CS 243 Homework 7 Solutions

Winter 2024

Due: March 13, 2024 at 11:59pm

Directions:

- Submit written answers via Gradescope.
- Submit the SMT file for Problem 6 via the separate Gradescope assignment.
- You may use up to two of your remaining late days for this assignment, for a late deadline of March 15, 2024 at 11:59pm.
- This is an individual assignment. You are allowed to discuss the homework with others, but you must write the solution individually. If you look up any material in the textbook or online, you should cite it appropriately.

Problem 1. For each of the following scenarios, what kind of memory management would you suggest (manual, reference counting, trace-based garbage collector)? If you recommend garbage collection, which garbage collector algorithm would you recommend? You must justify your answer to each question.

1. A long-running web server.

   Trace-based garbage collector. For a long-running web server, which involves memory allocation of many resources, it is necessary to use a garbage collector for automatic memory collection. Generational GC would be more preferable due to its efficiency. However, since it is not provided as an option in this context, we can also choose a simple trace-based garbage collector. Reference counting is considered troublesome in this scenario because it can lead to issues with cyclic references.

2. An embedded system for real-time signal processing.

   Manual. For real-time signal processing, we require low latency, and since embedded systems do not have abundant memory resources, it is feasible to manage memory manually.

3. Large-scale numerical scientific computation.

   Manual. For numerical scientific computation, which often involves large data structures/matrix operations, and where objects have a longer lifespan, there is no need to use a garbage collector for automatic memory management.

Problem 2. Garbage collection

The runtime for a fictional language uses a generational garbage collection with 2 generations. `x` is a root object. Initially, `a` is in the older generation and `b` is in the newer generation (nursery). Both are reachable from the stack. All objects on the stack are rooted. There are no other objects, and all fields are initially null.

You are given the list of operations performed by the mutator. For each operation, write the new state of the heap, distinguishing reachable and unreachable objects. Do not indicate collected (freed) objects.

The operation `minor-gc` indicates a GC of the newer generation, and `major-gc` indicates a GC of all generations. Objects that survive a minor GC are promoted to the older generation.

1. `c = new Object`

   reachable: x, a, b, c; unreachable: $\emptyset$

2. `a.f = c`

   reachable: x, a, b, c; unreachable: $\emptyset$

3. `a = null`

   reachable: x, b, c; unreachable: a

4. `d = new Object`

   reachable: x, b, c, d; unreachable: a

5. `c = null`

   reachable: x, b, d; unreachable: a, c

6. `minor-gc`

   reachable: x, b, d; unreachable: a, c

   The minor GC has no effect because a is in the older generation, and c is reachable from a

7. `x.f = d`

   reachable: x, b, d; unreachable: a, c

8. `b = null`

   reachable: x, d; unreachable: a, b, c

9. `minor-gc`

   reachable: x, d; unreachable: a, b, c

   Again no effect because b was promoted to the older generation in the previous minor GC.

10. `major-gc`

    reachable: x, d; unreachable: $\emptyset$

Problem 3. Let a be the number of live ranges in a program, b be the number of *merged* live ranges (as defined for register allocation), and c be the number of variables in the same program translated to static single-assignment (SSA) form. Express the relationship between a, b, and c using one or more inequalities. Explain your answer.

$b \leq a \leq c$. $b \leq a$ since there cannot be more *merged* live ranges than live ranges. $a \leq c$ since every live range corresponds to one variable definition, which corresponds to one SSA variable – but SSA can introduce more variables for the ϕ-nodes.

Problem 4. Recall that a formula `F` is satisfiable if there exists an interpretation I of the underlying variables and functions that makes the formula $\mathtt{F}^I$ true. Generally, SMT solvers are used to verify the satisfiability of an input formula `F` by doing the following in SMT notation (ignore declarations):

```
(assert F)
(check-sat)
```

If `(check-sat)` returns true, then `F` is satisfiable.

Given two input formulae `F1` and `F2`, explain how would you use SMT solvers to perform the following operations and provide a similar notation as the example above:

1. Verify the validity of `F1` (Recall that a formula `F` is valid if it is true under all interpretations)

   Formula `F1` is valid if $\neg$ `F1` is unsatisfiable. This can be checked using:

   `(assert (not F1))`

   `(check-sat)`

   If it returns unsat, then `F1` is valid.

2. Verify that for all models where `F1` is satisfied, `F2` is also satisfied.

   Here, we want to verify the validity of `F1` $\implies$ `F2`. Therefore, we need to check if `F1` $\wedge\neg$ `F2` is unsatisfiable. We basically want to find if there exists a case where `F1` is satisfied but `F2` is not. If there exists no such case, then our given statement holds.

   This can be checked using:

   `(assert (and F1 (not F2)))`

   `(check-sat)`

   If it returns unsat, then the given statement holds.

   Another acceptable solution:

   `(assert F1)`

   `(assert (not F2))`

   `(check-sat)`

   If it returns unsat, then the given statement holds.

3. For the following two cases, provide the SMT notation and explain the relationship between the two cases:

   (a) `F1` is satisfiable and `F2` is satisfiable

   (b) (`F1` $\wedge$ `F2`) is satisfiable

Case i:

```
(push)
(assert F1)
(check-sat)
(pop)
(push)
(assert F2)
(check-sat)
(pop)
```

Case ii:

```
(assert F1)
(assert F2)
(check-sat)
```

Alternative case ii:

```
(assert (and F1 F2))
(check-sat)
```

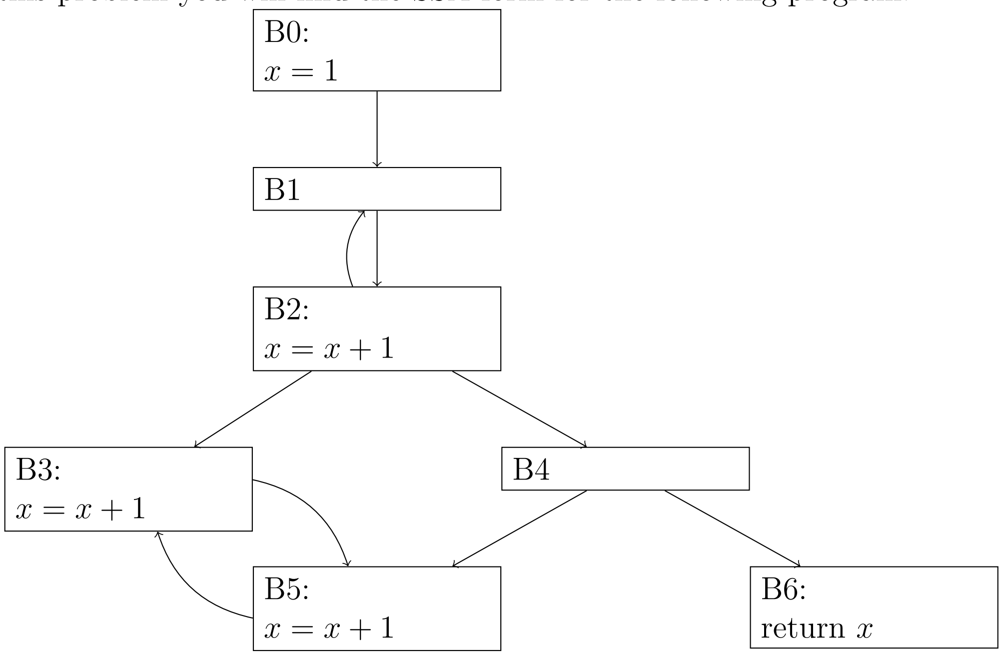
In case (ii), we check if $F1 \wedge F2$ is satisfiable, which means that both F1 and F2 are satisfiable at the same time (i.e. for the same interpretation). For case (i), we just check that they are individually satisfiable, but they may not be satisfiable together.

For example: lets consider a simple case:

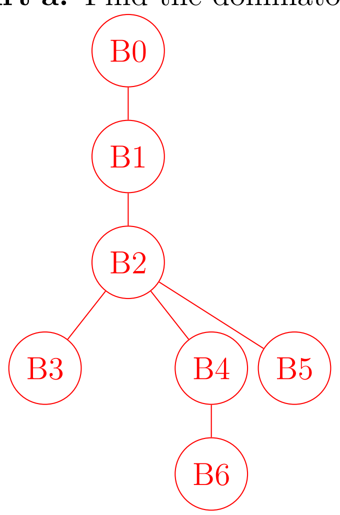
Given a variable x, let $F1 = x$, $F2 = \neg x$. Both F1 and F2 are individually satisfiable at $x = 1$ and $x = 0$ respectively, but $F1 \wedge F2 = 0$ is not satisfiable.

Problem 5. SSA Form.

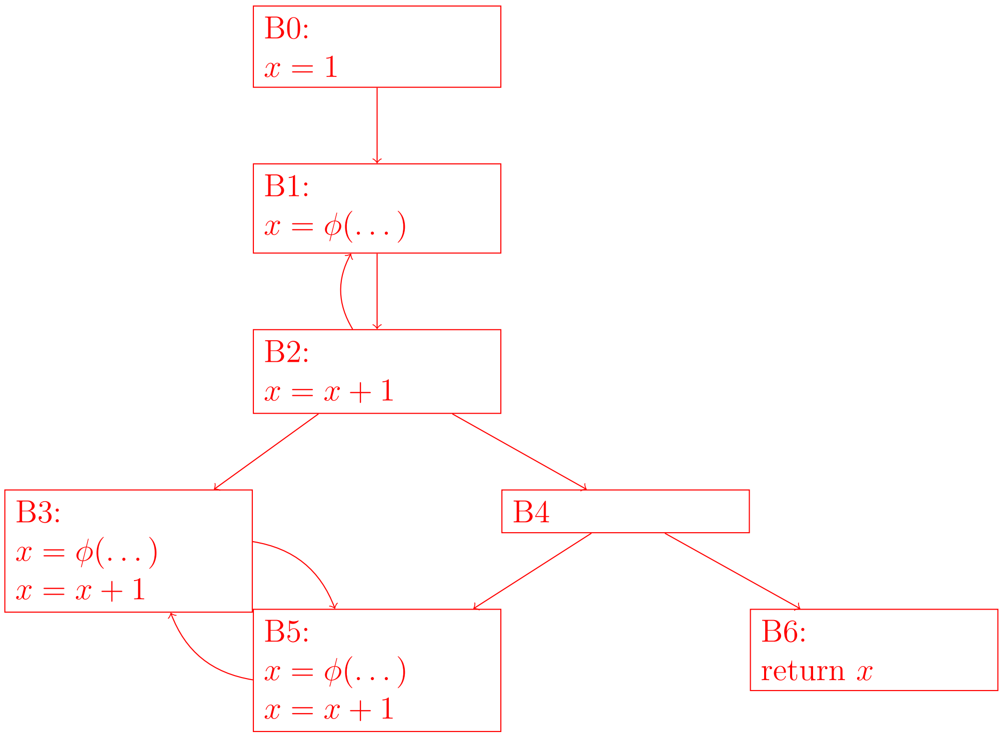
In this problem you will find the SSA form for the following program:



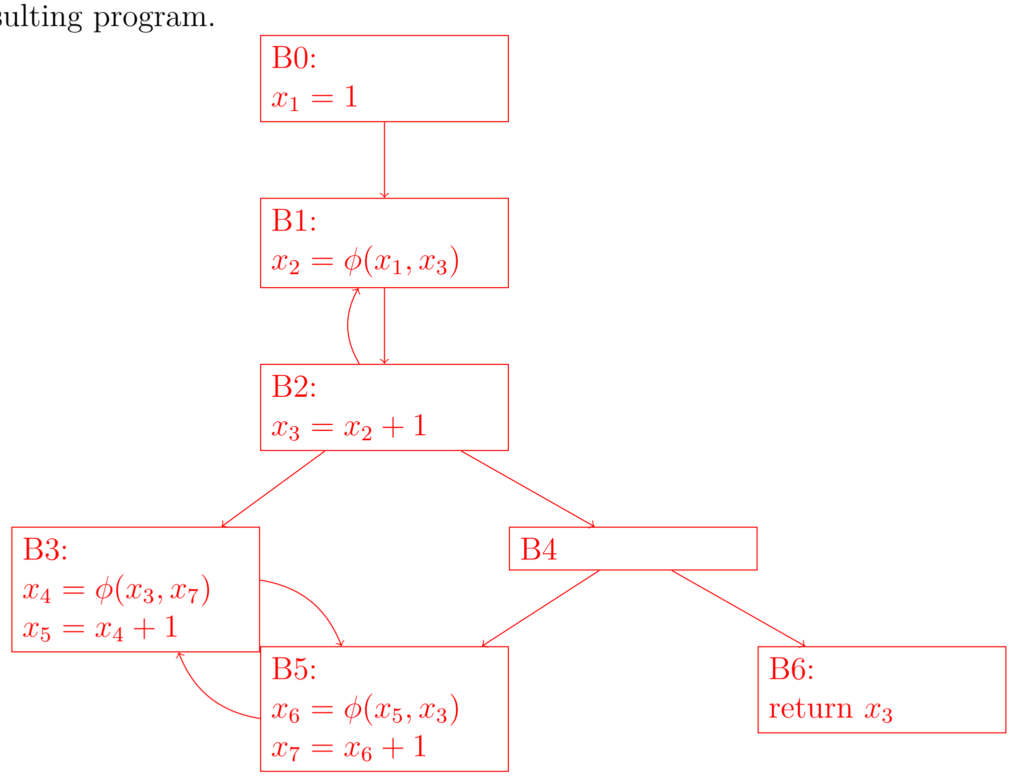
Part a. Find the dominator tree for the given program's CFG.



Part b. Insert ϕ-functions (without operands) and show the resulting program.



Part c. Rename variables by assigning unique numbers to each definition and show the resulting program.



Problem 6. Path-Sensitive Analysis with Satifisfiability Modulo Theories.

In this problem you will use an SMT solver to find test cases exhibiting a bug in the following C function:

```c
int func(int data[], int N, int x) {
    if (0 <= x && x < N && N > 0 && N < 5000) {
        int m = data[x + 1];     // access (buggy)
        if (m < 0 || m >= N) {
            m = 1;
        }
        int y = (x * m) % N;
        int i = data[y];         // access (fine)
        if (i < 0 || i >= N) {
            return 0;
        }
        k = data[data[y]];       // access (fine)
        int z = (x * m * i) % N;
        i = data[z];             // access (buggy)
        return (i + m) % N;
    } else {
        return 0;
    }
}
```

We are interested in checking whether the program can "crash" due to array out-of-bounds accesses. This means that the index with which we access the `data` array is either less than zero, or greater than `N-1`. To do so, first you will translate this function into an SMT formula. Then you will run an SMT solver on the formula, and interpret its output.

You may make the following assumptions about the program: `N` is at least 0, `data` is an array of length `N`, `int` refers to signed 32-bit integer, and the program terminates (crashes) as soon as the first out-of-bounds access happens.

Using SMT solvers. The research community has produced a large number of SMT solvers. For this assignment, you can use cvc5[1], a state-of-the-art SMT solver used in both research and industry. You can use cvc5's online interface to avoid the hassle of installation (link in footnote), or optionally you can download the solver binaries from its website to run locally.

To get started with writing SMT formulae, you can refer to an introductory guide[2]. The full SMT language specification is available at the SMT-LIB website[3]. For this assignment, you can use any feature defined in SMT-LIB, but the following features should be sufficient:

Commands: `assert`, `check-sat`, `declare-const`, `declare-func`, `get-model`, `pop`, `push`

[1] https://cvc5.github.io/; online interface at https://cvc5.github.io/app/

[2] https://microsoft.github.io/z3guide/docs/logic/intro

[3] https://smtlib.cs.uiowa.edu/papers/smt-lib-reference-v2.6-r2021-05-12.pdf

Sorts: Array, BitVec, Bool
Core functions: =, and, false, ite, not, or, true
Array functions: select, store
BitVec functions: bvadd, bvsub, bvmul, bvsdiv, bvsrem (% operator), bvsle/lt/ge/gt

Static single-assignment form. It is easier to translate the program into a SMT formula if we can assume that every variable is defined exactly once. One way to satisfy this assumption for imperative code is by transforming the program into *static single-assignment* (SSA) form. To do this, first assign every variable definition a unique suffix. Then, at each join point in the control-flow graph (e.g., after branching), introduce new definitions for the variables that are defined on either path using a ϕ-node.

See the following example which transforms the imperative code on the left into SSA form on the right (here, instead of normal ϕ-functions, we have used the actual branch condition so that the SMT solver can reason about which branch was taken):

```
if (i < next) {
    if (data[i] == cookie)
        i++;
    else
        process(data[i]);

    i++;

    if (i < next) {
        if (data[i] == cookie)
            i++;
        else
            process(data[i]);

        i++;
    }

}
```

```
if (φ1 = (i0 < next)) {
    if (φ2 = (data[i0] == cookie))
        i1 = i0 + 1;
    else
        process(data[i0]);
    i2 = φ2 ? i1 : i0;
    i3 = i2 + 1;

    if (φ3 = (i3 < next)) {
        if (φ4 = (data[i3] == cookie))
            i4 = i3 + 1;
        else
            process(data[i3]);
        i5 = φ4 ? i4 : i3;
        i6 = i5 + 1;
    }
    i7 = φ3 ? i6 : i3;
}
i8 = φ1 ? i7 : i0;
```

Follow the following steps to complete this problem.

1. **Rewrite the program in SSA form.** Notice that at different points within our function, variables v and z may refer to different definitions. Transform the function into SSA form as illustrated earlier.
2. **Translation to SMT.** The second step is to translate the program in SSA form into an SMT formula. Start your formula with the following lines:

```
(set-logic ALL)
(set-option :produce-models true)
(set-option :incremental true)
```

An assignment $x_3 = e$ becomes an assertion `(assert (= x3 E))` in SMT, where `E` is the translation of e. You need to translate `int` operations at the C level into bit vector operations at the SMT level. (Do not use the SMT-LIB `Int` sort, which models *unbounded* integers.) For example, the translation of $x = y + 1$ is:

```
(declare-const x (_ BitVec 32))
(declare-const y (_ BitVec 32))
(assert (= x (bvadd y #x00000001)))
(check-sat)
(get-model)
```

Upon this query, an SMT solver could respond with:

```
sat
(
(define-fun x () (_ BitVec 32) #b00000000000000000000000000000001)
(define-fun y () (_ BitVec 32) #b00000000000000000000000000000000)
)
```

meaning that the SMT formula is satisfiable with model $x = 1, y = 0$. Note that the variables x and y are given as functions of no arguments (which must be constants because there are no side effects), and the constants themselves are given in binary or hexadecimal.
To translate arrays, use variables of the sort `(Array (_ BitVec 32) (_ BitVec 32))`. Array dereferences like `data[i]` become `(select data i)` when translating a read, and `(store data i x)` when translating a write. Note that `(store data i x)` returns a new array, whose i^{th} element is now equal to x, and does not modify the original array.
To translate ϕ-nodes, you must use a logical expression that captures the condition under which the ϕ-node is evaluated. For example, given the following code in SSA form:

```
if (c)
    x1 = ...;
else
    x2 = ...;
x3 = φ(x1, x2);    // equivalent to x3 = c ? x1 : x2;
```

the translation of `x3` is `(ite c x1 x2)`. `ite` is short for if–then–else, and evaluates to the second or third argument based on the value of the first.

3. **Bounds checks.** The final step is to add an assertion to check each of the array accesses. You need to check that the signed value of the index is in bounds. Further, not all accesses are accessible on all paths, so you need to guard the assertion for a particular access. The assertion should express the execution reaches this access, and it is out of bounds. Note that this will possibly constrain some path variables if the access is nested inside an `if`-statement, for example. For each access, you should use the sequence `(push) (assert` C`) (check-sat) (pop)`, where C is the check for that access. Push/pop allows us to add C to our set of assertions temporarily, check satisfiability, and then remove it to add a different C. (If you added all the assertions together without push/pop, you would find a path which crash all points simultaneously, which is impossible.)

4. **Interpretation.** Once you find one or more bugs, add `(get-model)` after `(check-sat)` within the push/pop sequence for each satisfiable assertion, to print the model found for that bug. In writing, interpret the results. Does the result indicate a crash can occur on some concrete path? If not, does this mean there can be no crash for this access? If there is a crash, translate the model into a concrete (`x`, `data`) input pair in C syntax that would crash `func()` at the corresponding access.
   Hint: what happens if the program has more than one bug, and how should the model be interpreted in that case?
5. **Find the smallest value of N that can cause out-of-bounds accesses in line 14.** Do this by asserting the value of N in addition to asserting C described in part 3. Hint: use binary search to find the minimal value of N, as it can be quite large. $N = 1292$.
6. **Submission.** In the homework write-up, include the output of running the SMT solver on your input (`sat` or `unsat`, and, if `sat`, the produced model). Explain your interpretation of the results (part 4) and include answers to the questions in part 5.
   Additionally, submit your SMT-LIB input file to the separate Gradescope assignment.
   My SMT-LIB file:

```
(set-logic ALL)
(set-option :produce-models true)
(set-option :incremental true)
(declare-const data (Array (_ BitVec 32) (_ BitVec 32)))
(declare-const x (_ BitVec 32))
(declare-const N (_ BitVec 32))
(declare-const m (_ BitVec 32))
(declare-const y (_ BitVec 32))
(declare-const i (_ BitVec 32))
(declare-const z (_ BitVec 32))

(define-fun in_bounds ((ii (_ BitVec 32))) Bool
            (and (bvsle #x00000000 ii)
                 (bvslt ii N)))

(push)
(assert (bvsle #x00000000 x))
(assert (bvslt x N))
(assert (bvslt #x00000000 N))
(assert (bvslt N #x00001388))

(push)
(echo "line 3")
(assert (not (in_bounds (bvadd #x00000001 x))))
(check-sat)
(get-model)
(pop)
```

```
(assert (in_bounds (bvadd #x00000001 x)))
(assert (= m (select data x)))
(assert (in_bounds m))
(assert (= y (bvsrem (bvmul x m) N)))

(push)
(echo "line 8")
(assert (not (in_bounds y)))
(check-sat)
(pop)

(assert (in_bounds y))
(assert (= i (select data y)))
(assert (in_bounds i))

(push)
(echo "line 12")
(assert (not (in_bounds (select data y))))
(check-sat)
(pop)

(assert (in_bounds (select data y)))
(assert (= k (select data (select data y))))
(assert (= z (bvsrem (bvmul (bvmul x m) i) N)))

(push)
(echo "line 14")
(assert (not (in_bounds z)))
(check-sat)
(get-model)
(pop)

(pop)
```