

Modular Heap Abstraction-Based Memory Leak Detection for Heap-Manipulating Programs

Longming Dong Ji Wang Liqian Chen

National University of Defense Technology, Changsha, China

05/12/2012 – APSEC 2012

Outline

- Motivation
- Field-sensitive heap abstraction
- Memory leak detection
- Implementation and experiments
- Conclusion

Motivation

Motivation

Dynamic allocated data structures

- examples: lists, trees, etc.
- widely used in practice
 - e.g., operating systems, device drivers, etc.
- **error-prone**
 - **memory leak**
 - degrade performance
 - cause memory-intensive or long-time running software to crash
 - dangling reference
 - double free
 - null pointer dereference
 - ...

Motivation

```
typedef struct list{
    int d;
    struct list* n;
}List;

void f(){
1: List* x=(List*) malloc(sizeof(List));
2: x→n =(List*) malloc(sizeof(List));
3: free(x);
} Memory leak on  $x \rightarrow n$ 
```

Field sensitive analysis of heap manipulating programs

- **problem:** high cost for exact memory layout
- **solution:** **abstraction** to make the problem tractable
 - proper abstraction according to properties to check
→ simplify the problem & be precise enough

Field-sensitive heap abstraction

Concrete heap state

Shape graph $\langle H, S \rangle$

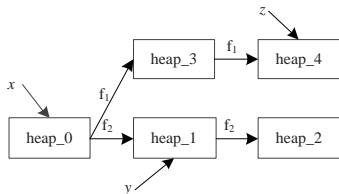
- the topological structure of heap memory can be described by a directed graph $H = \langle V, E \rangle$ where
 - V denotes the set of heap cells
 - $E : V \xrightarrow{F} V$ denotes the points-to relations between cells via their fields F
- $S : PVar \rightarrow V$ where $PVar$ denotes the set of pointer variables

$$PVar = \{x, y, z\}$$

$$V = \{heap_0, heap_1, heap_2, heap_3, heap_4\}$$

$$S = \{\langle x, heap_0 \rangle, \langle y, heap_1 \rangle, \langle z, heap_4 \rangle\}$$

$$E = \{heap_0 \xrightarrow{f_1} heap_3, heap_0 \xrightarrow{f_2} heap_1, heap_1 \xrightarrow{f_2} heap_2, heap_3 \xrightarrow{f_1} heap_4\}$$



It may cause high memory cost! \Rightarrow abstraction

Field-sensitive heap abstraction

An abstract domain of member-access distances $\langle D, +, - \rangle$

- elements: a set of abstract distances $D = \{0, 1, 2\}$
 - 0: the current cell **itself** (p)
 - 1: member-access with depth **1** ($p \rightarrow f$)
 - 2: member-access with depth **more than 1** ($p \rightarrow f \rightarrow^*$).
- operations: $+, -$ over D (defined in Table 1)

Table 1: Operations over D

$+$	0	1	2	$-$	0	1	2
0	{0}	{1}	{2}	0	\perp	\perp	\perp
1	{1}	{2}	{2}	1	{1}	{0}	\perp
2	{2}	{2}	{2}	2	{2}	{1,2}	\top

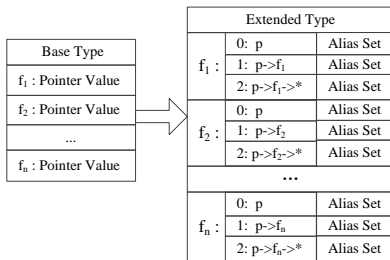
(\perp : the operator cannot be applied to these operands; \top : $\{0, 1, 2\}$)

Field-sensitive heap abstraction

An extended pointer structure of a pointer p

$$\tau_p^\# \triangleq \{f_1 : \langle D, 2^{PVar} \rangle; f_2 : \langle D, 2^{PVar} \rangle; \dots; f_n : \langle D, 2^{PVar} \rangle\}$$

- f_1, f_2, \dots, f_n denote the n **pointer fields** of the structure that p points to
- $D = \{0, 1, 2\}$ denotes the set of **abstract distances**
- 2^{PVar} denotes the **alias set** of accessing field f_i of pointer p with distance $d \in D$



Field-sensitive heap abstraction

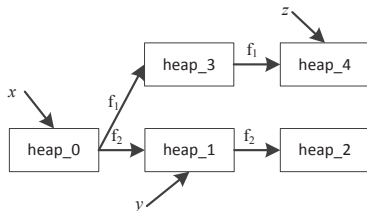
Example:

$$PVar = \{x, y, z\}$$

$$V = \{heap_0, heap_1, heap_2, heap_3, heap_4\}$$

$$S = \{\langle x, heap_0 \rangle, \langle y, heap_1 \rangle, \langle z, heap_4 \rangle\}$$

$$E = \{heap_0 \xrightarrow{f_1} heap_3, heap_0 \xrightarrow{f_2} heap_1, \\ heap_1 \xrightarrow{f_2} heap_2, heap_3 \xrightarrow{f_1} heap_4\}$$



$$\tau_x^\# : \{f_1 : \langle 0, \emptyset \rangle, \langle 1, \emptyset \rangle, \langle 2, \{z\} \rangle; \quad f_2 : \langle 0, \emptyset \rangle, \langle 1, \{y\} \rangle, \langle 2, \emptyset \rangle\}$$

$$\tau_y^\# : \{f_1 : \langle 0, \emptyset \rangle, \langle 1, \perp \rangle; \quad f_2 : \langle 0, \emptyset \rangle, \langle 1, \emptyset \rangle, \langle 2, \perp \rangle\}$$

$$\tau_z^\# : \{f_1 : \langle 0, \emptyset \rangle, \langle 1, \perp \rangle; \quad f_2 : \langle 0, \emptyset \rangle, \langle 1, \perp \rangle\}$$

Field-sensitive heap abstraction

Definition (Abstract Heap State)

The abstract heap state $\mathbb{S}^\#$ at each program point of a program HP consists of a set of extended structures of all pointer variables:

$$\mathbb{S}^\# = \{\tau_{p_i}^\# \mid p_i \in PVar(HP)\}$$

The number of abstract heap states: **finite**

$$\leq (fn \times 3 \times (2^{pn-1} + 1))^{pn}$$

- pn : the number of pointer variables
- fn : the maximum number of pointer fields

Field-sensitive heap abstraction

Alias bit-vector: using **bit-vector** to encode **alias set**

- maintain a variable ordering for all variables in $PVar$
- a alias bit-vector $\vec{r}_x^\# \in \{0, 1\}^{|PVar|}$ is defined as

$$\vec{r}_p^\#(f_m, d)[i] = 1 \Leftrightarrow v_i \text{ is an alias of accessing } f_m \text{ of } p \text{ with distance } d$$

Example: $PVar = \{x, y, z\}$ with variable ordering: $x \prec y \prec z$

$$\vec{r}_x^\# : \{f_1 : \langle 0, 0 \rangle, \langle 1, 0 \rangle, \langle 2, 001 \rangle; f_2 : \langle 0, 0 \rangle, \langle 1, 010 \rangle, \langle 2, 0 \rangle\}$$

$$\vec{r}_y^\# : \{f_1 : \langle 0, 0 \rangle, \langle 1, \perp \rangle; f_2 : \langle 0, 0 \rangle, \langle 1, 0 \rangle, \langle 2, \perp \rangle\}$$

$$\vec{r}_z^\# : \{f_1 : \langle 0, 0 \rangle, \langle 1, \perp \rangle; f_2 : \langle 0, 0 \rangle, \langle 1, \perp \rangle\}$$

Field-sensitive heap abstraction

Abstract heap state with a canonical form

Definition (Saturated abstract state)

An abstract state \mathbb{S}^\sharp is *saturated* iff it satisfies:

- 1 **Anti-reflexivity.** $\forall p_i. \vec{r}_{p_i}^\sharp(f_m, 0)[i] = 0.$
- 2 **Symmetry.** $\forall p_i, p_j. \vec{r}_{p_i}^\sharp(f_m, 0)[j] = 1 \rightarrow \vec{r}_{p_j}^\sharp(f_m, 0)[i] = 1.$
- 3 **Transitivity.** $\forall p_i, p_j, p_t. \vec{r}_{p_i}^\sharp(f_m, d_1)[j] = 1 \wedge \vec{r}_{p_j}^\sharp(f_m, d_2)[t] = 1$
 $\rightarrow \vec{r}_{p_i}^\sharp(f_m, d_1 + d_2)[t] = 1.$

Memory leak detection

Syntax of heap-manipulating programs

$p, q \in PVar$
 $f_1, f_2, \dots, f_i, \dots, f_m \in Fields$
 $AsgnStmt := p = null \mid p \rightarrow f_i = null \mid p = q \mid$
 $p = q \rightarrow f_i \mid p \rightarrow f_i = q \mid p = malloc() \mid p = free()$
 $SwitchStmt := switch\ e\ \{c_1 : n_1, \dots, c_j : c_j, \dots, c_k : n_k, \dots\}$
 $CallStmt := e = g(e_1, \dots, e_k)$
 $ReturnStmt := return\ e$
 $Stmt := AsgnStmt \mid SwitchStmt \mid CallStmt \mid ReturnStmt$
 $SequenceStmt := Stmt; Stmt$

Memory leak detection

Detecting memory leaks in assignments

- (a) check whether there are other pointers that can access the cell pointed to by the current pointer (p_u), such as **Rule 1, 3, 4, 6**;
- (b) check whether there are other pointers that points to the cell referenced by the pointer field of the current pointer ($p_u \rightarrow f_i$), such as **Rule 2, 5**;
- (c) check whether all pointer fields of the current pointer ($p_u \rightarrow f_i$) are *null* or pointed to by other pointers, like **Rule 7**.

Example

$$\begin{aligned}
 \textcircled{1} \quad [[p_u = \text{null}]](\mathbb{S}^\#) = & \{ \mathbb{S}^\#[\vec{r}_{p_u}^\#(f_m, d)[u] \leftarrow 0, \vec{r}_{p_u}^\#(f_m, d) \leftarrow \perp] \mid v \neq u \} \\
 & \text{if } \exists w \neq u \wedge l \in D. \vec{r}_{p_w}^\#(f_m, l)[u] \neq 0 \vee \vec{r}_{p_u}^\#(f_m, 0) = \perp \\
 & \{\text{memory_leak}\} \qquad \qquad \qquad \text{otherwise}
 \end{aligned}$$

Interprocedural memory leak detection

Big-step abstract semantics

$S_{Of}^\# = [[f(p_0, p_1, \dots, p_{k-1})]](S_{If}^\#)$, wherein $S_{Of}^\#$ is the postcondition after the running of the callee f under the precondition $S_{If}^\#$.

Procedural summary: $\langle S_{If}^\#, S_{Of}^\# \rangle$

- $S_{If}^\#$: abstract heap state over arguments and global pointers
- $S_{Of}^\#$: abstract heap state over return variable and global pointers

Example

```

typedef struct list{
    int d;
    struct list* n;
}List;
List* l;
List* f1(List* p){
1  if(p!=NULL){
2    l=p;
3  }
4  p=(List*) malloc(sizeof(List));
5  p->n=(List*)malloc(sizeof(List));
6  return p;
}

void g1(){
1  l=NULL;
2  List* x=(List*)malloc(sizeof(List));
3  List* y=f1(x);
4  free(y);
5  free(l);
}

void g2(){
1  l=NULL;
2  List* x=NULL;
3  List* z=f1(x);
4  free(z);
}
    
```

Table 2: procedural summary for $f1$

Precondition	Postcondition
$\overline{\mathcal{A}}_p^{\#} : \{n : \langle 0, \perp \rangle\}$	$\overline{\mathcal{A}}_{ret_{f1}}^{\#} : \{n : \langle 0, \emptyset \rangle, \langle 1, \emptyset \rangle, \langle 2, \perp \rangle\}$
$\overline{\mathcal{A}}_l^{\#} : \{n : \langle 0, \perp \rangle\}$	$\overline{\mathcal{A}}_l^{\#} : \{n : \langle 0, \perp \rangle\}$
$\overline{\mathcal{A}}_p^{\#} : \{n : \langle 0, \emptyset \rangle, \langle 1, \perp \rangle\}$	$\overline{\mathcal{A}}_{ret_{f1}}^{\#} : \{n : \langle 0, \emptyset \rangle, \langle 1, \emptyset \rangle, \langle 2, \perp \rangle\}$
$\overline{\mathcal{A}}_l^{\#} : \{n : \langle 0, \perp \rangle\}$	$\overline{\mathcal{A}}_l^{\#} : \{n : \langle 0, \emptyset \rangle, \langle 1, \perp \rangle\}$
$\overline{\mathcal{A}}_p^{\#} : \{n : \langle 0, \emptyset \rangle, \langle 1, \emptyset \rangle, \langle 2, \perp \rangle\}$	$\overline{\mathcal{A}}_{ret_{f1}}^{\#} : \{n : \langle 0, \emptyset \rangle, \langle 1, \emptyset \rangle, \langle 2, \perp \rangle\}$
$\overline{\mathcal{A}}_l^{\#} : \{n : \langle 0, \perp \rangle\}$	$\overline{\mathcal{A}}_l^{\#} : \{n : \langle 0, \emptyset \rangle, \langle 1, \emptyset \rangle, \langle 2, \perp \rangle\}$
$\overline{\mathcal{A}}_p^{\#} : \{n : \langle 0, \emptyset \rangle, \langle 1, \emptyset \rangle, \langle 2, \emptyset \rangle\}$	$\overline{\mathcal{A}}_{ret_{f1}}^{\#} : \{n : \langle 0, \emptyset \rangle, \langle 1, \emptyset \rangle, \langle 2, \perp \rangle\}$
$\overline{\mathcal{A}}_l^{\#} : \{n : \langle 0, \perp \rangle\}$	$\overline{\mathcal{A}}_l^{\#} : \{n : \langle 0, \emptyset \rangle, \langle 1, \emptyset \rangle, \langle 2, \emptyset \rangle\}$

Fixpoint iteration algorithm for analysis

- to compute the abstract heap state for each program point
- worklist-based
- always terminate (without need of widening)
 - maximum number of heap abstract states: $(fn \times 3 \times (2^{pn-1} + 1))^{pn}$

Implementation and experiments

Prototype

Heapcheck

- a **field** and **context** sensitive interprocedural memory leak detector
- based on Crystal (a program analysis system for C) ¹
- preprocessing process
 - slicing
 - transforming the input program into a SSA-like form by instrumenting new pointer variables

Pointer assignments	SSA-like assignments
$p \rightarrow f_i = q \rightarrow f_j$	$pt_0 = q \rightarrow f_j; p \rightarrow f_i = pt_0$
$p = p \rightarrow f_i$	$pt_1 = p \rightarrow f_i; p = pt_1$
$p \rightarrow f_i = \text{malloc}$	$pt_2 = \text{malloc}; p \rightarrow f_i = pt_2$
$p = q \rightarrow f_i \rightarrow f_j$	$pt_3 = q \rightarrow f_i; p = pt_3 \rightarrow f_j$
$p \rightarrow f_i = \text{free}()$	$pt_4 = p \rightarrow f_i; pt_4 = \text{free}()$

¹<https://www.cs.cornell.edu/projects/crystal/>

Experiments

Results on benchmark programs (memory leak)

Programs	Size (Kloc)	Preprocess time (Sec)	Analysis time (Sec)		Memory (MB)		Reported alarms (#fp/#total)
			Without sum.	Sum.-based	Without sum.	Sum.-based	
164.gzip	7.7	1.19	0.31	0.33	27	6	0
175.vpr	17	1.84	2.83	1.11	194	86.7	1/1
179.art	1.2	0.32	0.1	0.1	34.4	33	0
186.crafty	21.7	3.13	7.56	6.98	295	258	0
188.amp	13.2	1.88	1.22	0.21	135	60.2	0
300.twolf	19.9	3.05	7.38	4.31	442	195	0/3
176.gcc	210	8.35	106.62	61.04	4596	920	2/17
tar-1.12	11.7	1.08	18.98	9.09	239	178	0/5
openssh	58.3	20.55	2.61	1.44	186	144.3	2/14
openssl	36	8.47	0.46	0.44	73.5	40.7	6/11

- Real bugs found ($\#total - \#fp$)
 - 1 ignoring judging whether all the **sub-level pointer fields** are null when deallocating the heap cell pointed to by a pointer
 - 2 the heap cell pointed to by a **local** pointer variable is not deallocated at the return site

Experiments

Results on benchmark programs (memory leak)

Programs	Size (Kloc)	Preprocess time (Sec)	Analysis time (Sec)		Memory (MB)		Reported alarms (#fp/#total)
			Without sum.	Sum.-based	Without sum.	Sum.-based	
164.zip	7.7	1.19	0.31	0.33	27	6	0
175.vpr	17	1.84	2.83	1.11	194	86.7	1/1
179.art	1.2	0.32	0.1	0.1	34.4	33	0
186.crafty	21.7	3.13	7.56	6.98	295	258	0
188.amp	13.2	1.88	1.22	0.21	135	60.2	0
300.twolf	19.9	3.05	7.38	4.31	442	195	0/3
176.gcc	210	8.35	106.62	61.04	4596	920	2/17
tar-1.12	11.7	1.08	18.98	9.09	239	178	0/5
openssh	58.3	20.55	2.61	1.44	186	144.3	2/14
openssl	36	8.47	0.46	0.44	73.5	40.7	6/11

- Precision

- false positive rate: about 32% on openssh and openssl
 - compared with [B. Hackett, R. Rugina. Region-based shape analysis with tracked locations. POPL05]: about 64% (openssh: 16/26; openssl: 9/13)

Conclusion

Conclusion

Summary

- a **field** sensitive heap abstraction based on **member-access distances** and **alias bit-vector** domain
- a **field** and **context** sensitive interprocedural memory leak detection algorithm based on **summaries**
- experimental evaluations
 - our approach is scalable with satisfied precision in detecting **memory leaks** for large heap-manipulating programs

THANK YOU!

QUESTIONS?