C++ Programming

Lecture 12
Software Engineering Group

Philipp D. Schubert
Contents

1. Cyber attacks
2. Program analysis
3. Do we need analysis?
4. Static analysis
5. Designing code analysis (DECA)
6. Job offer
Usual Cyber-Attack

- Find vulnerability or place backdoor
- Inject malicious code
- Circumvent detection
- Surveil system, infect more machines if required
- Execute payload
Who is responsible?

According to a recent study by the DHS, more than 90% of all current cyber attacks succeed because of vulnerabilities in the application layer!
How can we reduce the number of vulnerabilities?

- Write your code carefully
- Test your code excessively
- Use dynamic analysis
- Use static analysis
- Use manual reviews
Problem: Rice’s theorem

- Program analysis is mathematically provable hard
- All non-trivial semantic properties of programs are undecidable!
- A semantic property is one about the behavior
  - An example
    - Does a program terminate for all inputs?
  - A property is non-trivial if
    - It is neither true for every program
    - Nor for no program
    - That are quite a lot of properties!
  - Contains proof sketch as well as complete proof
- We have to use an over-approximation then!
Analyzing programs

- Static “white box” analysis
- Dynamic “black box” analysis
Static versus dynamic analysis

**Static analysis**
- Retrieve information about a program without executing it
- Analysis performed on source code or intermediate code
- Over-approximation of the program behavior

**Dynamic analysis**
- Retrieve information about a program behavior by executing it
- Execute on real or virtual processor
- Must be executed with sufficient test inputs (code coverage)
- Discover a set of possible behaviors
- Uncertainty principle, make sure code instrumentalization does not cause side-effects
- Inadequate testing leads to catastrophic failures → runtime error in Ariane 5 rocket

**Program understanding**
- Done by humans: program comprehension, code review, software walkthroughs

Uses of static analysis

- Bug finding, vulnerability detection, compiler optimization
- Popular in industry for companies producing their own software
  - A. Part of nightly builds
  - B. At the end of the development phase

Relative cost to fix, based on time of detection

Source: National Institute of Standards and Technology
Static analysis

I SEE BUGS EVERYWHERE
C and C++

- Most common compilers
  - GCC
  - Clang/ LLVM
  - MSVC

- “Toward Understanding Compiler Bugs in GCC and LLVM”
  - Most buggy component?
    - C++
      - In both compilers
      - 20% of the total bugs
    - Bugs can be found in programs << 45 lines of code

100 bugs in Open Source C/C++ projects [https://www.viva64.com/en/a/0079/]

- Andrey Karpov (MS MVP) & Evgeniy Ryzhkov
- Found bugs in projects using static analysis
- Projects include
  - Apache HTTP Sever
  - Chromium
  - Cmake
  - MySQL
  - Qt
  - TortoiseSVN
  - ...
Passing the wrong size [https://www.viva64.com/en/a/0079/](https://www.viva64.com/en/a/0079/)

- Wolfenstein 3D project

```c
void CG_RegisterItemVisuals( int itemNum ) {
    ...
    itemInfo_t *itemInfo;
    ...
    memset( itemInfo, 0, sizeof( &itemInfo ) );
    ...
}
```
Is sizeof that hard? [https://www.viva64.com/en/a/0079/](https://www.viva64.com/en/a/0079/)

- ReactOS project

```c
static const PCHAR Nv11Board = "NV11 (GeForce2) Board";
static const PCHAR Nv11Chip = "Chip Rev B2";
static const PCHAR Nv11Vendor = "NVidia Corporation";

BOOLEAN
IsVesaBiosOk(...) {
    ...
    if (!strncmp(Vendor, Nv11Vendor, sizeof(Nv11Vendor))) &&
        !(strncmp(Product, Nv11Board, sizeof(Nv11Board))) &&
        !(strncmp(Revision, Nv11Chip, sizeof(Nv11Chip))) &&
        (OemRevision == 0x311))
    ...
}
```

- At least the original intention was good
- Do not using strcmp
If bias shall be accessed just say so [https://www.viva64.com/en/a/0079/](https://www.viva64.com/en/a/0079/)

- VirtualDub project

```c
struct ConvoluteFilterData {
    long m[9];
    long bias;
    void *dyna_func;
    DWORD dyna_size;
    DWORD dyna_old_protect;
    BOOL fClip;
};

static unsigned long __fastcall do_conv(
    unsigned long *data,
    const ConvoluteFilterData *cfd,
    long sflags, long pit)
{
    long rt0=cfd->m[9], gt0=cfd->m[9], bt0=cfd->m[9];
    ...
}
```
This is not how `clear()` works [https://www.viva64.com/en/a/0079/](https://www.viva64.com/en/a/0079/)

- TortoiseSVN project

```cpp
CMailMsg& CMailMsg::SetFrom(string sAddress,
                             string sName)
{
    if (initIfNeeded())
    {
        // only one sender allowed
        if (m_from.size())
            m_from.empty();
        m_from.push_back(TStrStrPair(sAddress, sName));
    }
    return *this;
}
```

- Learn the STL!
This is not how `clear()` works. See [cppreference.com](https://www.viva64.com/en/a/0079/) for more information.

### `std::vector::empty`

```cpp
bool empty() const;  // (until C++11)
bool empty() const noexcept;  // (since C++11)
```

Checks if the container has no elements, i.e. whether `begin() == end()`.

**Parameters**

(None)

**Return value**

- `true` if the container is empty, `false` otherwise

**Complexity**

Constant.
Check twice to be sure  https://www.viva64.com/en/a/0079/

- Notepad++ project

```cpp
bool _isPointXValid;
bool _isPointYValid;
...
bool isPointValid() {
    return _isPointXValid && _isPointYValid;
};
```
Better check if `unsigned` works [https://www.viva64.com/en/a/0079/](https://www.viva64.com/en/a/0079/)

- `size_t` is usually something like `unsigned long long`.

- Qt project

```cpp
bool equals( class1* val1, class2* val2 ) const{
    ...
    size_t size = val1->size();
    ...
    while ( --size >= 0 ){
        if ( !comp(*itr1,*itr2) )
            return false;
        itr1++;
        itr2++;
    }
    ...}
```
Checking for `nullptr` is usually good [https://www.viva64.com/en/a/0079/](https://www.viva64.com/en/a/0079/)

- Ultimate TCP/IP project

```c
char *CUT_CramMd5::GetClientResponse(LPCSTR ServerChallenge)
{
    ...
    if (m_szPassword != NULL)
    {
        ..
        if (m_szPassword != '\0')
        {
            ...
        }
    }
    else
    {
        // Check `char` for null two times?
        // Probably:
        // Password not empty was meant
        *m_szPassword != '\0';
    }
}
```
Dereferencing `nullptr` [https://www.viva64.com/en/a/0079/](https://www.viva64.com/en/a/0079/)

- Chromium project

```cpp
bool ChromeFrameNPAPI::Invoke(...) {
    ChromeFrameNPAPI* plugin_instance =
        ChromeFrameInstanceFromNPObjec(header);
    if (!plugin_instance &&
        (plugin_instance->automation_client_.get()))
        return false;
    ...
}
```
Again sizeof [https://www.viva64.com/en/a/0079/](https://www.viva64.com/en/a/0079/)

- SMTP client with SSL/TLS project

```c
void MD5::finalize () {
    ...
    uint1 buffer[64];
    ...
    // Zeroize sensitive information
    memset (buffer, 0, sizeof(*buffer));
    ...
}
```
Forgot to dereference [https://www.viva64.com/en/a/0079/](https://www.viva64.com/en/a/0079/)

- Miranda IM project
- String-end determined incorrectly

```c
static char * _skipblank(char * str)
{
    char * endstr=str+strlen(str);
    while ((*str==' ' || *str=='\t') && str!='\0') str++;
    while ((*endstr==' ' || *endstr=='\t') &&
           endstr!='\0' && endstr<str)
        endstr--;
    ...
}
```
Checking things twice [https://www.viva64.com/en/a/0079/]

- Intel AMT SDK project

```c
static void
wsman_set_subscribe_options(...) {
    ...
    if (options->delivery_certificateThumbprint ||
        options->delivery_password ||
        options->delivery_password) {
        ...
    }
}
```

- Does not check the presence of a username
Characteristics of C and C++

- What comes in my mind:
  - Old
  - Very powerful
  - Expressive
  - Complex
  - Expert friendly
    - A million rules to remember!
- In reality (oftentimes):
  - Un-concentrated developers
    - Call wrong function
    - Forgot to dereference pointer
    - `sizeof` operator
- "What is a protected abstract virtual base pure virtual private destructor and when was the last time you needed one?" Tom Cargill (1990)

```cpp
class Base {
private:
    virtual ~Base() = 0;
};

class Derived : protected virtual Base {
    ...
};
```
Characteristics of C and C++

- What comes in my mind:
  - Old
  - Very powerful
  - Expressive
  - Complex
  - Expert friendly
    - A million rules to remember!

- In reality (oftentimes):
  - Un-concentrated developers
    - Call wrong function
    - Forgot to dereference pointer
    - `sizeof` operator

  “What is a protected abstract virtual base pure virtual private destructor and when was the last time you needed one?” Tom Cargill (1990)
C and C++ are basically everywhere!

- When performance matters
  - Operating systems
  - Embedded systems
  - Simulations
  - Real-time systems
  - Browsers
  - and so on …
Use static analysis

- … to reduce the number of coding mistakes!
- How?
- Recipe
  1. Become aware of a bug (or mistake)
  2. Understand the bug
  3. Write a static analysis
  4. Run the analysis on code
  5. Obtain findings of potential bugs
  6. If it is a bug → fix it!
- Usually integrated within a build pipeline
Data-flow analysis

- Intra-procedural analysis

Does the property $\varphi$ hold at statement $s$?
Some properties & corresponding analysis

- Is this variable still used later on?
  - Live-variables analysis
- Can this code ever execute?
  - Dead-code analysis
- Can this pointer ever be null?
  - Nullness analysis
- Is this file handle ever closed?
  - Typestate analysis
- Can sensitive data leak?
  - Taint analysis
- There are many more
### Data-flow analysis: workflow

- **Workflow in static analysis:**
  1. Parse method (as source code, bytecode or some other intermediate representation)
  2. Convert into control-flow graph
  3. Perform an analysis on the CFG
  4. Find interesting properties

```java
int x = 1;
p1int(x);
if (z > 0) {
    x = 2;
}
p1int(x);
```

```
int x = 1;
p1int(x);
if (z > 0)
    x = 2;
p1int(x);
```
Intermediate representations

- Analysis usually performed on an intermediate representation (IR)
  - Simpler than source language
  - Just a few op codes
  - Uses jumps (goto) to represent loops
Compiler infrastructure

Provide many helpful mechanisms to write:

- Compiler optimizations
- Static analyses

Intermediate representation LLVM IR

- Abstracted from concrete input language

Pros

- No nesting
- Looping/branches through jumps (goto)
- Simple basic operations
- 3 address code

Cons

- No direct mapping from LLVM IR back to source
Control-flow graph

```c
int y = x;
if (p) x = y;
if (!p) z = 2;
b = y;
```

- Depending on complexity of `p`:
  - Mutual exclusiveness cannot be inferred
- CFGs are conservative:
  - If control may flow from stmt A to stmt B then there is an edge from A to B
    - Opposite is not true!
  - Problem is undecidable
    - Over-approximation
  - Real CFGs contain exceptional edges as well
    - Otherwise unsound
Perform an analysis

- Analysis: What values are printed?
- Reaching definition analysis
  ```java
  int x = 1;
  print(x);
  if (z > 0) {
    x = 2;
  }
  print(x);
  ```
- Use data flow analysis

```java
int x = 1;
print(x);
if (z > 0) {
  x = 2;
}
print(x);
```
A more advanced example

- Which assignments are unnecessary?
- Live variables analysis

```c
int z = /* some value */;
int x = 1;
int y = 2;
if (z > 0) {
    y = z;
    if (z > 1) {
        z = 7;
    }
}
print(y);
```

- Findings
  - Assignment to x and z can be eliminated

[Example borrowed from SEPL Goethe University Frankfurt]
That easy? Loops

- Reaching definitions revisited
  ```c
  int x = 1;
  while (/* some property */) {
      ++x;
  }
  print(x);
  ```

- Problem!
  - This does not terminate
  - Number of iterations must be bound
  - We have a mathematical theory
    - Monotone framework
Monotone framework

1. Analysis direction (forward or backward)
2. Analysis domain (lattice)
3. Effect of statements on information (flow functions)
4. Initial values (values of the lattice)
5. Merge information (binary operator on sets of lattice values)

- Algorithm fits on one slide
  - It is not that hard, right?
Monotone framework

1. Analysis direction (forward or backward)
2. Analysis domain (lattice)
3. Effect of statements on information (flow functions)
4. Initial values (values of the lattice)
5. Merge information (binary operator on sets of lattice values)

Flow functions

- Concrete
- Static

Analysis algorithm

- Algorithm fits on one slide
- It isn’t that hard, right?

**INPUT:** An instance of a Monotone Framework: 
\((L, F, E, \iota, f, \cdot)\)

**OUTPUT:** \(MFP_0, MFP_*\)

**METHOD:**

**Step 1:** Initialisation (of \(W\) and Analysis)
\[W := \text{nil};\]
\[\text{for all }(\ell, \ell') \text{ in } F \text{ do}\]
\[W := \text{cons}((\ell, \ell'), W);\]
\[\text{for all } \ell \text{ in } F \text{ or } E \text{ do}\]
\[\text{if } \ell \in E \text{ then } \text{Analysis}[\ell] := \iota\]
\[\text{else } \text{Analysis}[\ell] := \bot_L;\]

**Step 2:** Iteration (updating \(W\) and Analysis)
\[\text{while } W \neq \text{nil do}\]
\[\ell := \text{fst(head}(W)); \ell' = \text{snd(head}(W));\]
\[W := \text{tail}(W);\]
\[\text{if } f_\ell(\text{Analysis}[\ell]) \not\subseteq \text{Analysis}[\ell'] \text{ then}\]
\[\text{Analysis}[\ell'] := \text{Analysis}[\ell'] \sqcup f_\ell(\text{Analysis}[\ell]);\]
\[\text{for all } \ell'' \text{ with } (\ell', \ell'') \text{ in } F \text{ do}\]
\[W := \text{cons}((\ell', \ell''), W);\]

**Step 3:** Presenting the result (\(MFP_0\) and \(MFP_*\))
\[\text{for all } \ell \text{ in } F \text{ or } E \text{ do}\]
\[MFP_0(\ell) := \text{Analysis}[\ell];\]
\[MFP_*(\ell) := f_\ell(\text{Analysis}[\ell]);\]

**Table 2.8:** Algorithm for solving data flow equations.
Intra- versus inter-procedural analysis: A call graph

```c
void foo() {
    bar();
}

void bar() {}

int main() {
    foo();
    bar();
    return 0;
}
```

- It is not always that easy!
  - Function pointers
  - Virtual dispatch
Calling contexts

- Caution
  - Monotone framework cannot be used directly for inter-procedural analysis: too imprecise!
  - We have to consider calling contexts
Problems with usable static analysis

- Precision versus performance
- Static analysis often does not scale well
  - Massive runtimes & memory consumption
- Sophisticated solutions
  - Often require (really) complex algorithms
- Abstractions making it even more difficult
- Undecidable problems
Pros and cons of automated static analysis

- Fast(er than manual audit)
- Cheaper than manual audits
- Finds almost as much as manual audits
  - Efficient for obvious vulnerabilities
  - Detects useful hints for more complex programs
- Only requires basic knowledge of security to review warnings

- Less flexible than human analysts
  - Difficulties staging complex attacks
  - Cannot interpret human language
- Yields too many results
  - False positives
  - Irrelevant results
- Tough to implement
Why do we have little tooling?

- C and C++ are old
  - We know how to write compilers!
- Static analysis is old
  - We know how it works!
- Why are there so little static analysis tools?
  - Sophisticated static analysis is hard
  - Many problems are undecidable
    - Fight the complexity dragon
  - We have to convince companies and developers to use it!
Designing code analysis for large-scale software systems (DECA)

- Summer semester 2018
- Lead by Eric Bodden
- Contents
  - Intra-procedural data-flow analysis
  - Call-graph construction algorithms
  - Context-insensitive inter-procedural data-flow analysis
  - Context-sensitivity using the call-strings approach
  - Value-based context
  - Context-sensitivity using the functional approach
  - Efficiently solving distributed problems in the IFDS and IDE frameworks
  - Current challenges in inter-procedural static program analysis
  - Applications to software security
Overview and more examples

- SEPL Goethe University, Frankfurt am Main
- Lecture “Foundations of Programming Languages”, Christoph Reichenbach
- Have a look at the following youtube playlist (one lecture unit)
  - [https://www.youtube.com/watch?v=sxiFwiCgoVo&list=PLgJZZQPiH1mHIZAyIF1baZbMplzxXn90o](https://www.youtube.com/watch?v=sxiFwiCgoVo&list=PLgJZZQPiH1mHIZAyIF1baZbMplzxXn90o)
- Optimizations and static analysis
  - What?
  - Why?
  - How?
It is all about creativity

- You cannot define custom operators
- But how about that?

```cpp
#include <iostream>
using namespace std;
enum _multiplication { times, multiply, mult, $cool$ };

// tiny int wrapper to trick the type system
struct _int {
    int i;
    _int(int i) : i(i) {}  // implicit conversion operator
    operator int() const { return i; }
};

_int operator/(int i, _multiplication m) {
    return _int(i);
}

int operator/(_int j, int k) { return j * k; }

int main() {
    int a = 12 /times/ 2;
    int b = 144 /$cool$/ 3;
    int c = 4 /multiply/ 2 /mult/ 3;
    cout << "a= " << a << " b= " << b << ", c= " << c << 'n';
    return 0;
}
```

- This can be realized in C++
- Associativity (left or right) depends on how you overload `operator/`
- Of course do not use this in real projects!
Be critical

- Sadly not everything about C++ is great!
- A nice talk that sums up some of the biggest oddities of C++
- C++ bashing: “A case against C++”, Felix von Leitner
  - “Why C++ is bad for the environment, causes global warming and kills puppies”
  - https://media.ccc.de/v/cccamp07-en-1951-A_Case_Against_C++
- However, at the end I hope you find C++ somewhat useful!
Job offer

- Topic
  - Static analysis
  - C++ programming
  - LLVM compiler framework

- Benefits
  - Money ;-) 
  - Fun
  - Learn a lot
  - Invitations to our professional and social events
  - Opportunity for bachelor/ master thesis
  - Lots of career options
  - Working on an important topic

- Just send me an email
Recap

- Program analysis
- Real-world findings
- Static code analysis
- Custom operator hack

Next time

- Introduction to the final project
Thank you for your attention

Questions?