

Frameworks for Interprocedural Analysis

Helmut Seidl
Vesal Vojdani
Kalmer Apinis

MOD 2013

Frameworks for Interprocedural Analysis

The Goblint Approach

Helmut Seidl
Vesal Vojdani
Kalmer Apinis

MOD 2013

Analyzer Goblin

Certify absence of concurrency bugs in C !

Automotive

Avionics

Analyzer Goblint

Certify absence of concurrency bugs in C !

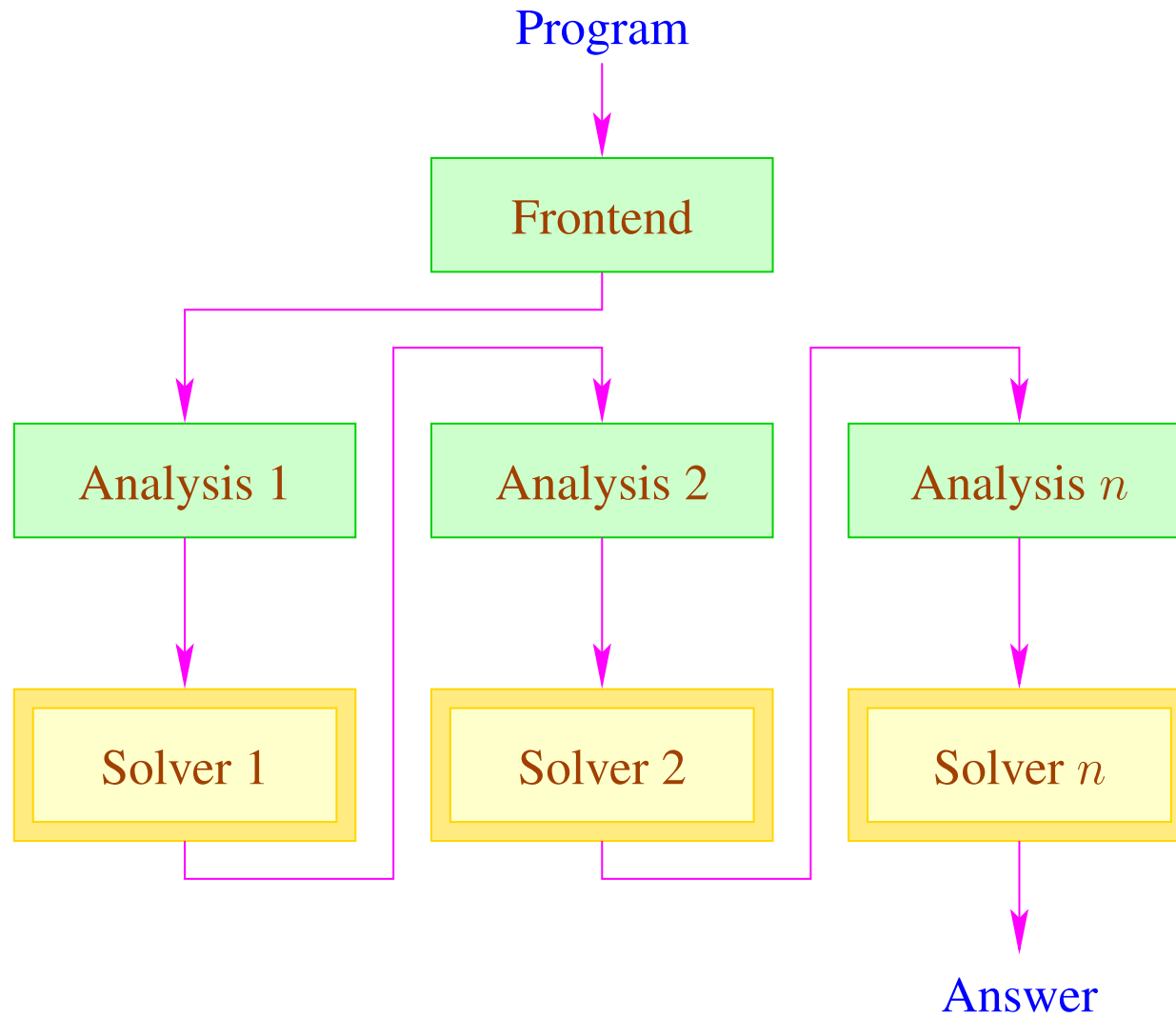
Automotive

Avionics

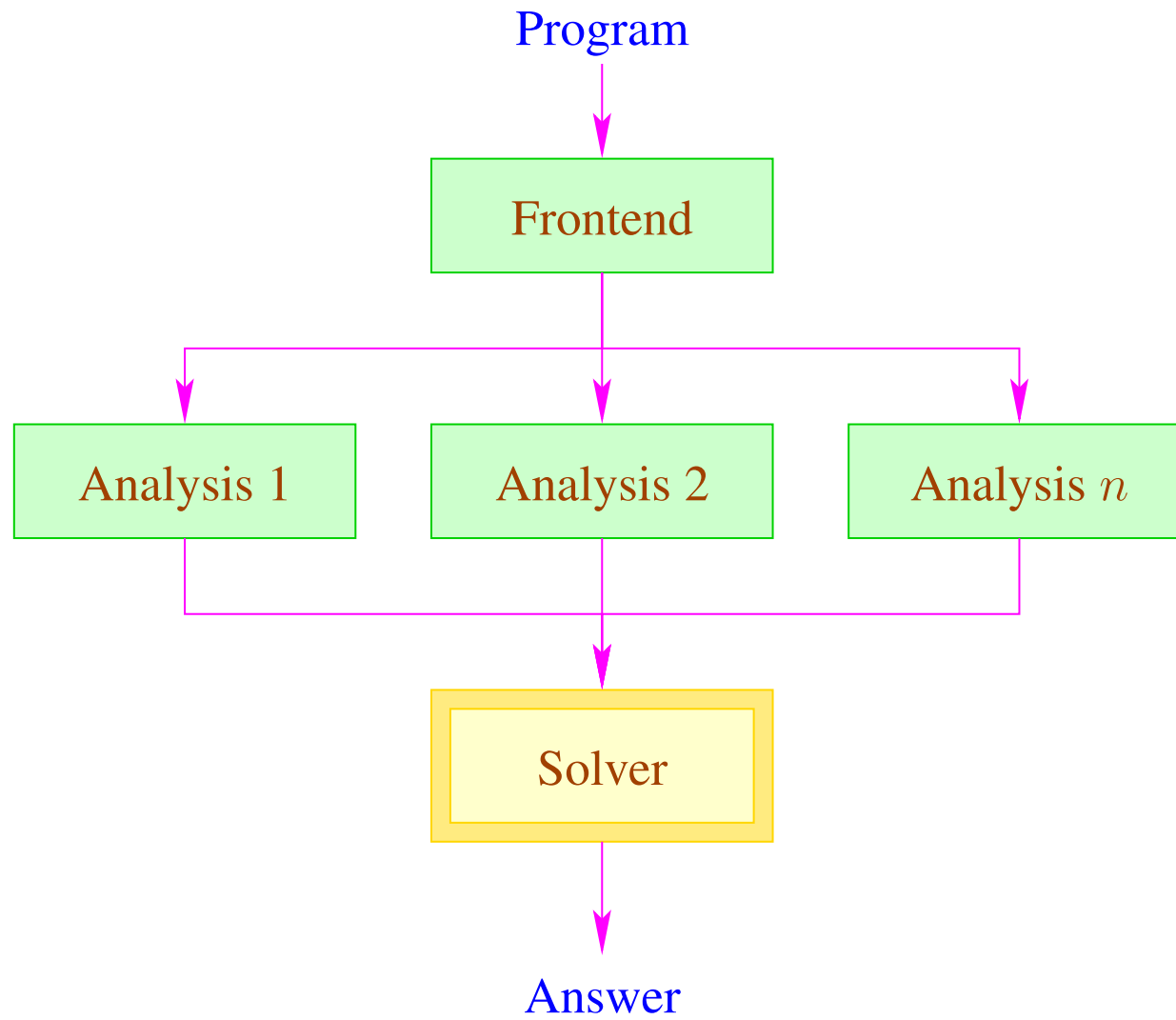
Challenges

- ⇒⇒ Control-flow may depend on data.
- ⇒⇒ Call-graph may depend on pointer analysis.
- ⇒⇒ Multi-threading.
- ⇒⇒ Decent scalability.

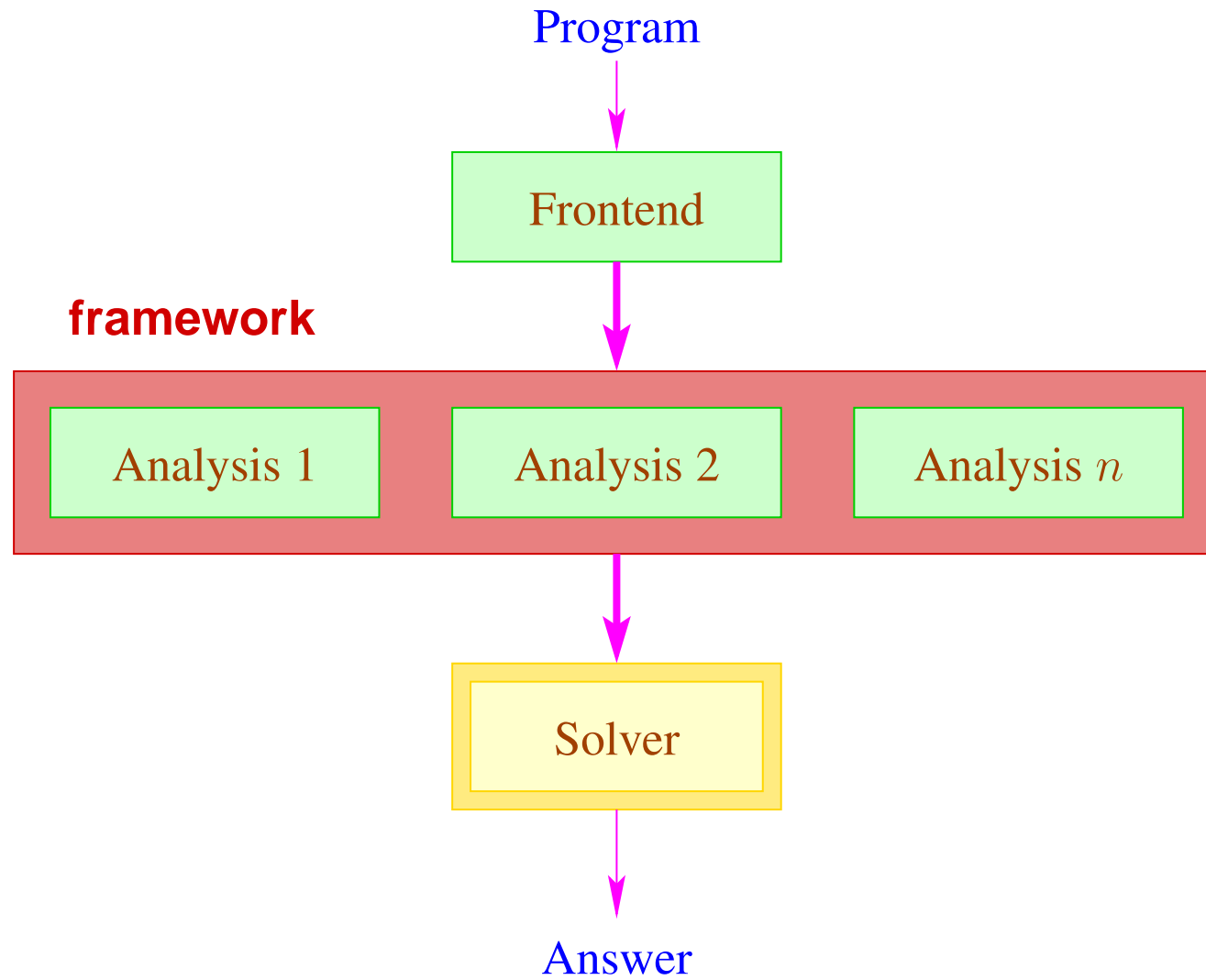
Architecture:



Dream:



Goblint:



Organization

Part 1: Constraint Systems

Part 2: Widening and Narrowing

Organization

Part 1: Constraint Systems

Part 2: Widening and Narrowing

Constraint Systems

Overview — The Problem

- Context-insensitive Analysis
- Context-sensitive Analysis
- Partially Context-sensitive Analysis

...

Overview — The Problem

- Context-insensitive Analysis
finite constraint systems
- Context-sensitive Analysis
- Partially Context-sensitive Analysis

...

Overview — The Problem

- Context-insensitive Analysis
finite constraint systems
- Context-sensitive Analysis
infinite constraint systems
- Partially Context-sensitive Analysis

...

Overview — The Problem

- Context-insensitive Analysis
finite constraint systems
- Context-sensitive Analysis
infinite constraint systems
- Partially Context-sensitive Analysis
infinite constraint systems, infinite dependences
- ...

Overview — Our Solution

- Side-effecting Constraint Systems
- Sharir-Pnueli
- Global Invariants
- Some Experiments
- Lessons learnt

Context-insensitive Analysis

- Represent program by control-flow graph
- Compute one abstract value per program point

a_0 : \mathbb{D} value at program start

$\llbracket s \rrbracket^\#$: $\mathbb{D} \rightarrow \mathbb{D}$ abstract semantics

$\mathcal{D}[v]$ \equiv value at program point v

The Corresponding Constraint System:

$$\begin{array}{lll} \mathcal{D}[v] & \sqsupseteq & a_0 \quad v \text{ entry point} \\ \mathcal{D}[v] & \sqsupseteq & [[s]]^\# (\mathcal{D}[u]) \quad (u, s, v) \text{ edge} \end{array}$$

Discussion

- The constraint system is **finite**.
- Right-hand sides are **monotonic**.
- Dependences between unknowns are **explicitly given** by the control-flow graph.

Discussion

- The constraint system is **finite**.
- Right-hand sides are **monotonic**.
- Dependences between unknowns are **explicitly given** by the control-flow graph

\implies any fixpoint algo will do !

Context-sensitive Analysis

For each procedure, determine a summary:

$$\mathbb{D} \rightarrow \mathbb{D}$$

that describes the **abstract** effect.

Context-sensitive Analysis

For each procedure, determine a summary:

$$\mathbb{D} \rightarrow \mathbb{D}$$

that describes the **abstract** effect.

Problem

- Practical summaries are **extremely complicated**.
- No succinct representations.
- There are **infinite ascending chains**.

Idea

Cousot '77, Sharir/Pnueli '81

Analyze each procedure f for each abstract context a separately !

$\mathcal{D}[v, a] \in \mathbb{D} \quad \equiv \quad$ value at program point v
within a call for context a .

Idea

Cousot '77, Sharir/Pnueli '81

Analyze each procedure f for each abstract context a separately !

$\mathcal{D}[v, a] \in \mathbb{D} \quad \equiv \quad$ value at program point v
within a call for context a .

Auxiliary Functions:

$\text{enter}^\# \quad : \quad \mathbb{D} \rightarrow \mathbb{D} \quad \text{passing of parameters}$

$\text{combine}^\# \quad : \quad \mathbb{D} \rightarrow \mathbb{D} \rightarrow \mathbb{D} \quad \text{returning of results}$

The Constraint System:

$$\begin{array}{ll} \mathcal{D}[st_f, a] \sqsupseteq a & st_f \text{ entry point of } f \\ \mathcal{D}[v, a] \sqsupseteq \llbracket s \rrbracket^\# (\mathcal{D}[u, a]) & (u, s, v) \text{ edge} \end{array}$$

The Constraint System:

$$\begin{aligned} \mathcal{D}[st_f, a] &\sqsupseteq a && st_f \text{ entry point of } f \\ \mathcal{D}[v, a] &\sqsupseteq \llbracket s \rrbracket^\# (\mathcal{D}[u, a]) && (u, s, v) \text{ edge} \\ \mathcal{D}[v, a] &\sqsupseteq \text{combine}^\# (\mathcal{D}[u, a]) (\mathcal{D}[rt_f, \text{enter}^\# (\mathcal{D}[u, a])]) && \\ &&& (u, f(), v) \text{ call} \\ &&& rt_f \text{ return point of } f \end{aligned}$$

Problem

- This constraint system is generally **infinite** ?
- The dependences between unknowns **change** ?
- Right-hand sides are no longer **monotonic** ?

Problem

- This constraint system is generally **infinite** ?
- The dependences between unknowns **change** ?
- Right-hand sides are no longer **monotonic** ?

Idea

- Do not solve it **completely** !
- Analyse only calls that are **queried** when determining the value of

$$\mathcal{D}[rt_{\text{main}}, \text{enter}^{\#} a_0]$$

Problem

- This constraint system is generally **infinite** ?
- The dependences between unknowns **change** ?
- Right-hand sides are no longer **monotonic** ?

Idea

- Do not solve it **completely** !
- Analyse only calls that are **queried** when determining the value of

$$\mathcal{D}[rt_{\text{main}}, \text{enter}^{\#} a_0]$$

\implies **local solving**

Generic Local Solving — Summary

- Start with the exploration of some **interesting** unknown.
- Identify dependences **on the fly** !
- Explore values of unknowns **on demand** !
- Use **accumulation** to deal with non-monotonicity !

Generic Local Solving — Summary

- Start with the exploration of some **interesting** unknown.
- Identify dependences **on the fly** !
- Explore values of unknowns **on demand** !
- Use **accumulation** to deal with non-monotonicity !

Simplification

$$x \sqsupseteq f_x \quad (x \in \text{Unknowns})$$

The Generic Local Solver LR

- Start with the exploration of some **interesting** unknown.
- If during evaluation of the right-hand side f_x of x , an unknown y is accessed, y is first solved recursively. Then x is added to $\text{infl}[y]$...

$\text{eval } x \ y = \text{solve } y;$

$\text{infl}[y] = \text{infl}[y] \cup \{x\};$

$D[y];$

\implies detection of dependences on the fly !!

The Function `solve` :

`solve x =`

`t = fx (eval x); t = D[x] ⊔ t;`

`if (t ≠ D[x]) {`

`W = infl[x];`

`D[x] = t;`

`iter solve W;`

`}`

`}`

Preventing Non-termination

- Recursion will cause infinite descent into unknowns and thus result in **non-termination** ??
- In order to prevent that, a set **Stable** of unknown is maintained for which **solve** just looks up their values
...
- Initially, **Stable** = \emptyset .

The Function `solve` :

```
solve  $x$  = if ( $x \notin \text{Stable}$ ) {  
     $\text{Stable} = \text{Stable} \cup \{x\}$ ;  
     $t = f_x(\text{eval } x)$ ;     $t = D[x] \sqcup t$ ;  
    if ( $t \neq D[x]$ ) {  
         $W = \text{infl}[x]$ ;  
         $D[x] = t$ ;  
         $\text{Stable} = \text{Stable} \setminus W$ ;  
        iter solve  $W$ ;  
    }  
}
```

The Function `solve` :

```
solve  $x$  = if ( $x \notin \text{Stable}$ ) {  
     $\text{Stable} = \text{Stable} \cup \{x\}$ ;  
     $t = f_x(\text{eval } x)$ ;     $t = D[x] \sqcup t$ ;  
    if ( $t \neq D[x]$ ) {  
         $W = \text{infl}[x]$ ;     $\text{infl}[x] = \emptyset$ ;  
         $D[x] = t$ ;  
         $\text{Stable} = \text{Stable} \setminus W$ ;  
        iter solve  $W$ ;  
    }  
}
```

Example:

Consider the system:

$$x_1 \supseteq \{a\} \cup x_3$$

$$x_2 \supseteq x_3 \cap \{a, b\}$$

$$x_3 \supseteq x_1 \cup \{c\}$$

A trace of the fixpoint algorithm then looks as follows:

solve x_2

eval $x_2 x_3$

solve x_3

eval $x_3 x_1$

solve x_1

eval $x_1 x_3$

solve x_3
stable!

$$\text{infl}[x_3] = \{x_1\} \\ \Rightarrow \emptyset$$

$$D[x_1] = \{a\}$$

$$\text{infl}[x_1] = \{x_3\} \\ \Rightarrow \{a\}$$

$$D[x_3] = \{a, c\}$$

$$\text{infl}[x_3] = \emptyset$$

solve x_1

eval $x_1 x_3$

solve x_3
stable!

$$\text{infl}[x_3] = \{x_1\} \\ \Rightarrow \{a, c\}$$

$$D[x_1] = \{a, c\}$$

$$\text{infl}[x_1] = \emptyset$$

solve x_3

eval $x_3 x_1$

solve x_1
stable!

$$\text{infl}[x_1] = \{x_3\} \\ \Rightarrow \{a, c\}$$

ok

$$\text{infl}[x_3] = \{x_1, x_2\} \\ \Rightarrow \{a, c\}$$

$$D[x_2] = \{a\}$$

Generic Local Solving — Discussion

- It is reasonably efficient !
- It terminates if \mathbb{D} has only finite ascending chains and only finitely many contexts are encountered.

Generic Local Solving — Discussion

- It is reasonably efficient !
- It terminates if \mathbb{D} has only finite ascending chains and only finitely many contexts are encountered.
- It does not require pre-processing of constraint system !

Generic Local Solving — Discussion

- It is reasonably efficient !
- It terminates if \mathbb{D} has only finite ascending chains and only finitely many contexts are encountered.
- It does not require pre-processing of constraint system !
- Solver LR is small, but intricate \implies verification

Hofmann, Karbyshev, S. 2010

So far:

- Context-insensitive analysis
- Context-sensitive analysis

Specification

complete lattice

unknowns

constraints

Solving

Local fixpoint algorithm LR

Incredients of Spec

- \mathbb{D} — complete lattice of invariants
 - \perp least element – **unreachability**
 - \top greatest element – **don't know**
 - \sqsubseteq ordering
 - \sqcup least upper bound
- $\llbracket i = j - 1; \rrbracket^\#$ — abstract semantics
 - transfer function $\mathbb{D} \rightarrow \mathbb{D}$
 - should preserve \perp

Incredients of Spec (cont.)

Unknowns

$\mathcal{D}[v]$ — invariant at program point v
// context-insensitive analysis

$\mathcal{D}[v, a]$ — invariant at program point v
for context a
// context-sensitive analysis

Constraints $x \sqsupseteq f$

$f : (\text{Unknowns} \rightarrow \mathbb{D}) \rightarrow \mathbb{D}$

Disadvantage of LR

Reachability information only **implicitly** given ...

Disadvantage

Reachability information only **implicitly** given by:

$$\text{Reach}[v] = \bigsqcup \{ \mathcal{D}[v, a] \mid a \text{ encountered context} \}$$

Disadvantage

Reachability information only **implicitly** given by:

$$\text{Reach}[v] = \bigsqcup \{ \mathcal{D}[v, a] \mid a \text{ encountered context} \}$$

Explicit Reachability

$$\mathcal{D}[st_{\text{main}}, a] \sqsupseteq a \quad a = \text{enter}^\# a_0$$

$$\mathcal{D}[st_f, a] \sqsupseteq \bigsqcup \{ a \mid a = \text{enter}^\# (\mathcal{D}[u, a']) \}$$

f procedure

Disadvantage

Reachability information only **implicitly** given by:

$$\text{Reach}[v] = \bigsqcup \{ \mathcal{D}[v, a] \mid a \text{ encountered context} \}$$

Explicit Reachability

$$\mathcal{D}[st_{\text{main}}, a] \sqsupseteq a \quad a = \text{enter}^\# a_0$$

$$\mathcal{D}[st_f, a] \sqsupseteq \bigsqcup \{ a \mid a = \text{enter}^\# (\mathcal{D}[u, a']) \}$$

f procedure

\implies Dependences have become infinite !?

Partially Context-sensitive Analysis

Assume that

$$\mathbb{D} = A \times \mathbb{B}$$

where only \mathbb{B} should go into the context.

Partially Context-sensitive Analysis

Assume that

$$\mathbb{D} = \mathbb{A} \times \mathbb{B}$$

where only \mathbb{B} should go into the context.

- Analyze each procedure f for each b only !
- Accumulate occurring contributions to the rest !

$\mathcal{D}[v, b]$ \equiv value at program point v
within calls for contexts (a, b) , $a \in \mathbb{A}$.

Example Contexts

- First parameter,
- Parameters of certain types,
- Extra information such as held locks, opened files ...
- Abstraction of the call-stack

⇒ call-strings

The Constraint System:

$$\mathcal{D}[st_{\text{main}}, b] \sqsupseteq (a, b) \quad (a, b) = \text{enter}^\# (a_0, b_0)$$

$$\mathcal{D}[st_f, b] \sqsupseteq \sqcup \{(a, b) \mid (a, b) = \text{enter}^\# (\mathcal{D}[u, b'])\}$$

f procedure

The Constraint System:

$$\mathcal{D}[st_{\text{main}}, b] \sqsupseteq (a, b) \quad (a, b) = \text{enter}^\# (a_0, b_0)$$

$$\mathcal{D}[st_f, b] \sqsupseteq \sqcup \{(a, b) \mid (a, b) = \text{enter}^\# (\mathcal{D}[u, b'])\}$$

f procedure

$$\mathcal{D}[v, b] \sqsupseteq \llbracket s \rrbracket^\# (\mathcal{D}[u, b]) \quad (u, s, v) \text{ edge}$$

$$\mathcal{D}[v, b] \sqsupseteq \text{combine}^\# (\mathcal{D}[u, b]) (\mathcal{D}[rt_f, \pi_2(\text{enter}^\# (\mathcal{D}[u, b]))])$$

$(u, f(), v)$ call

Discussion

- Explicit tracking of reachability is now mandatory !
- Since the system is infinite, ordinary fixpoint iteration fails.
- Since potential dependences between unknowns are infinite, local solving fails as well ?

Discussion

- Explicit tracking of reachability is now mandatory !
 - Since the system is infinite, ordinary fixpoint iteration fails.
 - Since potential dependences between unknowns are infinite, local solving fails as well ?
- ⇒ The system does not support solving, but may serve as a **specification** to be proven correct w.r.t. the concrete semantics

Side-effecting Constraint Systems

Re-formulating Partial Context-Sensitivity

- Remove critical constraints for entry points of procedures !
- Generate the contributions to entry points of procedures when they occur ...

Idea

Constraints now take the form (x, f) where:

$$f : (\text{Unknowns} \rightarrow \mathbb{D}) \rightarrow (\text{Unknowns} \rightarrow \mathbb{D} \rightarrow \mathbf{void}) \rightarrow \mathbb{D}$$

Idea

Constraints now take the form (x, f) where:

$$f : (\text{Unknowns} \rightarrow \mathbb{D}) \rightarrow (\text{Unknowns} \rightarrow \mathbb{D} \rightarrow \text{void}) \rightarrow \mathbb{D}$$

When evaluating f **get set**,

- the argument **get** returns for every variable x , the current value **get** $x \in \mathbb{D}$;

Idea

Constraints now take the form (x, f) where:

$$f : (\text{Unknowns} \rightarrow \mathbb{D}) \rightarrow (\text{Unknowns} \rightarrow \mathbb{D} \rightarrow \text{void}) \rightarrow \mathbb{D}$$

When evaluating f **get set**,

- the argument **get** returns for every variable x , the current value **get** $x \in \mathbb{D}$;
- the argument **set** when called for $x \in \text{Unknowns}$ and $d \in \mathbb{D}$ allows to produce as a **side effect** the contribution d to x .

The Constraint System with Side Effects:

$$\mathcal{D}[st_{\text{main}}, b] \sqsupseteq (a, b) \quad (a, b) = \text{enter}^\# (a_0, b_0)$$

$$\mathcal{D}[v, b] \sqsupseteq [[s]]^\# (\text{get } [u, b]) \quad (u, s, v) \text{ edge}$$

The Constraint System with Side Effects:

$$\mathcal{D}[st_{\text{main}}, b] \sqsupseteq (a, b) \quad (a, b) = \text{enter}^\# (a_0, b_0)$$

$$\mathcal{D}[v, b] \sqsupseteq [[s]^\# (\text{get } [u, b]) \quad (u, s, v) \quad \text{edge}]$$

$$\begin{aligned} \mathcal{D}[v, b] \sqsupseteq & \text{let } (a', b') = \text{enter}^\# (\text{get } [u, b]) \text{ in} \\ & \text{let } () = \text{set } [st_f, b'] (a', b') \\ & \text{in } \text{combine}^\# (\text{get } [u, b]) (\text{get } [rt_f, b']) \\ & \quad (u, f (), v) \quad \text{call} \end{aligned}$$

Results

- The new constraint system is equivalent to the original one
 \implies every solution satisfies the specification !
- Each constraint depends only on few variables !

Results

- The new constraint system is equivalent to the original one

⇒⇒ every solution satisfies the specification !

- Each constraint depends only on few variables

⇒⇒

Local solving, e.g., by LR, enhanced with side effects, can be used as solver !

Partial-Context Sharir/Pnueli

- Sharir/Pnueli's algorithm is **operationally** different from Cousot's.
- SPA propagates changes **forward** through the control-flow graph.
- At a call, the graph is expanded by a new copy ...

Forward propagation, though, can be **simulated** by local solving and side effects ...

The SPA Constraint System

$$\begin{aligned} \mathcal{D}[st_{\text{main}}, b] &\sqsupseteq (a, b) & (a, b) &= \text{enter}^\# (a_0, b_0) \\ \mathcal{D}[u, b] &\sqsupseteq \text{let } () = \text{set } [v, b] (\llbracket s \rrbracket^\# (\text{get } [u, b])) \\ &\text{in } \perp & (u, s, v) &\text{ edge} \end{aligned}$$

The SPA Constraint System

$$\mathcal{D}[st_{\text{main}}, b] \sqsupseteq (a, b) \quad (a, b) = \text{enter}^\# (a_0, b_0)$$

$$\mathcal{D}[u, b] \sqsupseteq \text{let } () = \text{set } [v, b] (\llbracket s \rrbracket^\# (\text{get } [u, b])) \\ \text{in } \perp \quad (u, s, v) \text{ edge}$$

$$\mathcal{D}[u, b] \sqsupseteq \text{let } (a', b') = \text{enter}^\# (\text{get } [u, b]) \text{ in} \\ \text{let } () = \text{set } [st_f, b'] (a', b') \text{ in} \\ \text{let } () = \text{set } [v, v] (\text{combine}^\# (\text{get } [u, b]) (\pi_2(\text{get } [rt_f, b']))) \\ \text{in } \perp \\ (u, f(), v) \text{ call}$$

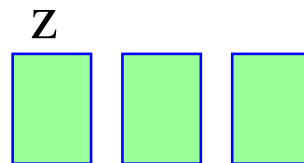
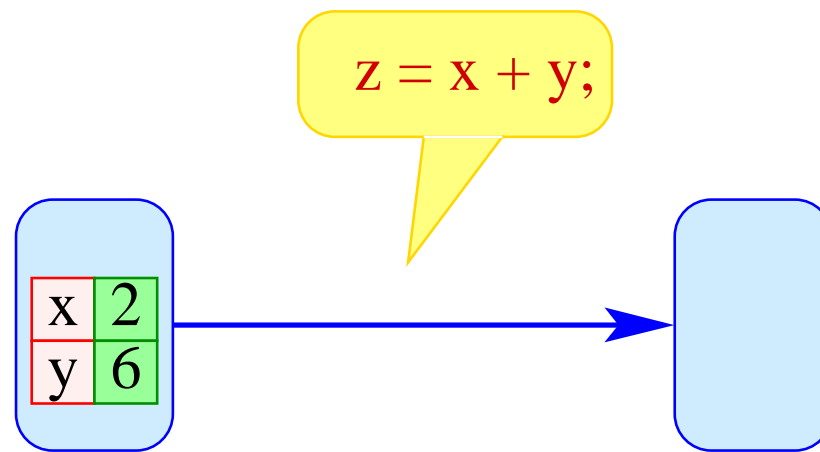
Discussion

- Local solving now starts with querying $\mathcal{D}[st_{\text{main}}, b]$, where $(_, b) = \text{enter}^\sharp(a_0, b_0)$.
- Besides in the initial constraint for $\mathcal{D}[st_{\text{main}}, b]$, each right-hand side just returns \perp .

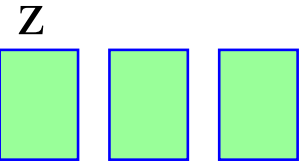
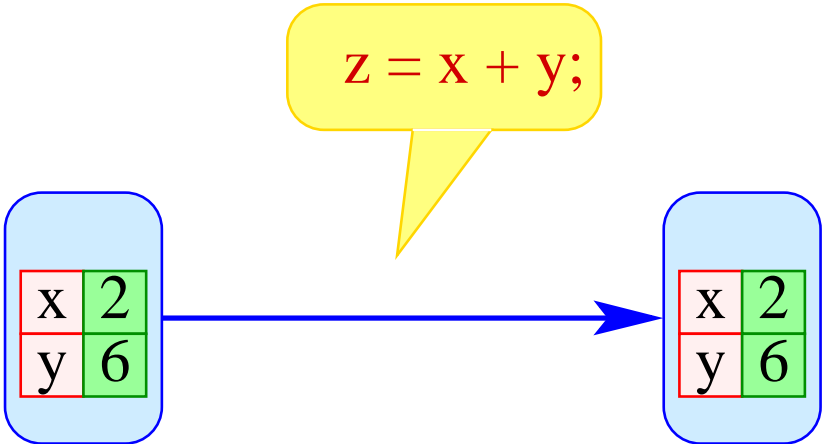
Discussion

- Local solving now starts with querying $\mathcal{D}[st_{\text{main}}, b]$, where $(_, b) = \text{enter}^\#(a_0, b_0)$.
- Besides in the initial constraint for $\mathcal{D}[st_{\text{main}}, b]$, each right-hand side just returns \perp .
- Subsequent program points receive their values through **side effects**
 \implies well-suited for the analysis of **binary code** !
- A quite non-standard formulation: Yet it enables to realize SPA by means of generic local solver :-)

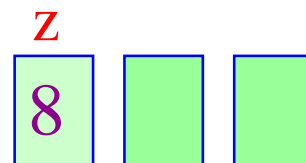
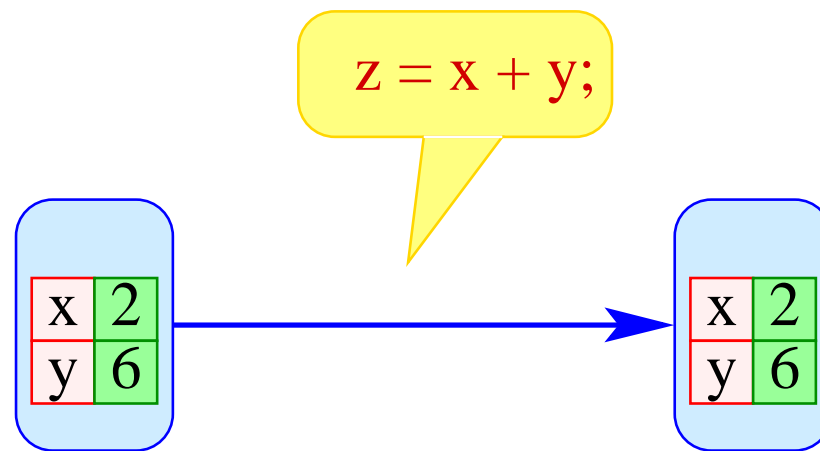
Flow-insensitive Invariants



Flow-insensitive Invariants



Flow-insensitive Invariants



Applications

- points-to information
- thread-safe information about globals

Applications

- points-to information
- thread-safe information about globals
 - ⇒ de-couples threads
 - ⇒ reduces thread analysis to sequential analysis !

Approach

- Only those contributions should be summarized which occur at **reachable** configurations.
- These become known only during fixpoint computation.

⇒ constraints with **side effects**

Experimental Evaluation

Goal

Accumulate for each global g , a set of definitely held mutexes when accessing g .

Experimental Evaluation

Goal

Accumulate for each global g , a set of definitely held mutexes when accessing g .

Prerequisites

- full constant propagation;
- may/must points-to analysis, ...
- context-sensitivity;
- side-effects to globals.

Benchmarks

aget multithreaded HTTP download accelerator

automount Autofs kernel-based automounter for Linux

ctrace C tracing library sample program

knot Knot web-server

pfscan Parallel file scanner

smtprc Smtpr relay checker

yplibind Linux NIS daemon

zfs-fuse ZFS filesystem for FUSE/Linux

Sizes: 1280 – 24097 LoC

Setup

- Cousot-style vs. SPA
- no/relevant/full context

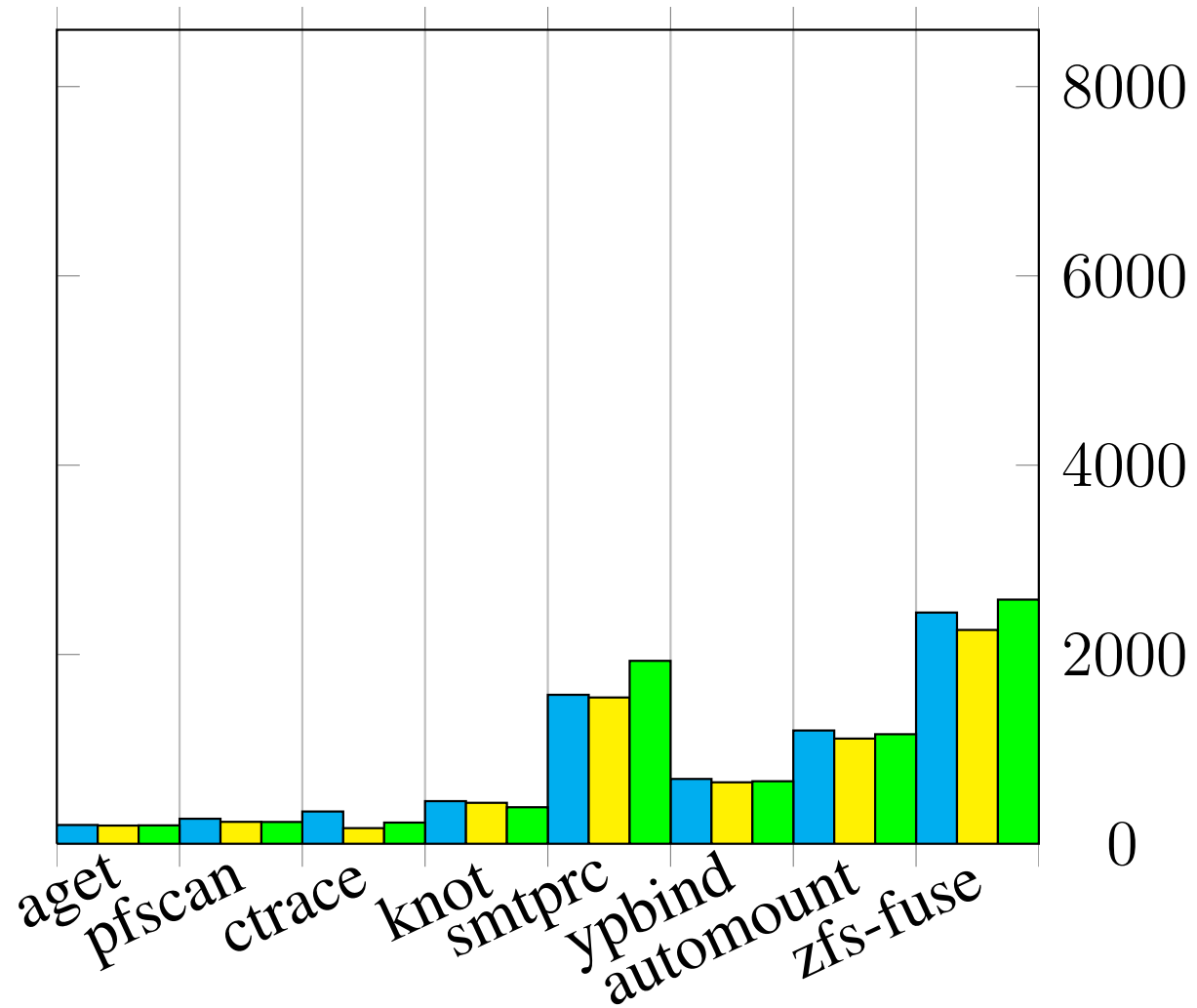
Setup

- Cousot-style vs. SPA
- no/relevant/full context

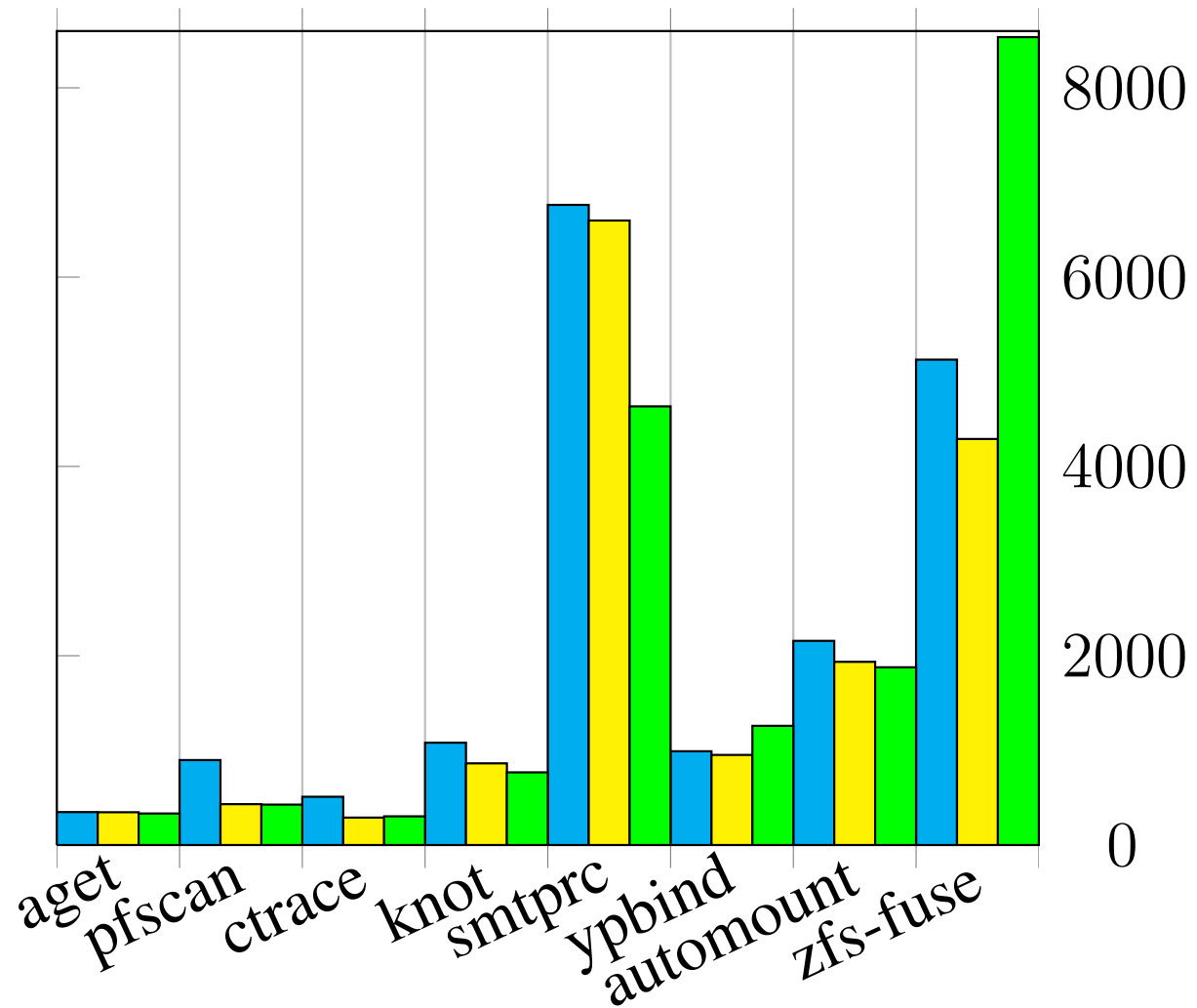
Results

- Overall, run-times are **decent** (3.5s for 24loc)
- For a **fair comparison**, we count evaluations of right-hand sides.

Cousot-style — # rhs



SPA — # rhs



Precision Results — Warnings

Name	Size	none	half	full
aget	1280	162	162	162
pfscan	1295	72	72	72
ctrace	1407	87	79	79
knot	2255	140	62	62
smtprc	5787	1068	636	636
ypbind	6596	251	244	244
automount	20624	505	480	480
zfs-fuse	24097	2319	2318	2318

Lessions learnt

- Static analyzers are complex software systems which themselves may contain bugs.
- A modular design allows to separate:
 - specification
 - solving

Lessions learnt

- Static analyzers are complex software systems which themselves may contain bugs.
- A modular design allows to separate:
 - specification
 - solving
- Side-effecting constraint systems are a convenient formalism to **freely combine**
 - flow-insensitive,
 - flow-sensitive analysis,
 - fine-tune the amount of tracked contexts.

Lessions learnt (cont.)

- Side-effecting constraint systems all can be solved by one dedicated local fixpoint algorithm.
- To ensure reliability of results, we provided:
 - a **correctness proof** of the algorithm
 - a **fixpoint checker** for analysis results
 - a **performance** checker for the overall system

Lessions learnt (cont.)

- The resulting implementation scales **reasonably well**.
- Some technical experiences:
 - **No** context is not necessarily faster than **some** context.
 - **More** context does not necessarily pay off.
 - Cousot is generally preferable to SPA.

goblint.in.tum.de