
<th>Live access paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>y = x.rptr</td>
<td>x, O₁, x, x.rptr</td>
<td></td>
</tr>
<tr>
<td>x.lptr = y</td>
<td>x, O₁, x, y, y</td>
<td></td>
</tr>
</tbody>
</table>
What Makes a Link Live?

Assuming that a statement must be executed, if nullifying a link read in the statement can change the semantics of the program, then the link is live.

Reading a link for comparing the address of the corresponding target object:

<table>
<thead>
<tr>
<th>Example</th>
<th>Objects read</th>
<th>Live access paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>if (x.lptr == null)</td>
<td>x, O₁</td>
<td>x, x -&gt; lptr</td>
</tr>
</tbody>
</table>

Key Idea #2: Transfer of Access Paths

Liveness Analysis

Program

Statement involving memory references

Effect of the statement on the access paths

Live Access Paths

Live Access Paths

Semantic Information

Generated \{x, x -> n, x -> n -> r\}

Killed \{x, x -> r\}

x after the assignment is same as x -> n before the assignment

Analysis

\{x, x -> r\}
**Key Idea #3: Liveness Closure Under Link Aliasing**

- **x = y**
- **x.n.d**

x and y are node aliases
x.n and y.n are link aliases
x⇒n is live ⇒ y⇒n is live

Nullifying y⇒n will have the side effect of nullifying x⇒n

**Explicit and Implicit Liveness**

- **x = y**
- **x.p = t**
- **y = y.p**

Explicit Liveness

1. \{x, y, y⇒p, y⇒p⇒q\}
2. \{y, y⇒p, y⇒p⇒q\}
3. \{y, y⇒q\}
4. use y.q.d

Required Liveness

1. \{x, y, t, t⇒q\}
2. \{y, y⇒p, y⇒p⇒q\}
3. \{y, y⇒q\}
4. use y.q.d

**Key Idea #4: Aliasing is Required with Explicit Liveness**

1. **x = y**
2. **x.p = t**
3. **y = y.p**
4. use y.q.d

Explicit Liveness

1. \{x, y, y⇒p, y⇒p⇒q\}
2. \{y, y⇒p, y⇒p⇒q\}
3. \{y, y⇒q\}
4. use y.q.d

Required Liveness

1. \{x, y, t, t⇒q\}
2. \{y, y⇒p, y⇒p⇒q\}
3. \{y, y⇒q\}
4. use y.q.d

The need of link alias closure of LHS
- Transferring liveness to RHS (soundness)
- Killing liveness (precision)

Link alias closure of RHS can be computed later for implicit liveness
Notation for Defining Flow Functions for Explicit Liveness

- Basic entities
  - Variables $u, v \in \mathbb{V}ar$
  - Pointer variables $w, x, y, z \in P \subseteq \mathbb{V}ar$
  - Pointer fields $f, g, h \in pF$
  - Non-pointer fields $a, b, c, d \in npF$

- Additional notation
  - Sequence of pointer fields $\sigma \in pF^*$ (could be $\epsilon$)
  - Access paths $\rho \in P \times pF^*$
  - Example: \{\(x, x \rightarrow f, x \rightarrow f \rightarrow g\)\}
  - Summarized access paths rooted at $x$ or $x \sigma$ for a given $x$ and $\sigma$
    - $x^* = \{x \sigma \mid \sigma \in pF^*\}$
    - $x \sigma^* = \{x \sigma \rightarrow \sigma' \mid \sigma' \in pF^*\}$

Data Flow Equations for Explicit Liveness Analysis

\[
In_n = (Out_n - Kill_n(Out_n)) \cup Gen_n(Out_n)
\]

\[
Out_n = \begin{cases} 
\text{BI} & n \text{ is End} \\
\bigcup_{s \in \text{succ}(n)} In_s & \text{otherwise}
\end{cases}
\]

Flow Functions for Explicit Liveness Analysis

Let $A$ denote May Aliases at the exit of node $n$

<table>
<thead>
<tr>
<th>Statement $n$</th>
<th>$Gen_n(X)$</th>
<th>$Kill_n(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = y$</td>
<td>${y \rightarrow \sigma \mid x \rightarrow \sigma \in X}$</td>
<td>$x \rightarrow \ast$</td>
</tr>
<tr>
<td>$x = y.f$</td>
<td>${y \rightarrow f \rightarrow \sigma \mid x \rightarrow \sigma \in X}$</td>
<td>$x \rightarrow \ast$</td>
</tr>
<tr>
<td>$x.f = y$</td>
<td>${y \rightarrow \sigma \mid z \rightarrow f \rightarrow \sigma \in X, z \in A(x)} \cup \bigcup_{z \in \text{Must}(A)(x)} z \rightarrow f \rightarrow \ast$</td>
<td></td>
</tr>
<tr>
<td>$x = \text{new}$</td>
<td>$\emptyset$</td>
<td>$x \rightarrow \ast$</td>
</tr>
<tr>
<td>$x = \text{null}$</td>
<td>$\emptyset$</td>
<td>$x \rightarrow \ast$</td>
</tr>
<tr>
<td>other</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
</tbody>
</table>

Flow Functions for Explicit Liveness Analysis

Let $A$ denote May Aliases at the exit of node $n$

<table>
<thead>
<tr>
<th>Statement $n$</th>
<th>$Gen_n(X)$</th>
<th>$Kill_n(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = y$</td>
<td>${y \rightarrow \sigma \mid x \rightarrow \sigma \in X}$</td>
<td>$x \rightarrow \ast$</td>
</tr>
<tr>
<td>$x = y.f$</td>
<td>${y \rightarrow f \rightarrow \sigma \mid x \rightarrow \sigma \in X}$</td>
<td>$x \rightarrow \ast$</td>
</tr>
<tr>
<td>$x.f = y$</td>
<td>${y \rightarrow \sigma \mid z \rightarrow f \rightarrow \sigma \in X, z \in A(x)} \cup \bigcup_{z \in \text{Must}(A)(x)} z \rightarrow f \rightarrow \ast$</td>
<td></td>
</tr>
<tr>
<td>$x = \text{new}$</td>
<td>$\emptyset$</td>
<td>$x \rightarrow \ast$</td>
</tr>
<tr>
<td>$x = \text{null}$</td>
<td>$\emptyset$</td>
<td>$x \rightarrow \ast$</td>
</tr>
<tr>
<td>other</td>
<td>$\emptyset$</td>
<td>$\emptyset$</td>
</tr>
</tbody>
</table>

- May link aliasing for soundness
- Must link aliasing for precision
Flow Functions for Explicit Liveness Analysis

Let $A$ denote May Aliases at the exit of node $n$

<table>
<thead>
<tr>
<th>Statement $n$</th>
<th>$Gen_n(X)$</th>
<th>$Kill_n(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f = y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\exists x \in Out_n$, we can do dead code elimination</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Why is $y \notin Gen_n(X)$ for $x.f = y$ when $x \notin X$?
  - If $\exists x \notin Out_n$, we can do dead code elimination
- Why is $y \notin Gen_n(X)$ for $x = y.f$ when $x \rightarrow \sigma \notin X$?
  - If $\exists x \rightarrow \sigma \in Out_n$, we can do dead code elimination
- Why is $x \notin Gen_n(X)$ for $x.f = y$?
  - If $\exists x \rightarrow f \rightarrow \sigma \in Out_n$, we can do dead code elimination
  - If $\exists x \rightarrow f \rightarrow \sigma \in Out_n$, then $\exists x \in Out_n$
    It will not be killed, so no need of $x \in Gen_n$

Computing Explicit Liveness Using Sets of Access Paths

Anticipability of Heap References: An All Paths problem

Analysis

$\{x, x \rightarrow n, x \rightarrow n \rightarrow r\}$

$\{x, x \rightarrow r\}$

$\{x, x \rightarrow r\}$

$\ldots = x.r.d$

Analysis

$\{x, x \rightarrow n\}$

$\{x\}$

$\{x, x \rightarrow r\}$

$\ldots = x.r.d$
Computing Explicit Liveness Using Sets of Access Paths

Liveness of Heap References: An *Any Path* problem

\[ \{x, x \rightarrow n, x \rightarrow n \rightarrow n, x \rightarrow n \rightarrow r, x \rightarrow n \rightarrow r \rightarrow \} \]

\[ \{x, x \rightarrow r\} \cup \{x, x \rightarrow n, x \rightarrow n \rightarrow r\} \]

\[ \{x, x \rightarrow r\} \]

\[ \ldots = x \cdot r \cdot d \]

---

Key Idea #5: Using Graphs as Data Flow Values

Different occurrences of n’s in an access path are **Indistinguishable**

(pattern of subsequent dereferences remains same)

Access Graph : \( x \rightarrow n \rightarrow n \rightarrow n \rightarrow n \rightarrow \ldots \)

**Finite Number of Bounded Structures**

---

Key Idea #6: Include Program Point in Graphs

Different occurrences of n’s in an access path are **Distinct**

(pattern of subsequent dereferences could be distinct)

Access Graph : \( x \rightarrow n \rightarrow n_1 \rightarrow n_2 \rightarrow f_2 \rightarrow \)
Inclusion of Program Point Facilitates Summarization

$G_1 = G_2 \sqcup G_3$

Analysis

Iteration #1

Analysis

1. $x = x.n$
2. $\ldots = x.r$

Iteration #2

Analysis

1. $x = x.n$
2. $\ldots = x.r$

Iteration #3

Analysis

1. $x = x.n$
2. $\ldots = x.r$

Sep 2017 IIT Bombay
Access Graph and Memory Graph

Program Fragment

Memory Graph

Access Graphs

Memory Graph: Nodes represent locations and edges represent links (i.e. pointers).
Access Graphs: Nodes represent dereference of links at particular statements. Memory locations are implicit.

Lattice of Access Graphs

- Finite number of nodes in an access graph for a variable
- $\cup$ induces a partial order on access graphs
  - a finite (and hence complete) lattice
  - All standard results of classical data flow analysis can be extended to this analysis.
    
    *Termination and boundedness, convergence on MFP, complexity etc.*

Access Graph Operations

- Union. $G \cup G'$
- Path Removal
  
  $G \ominus R$ removes those access paths in $G$ which have $\rho \in R$ as a prefix
- Factorization (/)
- Extension

Defining Factorization

Given statement $x.n = y$, what should be the result of transfer?

<table>
<thead>
<tr>
<th>Live AP</th>
<th>Memory Graph</th>
<th>Transfer</th>
<th>Remainder</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x \rightarrow n \rightarrow r$</td>
<td>$x \rightarrow n \rightarrow r \rightarrow y \rightarrow r$</td>
<td>$r$ (LHS is contained in the live access path)</td>
<td></td>
</tr>
<tr>
<td>$x \rightarrow n$</td>
<td>$x \rightarrow n \rightarrow r \rightarrow y$</td>
<td>$\epsilon$ (LHS is contained in the live access path)</td>
<td></td>
</tr>
<tr>
<td>$x$</td>
<td>$x \rightarrow n \rightarrow r \rightarrow y$</td>
<td>no transfer</td>
<td>?? (LHS is not contained in the live access path) Quotient is empty So no remainder</td>
</tr>
</tbody>
</table>
### Semantics of Access Graph Operations

- \( P(G) \) is the set of all paths in graph \( G \)
- \( P(G, M) \) is the set of paths in \( G \) terminating on nodes in \( M \)
- \( S \) is the set of remainder graphs
- \( P(S) \) is the set of all paths in all remainder graphs in \( S \)

<table>
<thead>
<tr>
<th>Operation</th>
<th>Access Paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Union</td>
<td>( G_3 = G_1 \cup G_2 ) ( P(G_3) \supseteq P(G_1) \cup P(G_2) )</td>
</tr>
<tr>
<td>Path Removal</td>
<td>( G_2 = G_1 \ominus X ) ( P(G_2) \supseteq P(G_1) - {\rho \rightarrow \sigma \mid \rho \in X, \rho \rightarrow \sigma \in P(G_1)} )</td>
</tr>
<tr>
<td>Factorization</td>
<td>( S = G_1/\rho ) ( P(S) = {\sigma \mid \rho \rightarrow \sigma \in P(G_1)} )</td>
</tr>
<tr>
<td>Extension</td>
<td>( G_2 = (G_1, M)#S ) ( P(G_2) \supseteq P(G_1, M), \sigma \in P(S) )</td>
</tr>
</tbody>
</table>

**Remainder is empty**

**Quotient is empty**

### Access Graph Operations: Examples

**Program**

| 1 | \( x = x./l \) |
| 2 | \( y = x.r.d \) |

**Access Graphs**

1. \( g_4 \)
2. \( g_5 \)
3. \( g_6 \)
4. \( g_7 \)

**Remainder Graphs**

1. \( r_{g_1} \)
2. \( r_{g_2} \)
3. \( r_{g_3} \)

<table>
<thead>
<tr>
<th>Operation</th>
<th>Path Removal</th>
<th>Factorisation</th>
<th>Extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Union</td>
<td>( g_4 \cup g_4 = g_4 ) ( g_6 \ominus {x \rightarrow l} = g_2 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( g_2 \cup g_4 = g_4 ) ( g_5 \ominus {x} = \mathcal{E}_G )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( g_5 \ominus {x, h_1} # {r_{g_1}} = g_6 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( g_2 \ominus {r_{g_2}} # {r_{g_2}} = g_6 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( g_6 \ominus {x \rightarrow r} = g_4 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( g_5 \ominus {x \rightarrow r} = \mathcal{E}_{RG} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( g_6 \ominus {x \rightarrow r} = \emptyset )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( g_2 \ominus {r_{g_2}} # \emptyset = \mathcal{E}_G )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Flow Equations for Explicit Liveness Analysis: Access Graphs Version

\[ \text{In}_n = (\text{Out}_n \circ \text{Kill}_n(\text{Out}_n)) \uplus \text{Gen}_n(\text{Out}_n) \]

\[ \text{Out}_n = \begin{cases} \text{BI} & \text{n is End} \\ \text{In}_s & \text{otherwise} \end{cases} \]

- \text{In}_n, \text{Out}_n, \text{and} \text{Gen}_n \text{ are access graphs}
- \text{Kill}_n \text{ is a set of access paths}

Flow Functions for Explicit Liveness Analysis: Access Paths Version

Let \( A \) denote May Aliases at the exit of node \( n \)

<table>
<thead>
<tr>
<th>Statement ( n )</th>
<th>Gen(_n)(( X ))</th>
<th>Kill(_n)(( X ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = y )</td>
<td>( { y \rightarrow \sigma</td>
<td>x \rightarrow \sigma \in X } )</td>
</tr>
<tr>
<td>( x = y.f )</td>
<td>( { y \rightarrow f \rightarrow \sigma</td>
<td>x \rightarrow \sigma \in X } )</td>
</tr>
<tr>
<td>( x.f = y )</td>
<td>( { y \rightarrow \sigma</td>
<td>z \rightarrow f \rightarrow \sigma \in X, z \in A(x) } \cup { z \rightarrow f \rightarrow \ast } )</td>
</tr>
<tr>
<td>( x = \text{new} )</td>
<td>( \emptyset )</td>
<td>( x \rightarrow \ast )</td>
</tr>
<tr>
<td>( x = \text{null} )</td>
<td>( \emptyset )</td>
<td>( x \rightarrow \ast )</td>
</tr>
<tr>
<td>other</td>
<td>( \emptyset )</td>
<td>( \emptyset )</td>
</tr>
</tbody>
</table>

May link aliasing for soundness
Must link aliasing for precision

Flow Functions for Explicit Liveness Analysis: Access Graphs Version

- \( A \) denotes May Aliases at the exit of node \( n \)
- \( \text{mkGraph}(\rho) \) creates an access graph for access path \( \rho \)

<table>
<thead>
<tr>
<th>Statement ( n )</th>
<th>Gen(_n)(( X ))</th>
<th>Kill(_n)(( X ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = y )</td>
<td>( \text{mkGraph}(y)#(X/x) )</td>
<td>( { x } )</td>
</tr>
<tr>
<td>( x = y.f )</td>
<td>( \text{mkGraph}(y\rightarrow f)#(X/x) )</td>
<td>( { x } )</td>
</tr>
<tr>
<td>( x.f = y )</td>
<td>( \text{mkGraph}(y)\left( \cup_{z \in A(x)} (X/(z \rightarrow f)) \right) \cup { z \rightarrow f</td>
<td>z \in \text{Must}(A)(x) } )</td>
</tr>
<tr>
<td>( x = \text{new} )</td>
<td>( \emptyset )</td>
<td>( { x } )</td>
</tr>
<tr>
<td>( x = \text{null} )</td>
<td>( \emptyset )</td>
<td>( { x } )</td>
</tr>
<tr>
<td>other</td>
<td>( \emptyset )</td>
<td>( \emptyset )</td>
</tr>
</tbody>
</table>
Liveness Analysis of Example Program: 1st Iteration

1. \( w = x \)

2. while (x.data < max)

4. \( y = x\.lptr \)

5. \( z = \text{New class of } x \)

6. \( y = y\.lptr \)

7. \( z\.sum = x\.data + y\.data \)

Liveness Analysis of Example Program: 2nd Iteration

1. \( w = x \)

2. while (x.data < max)

4. \( y = x\.lptr \)

5. \( z = \text{New class of } x \)

6. \( y = y\.lptr \)

7. \( z\.sum = x\.data + y\.data \)

Liveness Analysis of Example Program: 3rd Iteration

1. \( w = x \)

2. while (x.data < max)

4. \( y = x\.lptr \)

5. \( z = \text{New class of } x \)

6. \( y = y\.lptr \)

7. \( z\.sum = x\.data + y\.data \)

Liveness Analysis of Example Program: 4th Iteration

1. \( w = x \)

2. while (x.data < max)

4. \( y = x\.lptr \)

5. \( z = \text{New class of } x \)

6. \( y = y\.lptr \)

7. \( z\.sum = x\.data + y\.data \)
Tutorial Problem for Explicit Liveness (1)
Construct access graphs at the entry of block 1 for the following programs

A
1 \( x = x.n \)
2 \( x = x.n \)
3 Use \( x.r.d \)

B
1 
2 \( x = x.n \)
3 Use \( x.r.d \)

C
1 
2 \( x = x.n \)
3 Use \( x.r.d \)

D
1 
2 \( x = x.n \)
3 Use \( x.r.d \)

E
1 
2 \( x = x.n \)
3 \( x = x.l \)
4 Use \( x.r.d \)

F
1 
2 \( x = x.n \)
3 \( x = x.l \)
4 Use \( x.r.d \)

Sep 2017
IIT Bombay

Tutorial Problem for Explicit Liveness (2)

- Unfortunately the student who constructed these access graphs forgot to attach statement numbers as subscripts to node labels and has misplaced the programs which gave rise to these graphs.
- Please help her by constructing CFGs for which these access graphs represent explicit liveness at some program point in the CFGs.

The final magic!!

Rotate each picture anti-clockwise by 90° and compare it with its access graph.

The structure of access graph of variable \( x \) is identical to the control flow structure between pointer assignments of \( x \).
Tutorial Problem for Explicit Liveness (3)

- Compute explicit liveness for the program.
- Are the following access paths live at node 1?
  Show the corresponding execution sequence of statements

  P1: \( y \to m \to l \)
  P2: \( y \to l \to n \to m \)
  P3: \( y \to l \to n \)
  P4: \( y \to n \to l \to n \)

\[ x = z_1 \]
\[ x = y \]
\[ x.n = y.m \]
\[ y = x.n \]
\[ \text{use } x.d \]

Which Access Paths Can be Nullified?

- Consider extensions of accessible paths for nullification.
  Let \( \rho \) be accessible at \( p \) (i.e., available or anticipable) for each reference field \( f \) of the object pointed to by \( \rho \) if \( \rho \to f \) is not live at \( p \) then
    Insert \( \rho \to f = \text{null at } p \) subject to profitability

- For simple access paths, \( \rho \) is empty and \( f \) is the root variable name.

Availability and Anticipability Analyses

- \( \rho \) is available at program point \( p \) if the target of each prefix of \( \rho \) is guaranteed to be created along every control flow path reaching \( p \).
- \( \rho \) is anticipable at program point \( p \) if the target of each prefix of \( \rho \) is guaranteed to be dereferenced along every control flow path starting at \( p \).
- Finiteness.
  ▶ An anticipable (available) access path must be anticipable (available) along every paths. Thus unbounded paths arising out of loops cannot be anticipable (available).
  ▶ Due to “every control flow path nature”, computation of anticipable and available access paths uses \( \cap \) as the confluence. Thus the sets are bounded.

\[ \Rightarrow \] No need of access graphs.

Availability Analysis of Example Program
### Anticipability Analysis of Example Program

```
{w} 1 w = x
{w} 2 while (x.data < max)
    {x} 3 x = x.rptr
    {x, y, z} 4 y = x.lptr
    {x, y, z} 5 z = New class of z
    {x, y, z} 6 y = y.lptr
    {x, y, z} 7 z.sum = x.data + y.data
∅
```

### Live and Accessible Paths

```
{x} 1 w = x
{x} 2 while (x.data < max)
    {x} 3 x = x.rptr
    {x} 4 y = x.lptr
    {x, y} 5 z = New class of z
    {x, y} 6 y = y.lptr
    {x, y} 7 z.sum = x.data + y.data
∅
```

### Creating null Assignments from Live and Accessible Paths

```
y = z = null
1 w = x
w = null
2 while (x.data < max)
    {x} 3 x = x.rptr
    x.lptr = null
    y.lptr = y.rptr = null
    z.lptr = z.rptr = null
    x.rptr = x.lptr.rptr = null
    x.lptr.lptr.lptr = null
    x.lptr.lptr.rptr = null
    y.lptr.lptr = y.lptr.rptr = null
    x = y = z = null
```

### The Resulting Program

```
y = z = null
1 w = x
w = null
2 while (x.data < max)
    {x} 3 x = x.rptr
    x.lptr = null
    y.lptr = y.rptr = null
    z.lptr = z.rptr = null
    x.rptr = x.lptr.rptr = null
    x.lptr.lptr.lptr = null
    x.lptr.lptr.rptr = null
    y.lptr.lptr = y.lptr.rptr = null
    x = y = z = null
```
Overapproximation Caused by Our Summarization

- The program allocates \( x \rightarrow p \) in one iteration and uses it in the next
- First \( x \rightarrow p \) used in statement 3
- Second \( x \rightarrow p \) used in statement 4
- Third \( p \) is reallocated
Non-Distributivity of Explicit Liveness Analysis

1. \( x.n = \text{null} \)
2. \( x = x.n \)
3. \( x.n.n = \text{null} \)
4. \( x = x.r \)
5. \( x.n.r = \text{null} \)
6. \( x = x.n \)
7. \( \text{use } x.n.d \)
8. \( \text{use } x.r.d \)

Access path \( x \rightarrow n \rightarrow r \) (shown in blue color) is a spurious access path that arises due to \( \sqcup \) and is not removed by the assignment in node 1.
Issues Not Covered

- Precision of information
  ▶ Cyclic Data Structures
  ▶ Eliminating Redundant null Assignments
- Properties of Data Flow Analysis:
  Monotonicity, Boundedness, Complexity
- Interprocedural Analysis
- Extensions for C/C++
- Formulation for functional languages
- Issues that need to be researched: Good alias analysis of heap

BTW, What is Static Analysis of Heap?

Abstract, Bounded, Single Instance
Concrete, Unbounded, Infinitely Many

Static

Program Code

Static Analysis

Summary

Heap Data

Dynamic

Profiling

Conclusions

- Unbounded information can be summarized using interesting insights
  ▶ Contrary to popular perception, heap structure is not arbitrary
  *Heap manipulations consist of repeating patterns which bear a close resemblance to program structure*
  Analysis of heap data is possible despite the fact that the mappings between access expressions and l-values keep changing