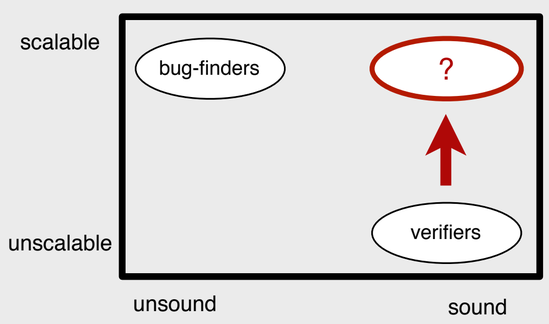
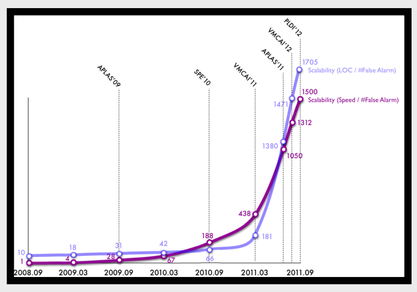
Design and Implentation of Sparse Global Analyses for C-like Languages

Hakjoo Oh, Kihong Heo, Wonchan Lee, Woosuk Lee, Kwangkeun Yi
Seoul National University

1. Motivation

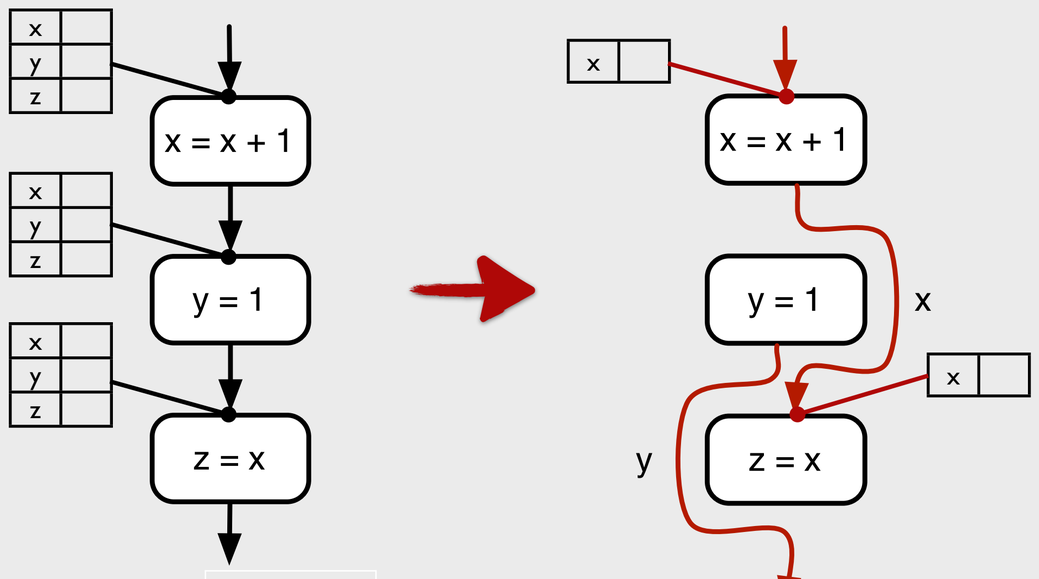
- Challenge : sound, precise, global yet scalable
- Sparrow, our commercialized bug-finder
  - sound in design
  - unsound yet scalable in reality
- Goal : scaling up its sound & global version





2. Overview

- Key : Right Part at Right Moment
  - Follow the data dependency, not the ctrl flow blindly



- Program : $\langle \mathbb{C}, \hookrightarrow \rangle$
  - $\mathbb{C}$: set of program points
  - $\hookrightarrow \subseteq \mathbb{C} \times \mathbb{C}$: control flow relation
- Abstract Semantics

$$[\![P]\!] \in \mathbb{C} \to \mathbb{S} = \mathbf{fix}F \qquad F \in (\mathbb{C} \to \mathbb{S}) \to (\mathbb{C} \to \mathbb{S})$$

$$\hat{\mathbb{S}} = \hat{\mathbb{L}} \to \hat{\mathbb{V}} \qquad \hat{F}(\hat{X}) = \lambda c \in \mathbb{C}.\hat{f}_c(\bigsqcup_{c' \hookrightarrow c} \hat{X}(c'))$$

- Precision Preserving Framework

$$\hat{F}(\hat{X}) = \lambda c \in \mathbb{C}.\hat{f}_c(\bigsqcup_{c' \hookrightarrow c} \hat{X}(c')) \to \hat{F}(\hat{X}) = \lambda c \in \mathbb{C}.\hat{f}_c(\bigsqcup_{c' \overset{l}{\rightsquigarrow}_a c} \hat{X}(c'))$$

$$\mathbf{fix}F = \mathbf{fix}F_s$$

3. Sparse Analysis Framework

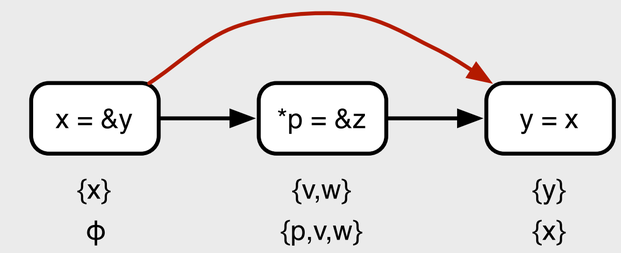
- Data Dependency

$$c_0 \overset{l}{\rightsquigarrow} c_n \triangleq \exists c_0 \ldots c_n \in \mathsf{Path}, l \in \mathbb{L}. \; l \in \mathsf{D}(c_0) \cap \mathsf{U}(c_n) \wedge \forall i \in (0,n).l \notin \mathsf{D}(c_i)$$

$$\mathsf{D}(c) \triangleq \{l \in \hat{\mathbb{L}} \mid \exists \hat{s} \sqsubseteq \bigsqcup_{c' \hookrightarrow c} (\mathbf{fix}\hat{F})(c').\hat{f}_c(\hat{s})(l) \neq \hat{s}(l)\}$$

$$\mathsf{U}(c) \triangleq \{l \in \hat{\mathbb{L}} \mid \exists \hat{s} \sqsubseteq \bigsqcup_{c' \hookrightarrow c} (\mathbf{fix}\hat{F})(c').\hat{f}_c(\hat{s})|_{\mathsf{D}(c)} \neq \hat{f}_c(\hat{s}\backslash l)|_{\mathsf{D}(c)}\}$$

x



- Approximated Data Dependency

$$c_0 \overset{l}{\rightsquigarrow}_a c_n \triangleq \exists c_0 \ldots c_n \in \mathsf{Path}, l \in \mathbb{L}. \; l \in \hat{\mathsf{D}}(c_0) \cap \hat{\mathsf{U}}(c_n) \wedge \forall i \in (0,n).l \notin \mathsf{D}(c_i)$$

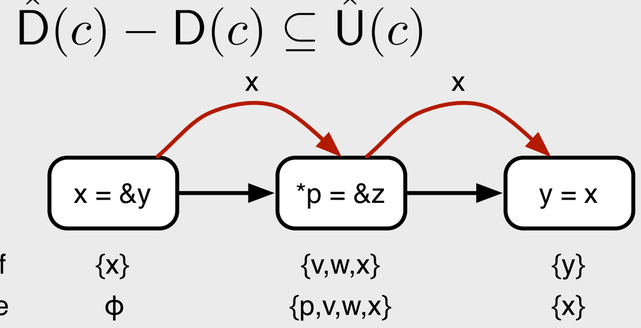
  - over-approximation

$$\hat{\mathsf{D}}(c) \supseteq \mathsf{D}(c) \wedge \hat{\mathsf{U}}(c) \supseteq \mathsf{U}(c)$$

  - spurious definitions should be included in uses

$$\hat{\mathsf{D}}(c) - \mathsf{D}(c) \subseteq \hat{\mathsf{U}}(c)$$

x x



4. Experiments

- Interval analysis

Program	LOC	Vanilla		Sparse		Spd↑	Mem ↓
		Time	Mem	Time	Mem		
make-3.76.1	27K	24240	1391	21	114	1154x	92%
wget-1.9	35K	44092	2546	11	85	4008x	97%
screen-4.0.2	45K	∞	∞	767	303	N/A	N/A
emacs-22.1	400K	∞	∞	37,830	7795	N/A	N/A
linux-3.0	710K	∞	∞	33,618	2,0529	N/A	N/A
ghostscript-0.00	1,363K	∞	∞	14,814	6384	N/A	N/A

- Octagon analysis

Program	LOC	Vanilla		Sparse		Spd↑	Mem ↓
		Time	Mem	Time	Mem		
gzip-1.2.4a	7K	2078	2832	21	269	98x	91%
bc-1.06	13K	9536	6987	55	358	173x	95%
make-3.76.1	28K	∞	∞	331	666	N/A	N/A
wget-1.9	35K	∞	∞	288	646	N/A	N/A
screen-4.02	45K	∞	∞	1,6433	9199	N/A	N/A
sendmail-8.13.6	130K	∞	∞	6,4808	2,9658	N/A	N/A

5. Conclusions

- A general framework for sparse analysis
  - data dependency
  - safe & correct def/use sets
- Scale-up of an sound & global static analyzer
  - 1.3 million LOC with interval domain
  - 0.13 million LOC with octagon domain