Analyzing C/C++ with cclyzer

George Balatsouras

LLVM Intermediate Representation

LLVM Bitcode

- low-level IR similar to assembly
- strongly typed RISC instruction set
- core of the LLVM umbrella project

LLVM IR - Basic Instructions

❏ Stack allocations	(1)	`p = alloca [type]`
❏ Heap allocations	(2)	`p = malloc nbytes`
❏ Load from address	(3)	`v = load p`
❏ Store to address	(4)	`store v, p`
❏ Address-of-field	(5)	p_{offset} `= &(p->f)`
❏ Address-of-array-index	(6)	p_{offset} `= &(p[i])`
❏ Function call	(7)	`v = call fn (`arg_1`, `arg_2`, ...)`
❏ No-op cast	(8)	`v = bitcast p to [type]`

LLVM Bitcode vs Java Bytecode

I. Addresses of Fields

LLVM Bitcode

- As in C, an instance field can have its *address taken*
- ... and then *loaded* elsewhere.
- By elsewhere, we mean even in a different function
- Expression `'p->f'` in fact translates to:

  p_{offset} `= &(p->f)`

  `v = load` p_{offset}

Java Bytecode*

- Impossible in Java
- May only allocate objects and then load from or store to some field
- Load/store instructions hence are ternary, containing an extra *field operand*
- Jimple* : stackless simplified format from the Soot framework

LLVM Bitcode vs Java Bytecode

II. Virtual registers

LLVM Bitcode

- All source-level variables become pointers
  … unless optimized away
- E.g., ‘`int p = 3;`’ becomes:

```
%p = alloca i32
store i32 3, i32* %p
```

- ‘`&p`’ becomes just ‘`%p`’
- Subsequent assignments to ‘`p`’ become store instructions to ‘`%p`’
- Additional temporary variables are introduced for intermediate expressions (e.g., ‘`%1`’, ‘`%2`’)
- Both ‘`%p`’ and ‘`%1`’, ‘`%2`’ are *virtual registers*.
- At register allocation:
  i. some will be replaced by *physical registers*
  ii. some will be *spilled*.

LLVM Bitcode vs Java Bytecode

III. Function Pointers

LLVM Bitcode

- `%v = call %fn (%arg`$_1$`, ...)`
- '`%fn`' can be either a *constant* or a *variable*
- Function pointers are actually used to implement C++ *dynamic dispatch*
    - v-table represented as global array, containing function pointers
    - More on this later...

Java Bytecode

- No such thing as a function pointer in Java
- Not even first class citizen methods (except via *reflection*)
- Invoke instruction variants:
  (i) `invokevirtual`, (ii) `invokespecial`,
  (iii) `invokestatic`, (iv) `invokeinterface`
- Much more high-level

LLVM Bitcode vs Java Bytecode

IV. Constant Expressions

LLVM Bitcode

- “A constant value that is initialized with an expression using other constant values.”
- ```
  @_ZTV1B = constant [5 x i8*] [
    null,
    bitcast (@_ZTI1B to i8*),
    bitcast (@_ZN1A3barEv to i8*),
    bitcast (@_ZN1B3fooEv to i8*),
    bitcast (@_ZN1B6foobarEv to i8*)
  ]
  ```
- Used in constant initializations for global variables, structs, arrays, etc.

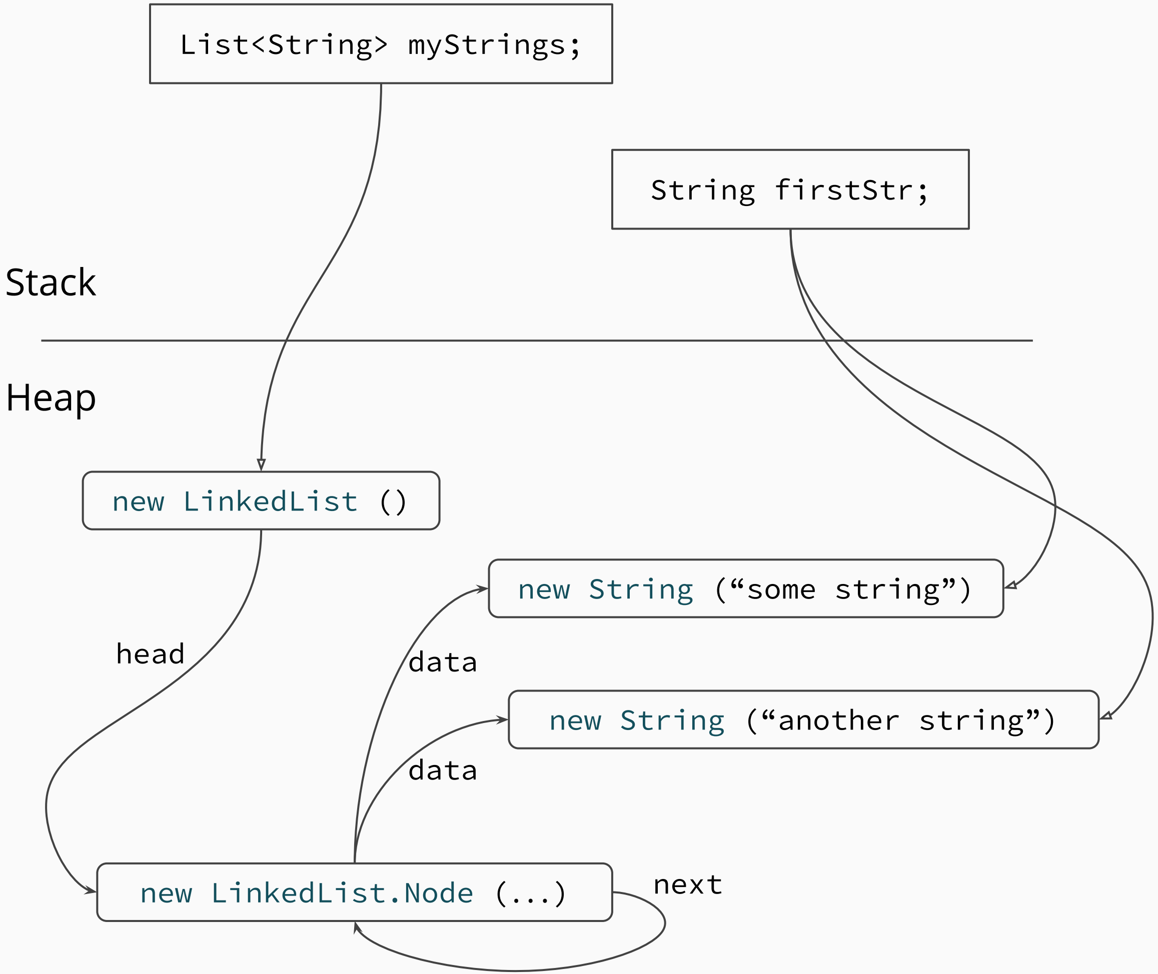
Java Bytecode

- “Nope, nothing wrong here.”
- `<clinit>` method for class initialization
  ```

Pointer Analysis on LLVM bitcode

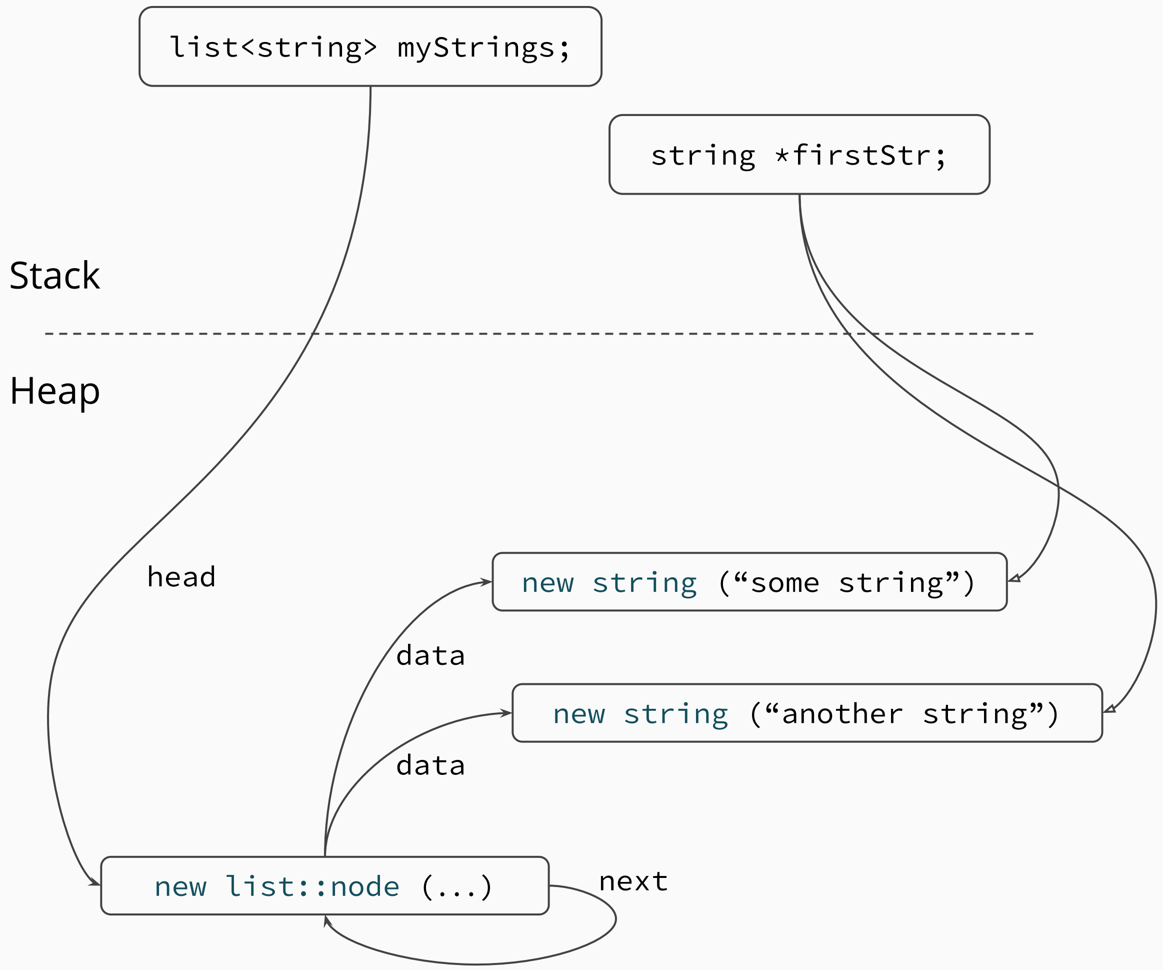
Java Memory Abstraction

- Clear distinction
  - variables reside on stack
  - allocated objects reside on heap
- Pointer analysis
  - variables *point-to* heap objects
  - heap objects *point-to* other heap objects through some field



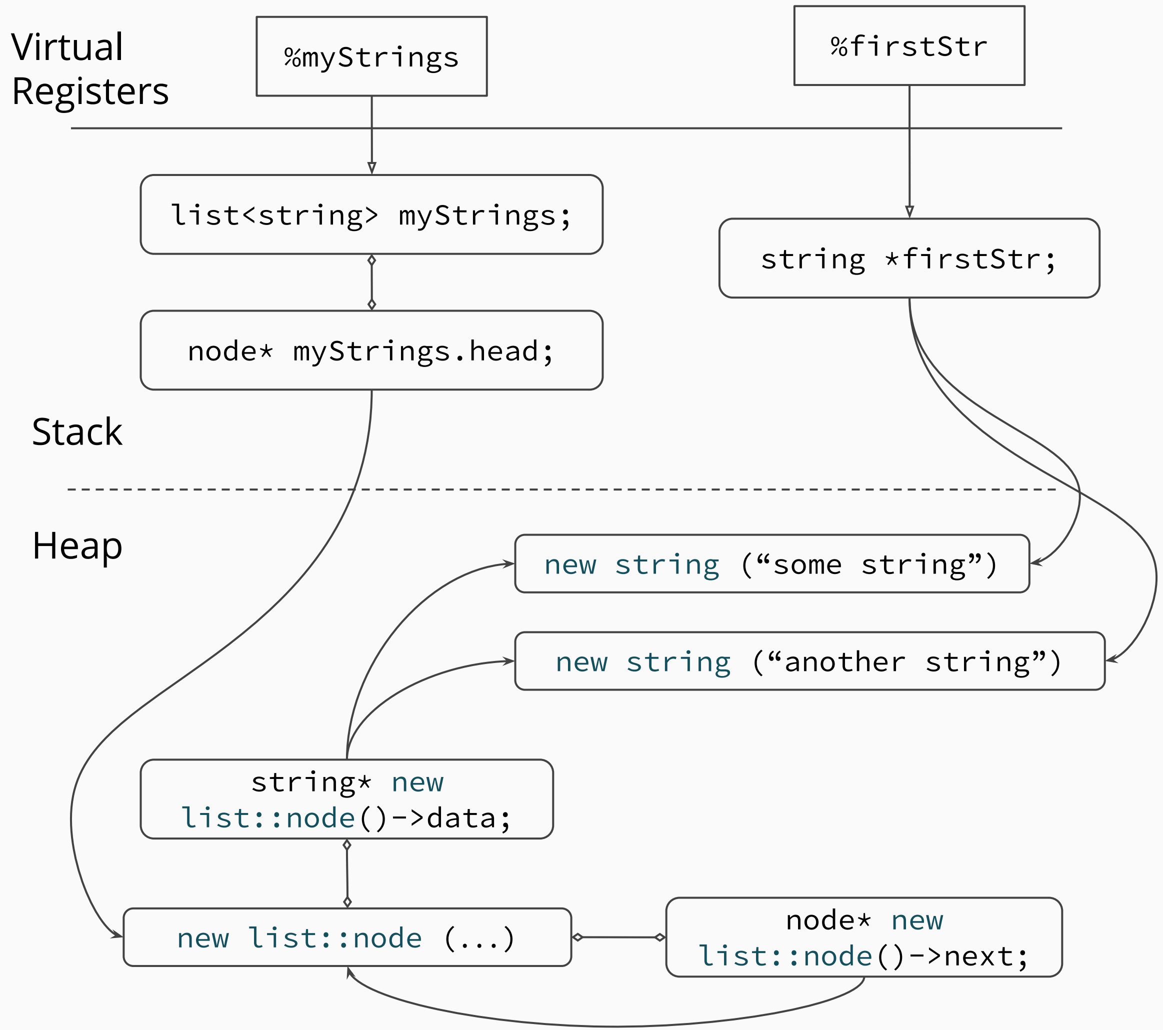
C/C++ Memory Abstraction

- Objects may be allocated:
  1. either on the heap
  2. or on the stack
- Pointer analysis
  - Dereference edges from abstract object to another abstract object
- What about field edges?
  - Objects contain other objects; unlike Java
  - Recall: we can take the address of a field



Our LLVM Memory Abstraction

- Decouple a variable from its stack allocation
- From now on, by *variable*, we mean virtual register
- Pointer analysis
  - Variables point-to (abstract) objects
  - Objects, when *dereferenced* point-to other objects
  - Fields of objects are objects themselves



Analyzing C/C++ code with cclyzer

https://github.com/plast-lab/cclyzer

Revisiting points-to

Simple Example

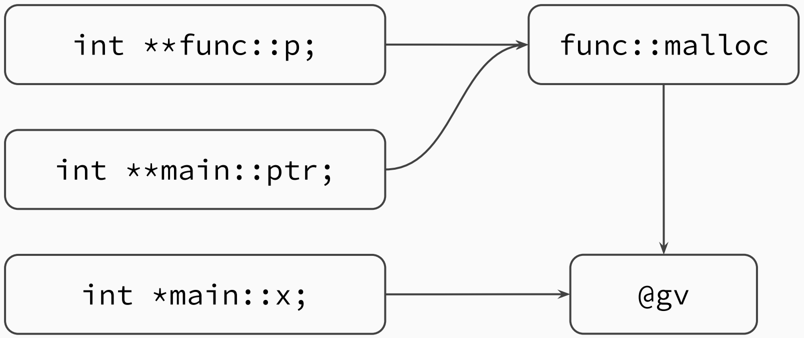
C Source Code

```
int gv;

int **func() {
   int **p = malloc(sizeof(int*));
   *p = &gv;
   return p;
}

int main(void) {
   int **ptr = func();
   int *x = *ptr;
   int y = *x;
}
```

What do we want to establish?



Revisiting points-to

LLVM Bitcode Translation

C Source Code

```
int gv;

int **func() {
    int **p = malloc(sizeof(int*));
    *p = &gv;
    return p;
}

int main(void) {
    int **ptr = func();
    int *x = *ptr;
    int y = *x;
}
```

LLVM Bitcode

```
int* @gv = global int 0;

int func() {
    int*** %p = alloca [int **];
    void*  %1 = call @malloc(8);
    int**  %2 = bitcast %1 to int**;

    store %2, %p;
    int**  %3 = load %p;
    store @gv, %3;

    ret %3;
}
```

Simple Example: Computing Points-to

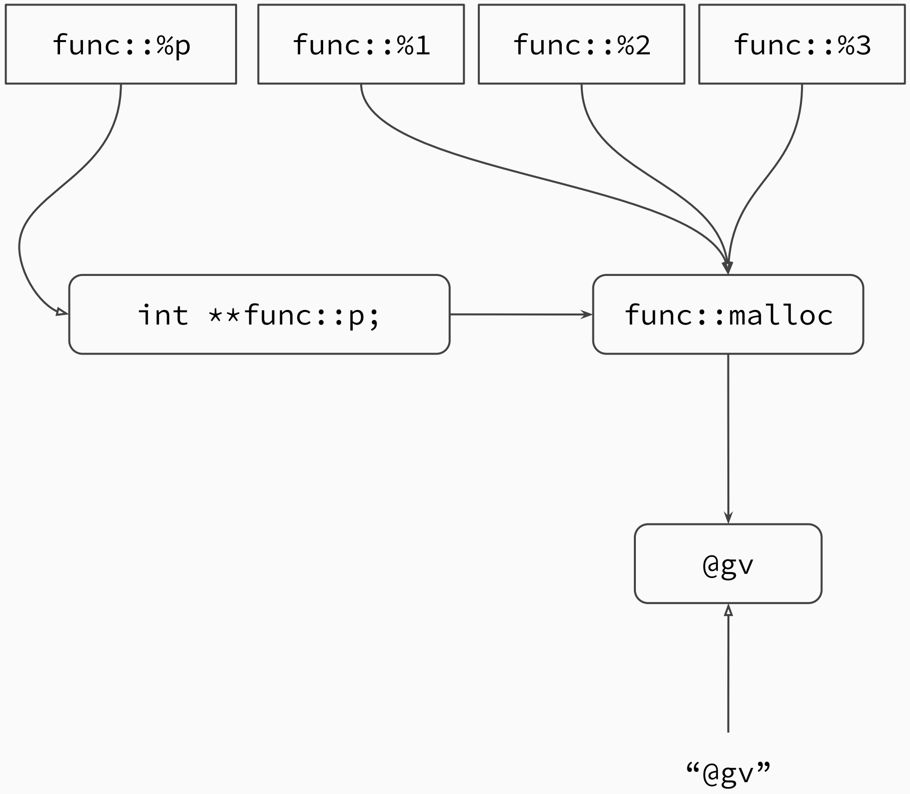
LLVM Bitcode

```
int* @gv = global int 0;

int func() {
    int*** %p = alloca [int **];
    void*  %1 = call @malloc(8);
    int**  %2 = bitcast %1 to
int**;

    store %2, %p;
    int**  %3 = load %p;
    store @gv, %3;

    ret %3;
}
```



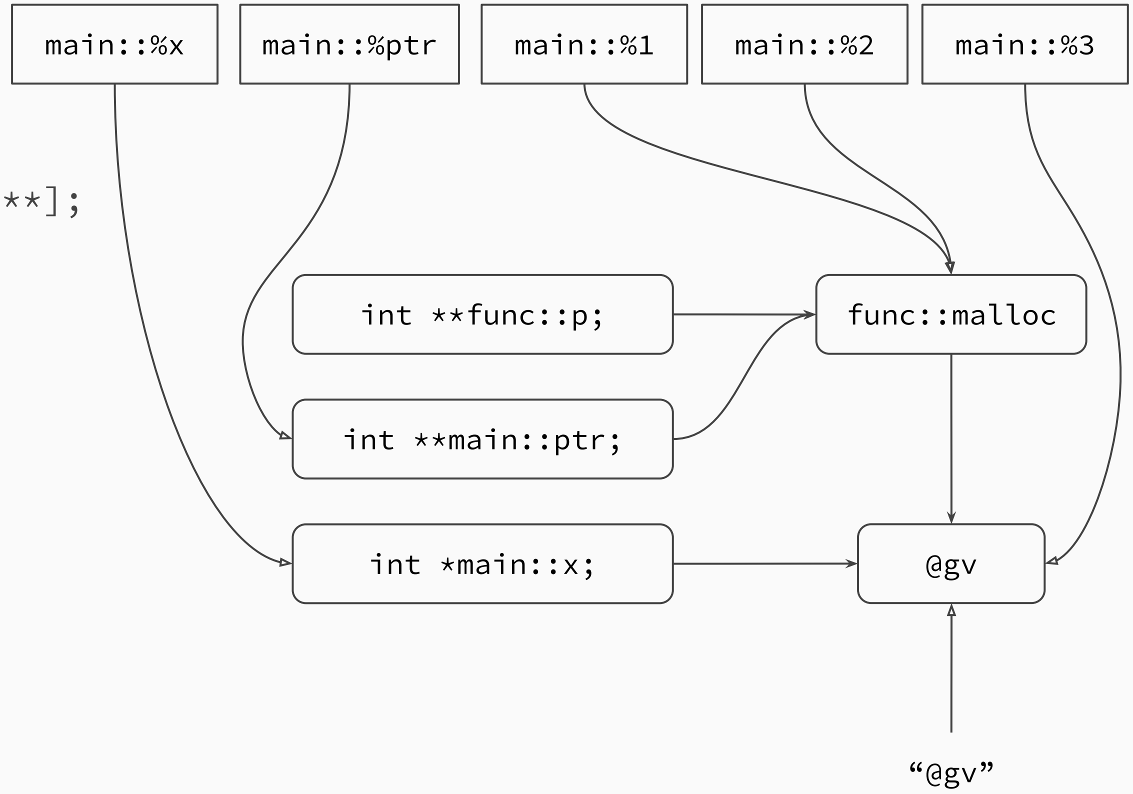
Simple Example: Computing Points-to

LLVM Bitcode

```
int main(void) {
    int*** %ptr = alloca [int **];
    int** %1 = call @func();
    store %1, %ptr;

    int** %x = alloca [int *];
    int** %2 = load %ptr;
    int*  %3 = load %2;
    store %3, %x;

    int* %y = alloca [int];
    int* %4 = load %x;
    int  %5 = load %4;
    store %5, %y;
}
```



Revisiting points-to

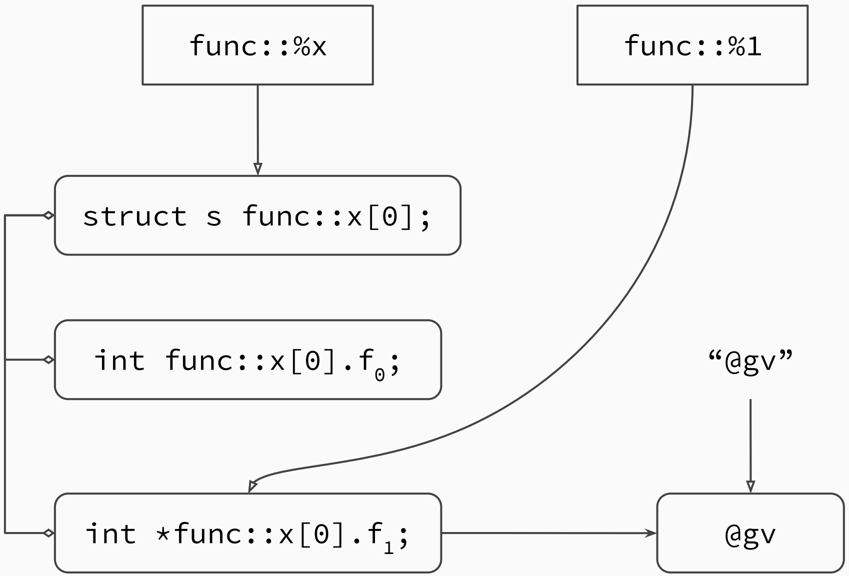
Field Sensitivity

LLVM Bitcode

```
int* @gv = global int 0;

%struct.s = type { int, int* }

void func() {
  %x = alloca [%struct.s];
  %1 = getelementptr %x, 0, 1; // &(x.f1)
  store @gv, %1;
}
```



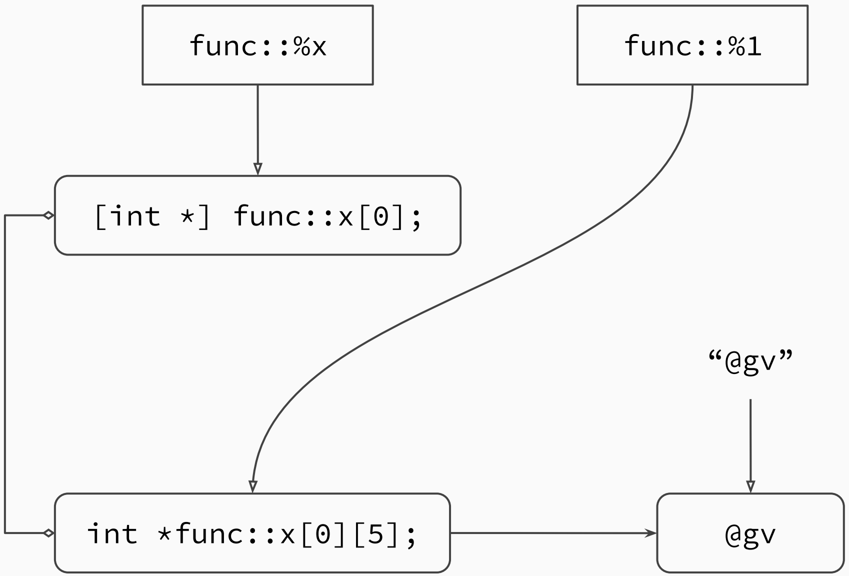
Revisiting points-to

Array Sensitivity

LLVM Bitcode

```
int* @gv = global int 0;


void func() {
  %x = alloca [100 x int*];
  %1 = getelementptr %x, 0, 5; // &(x
[5])
  store @gv, %1;
}
```



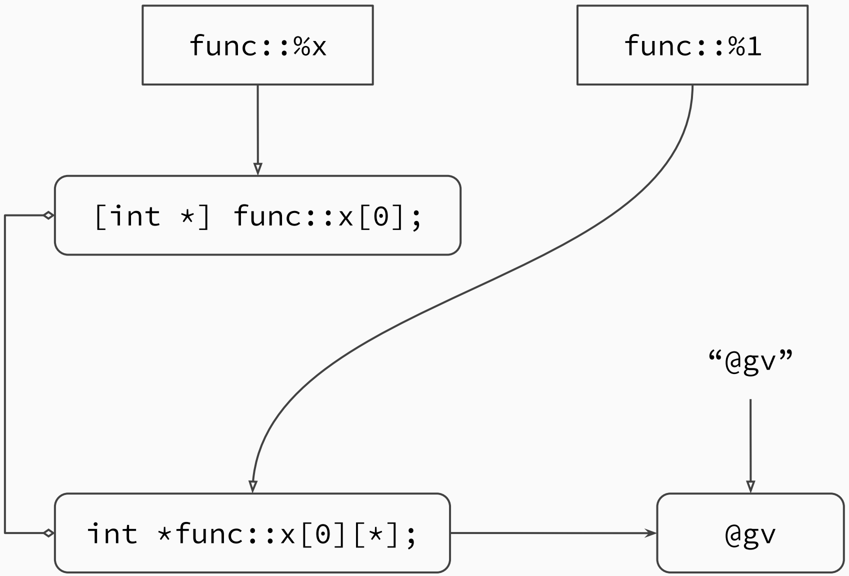
Revisiting points-to

Array Sensitivity

LLVM Bitcode

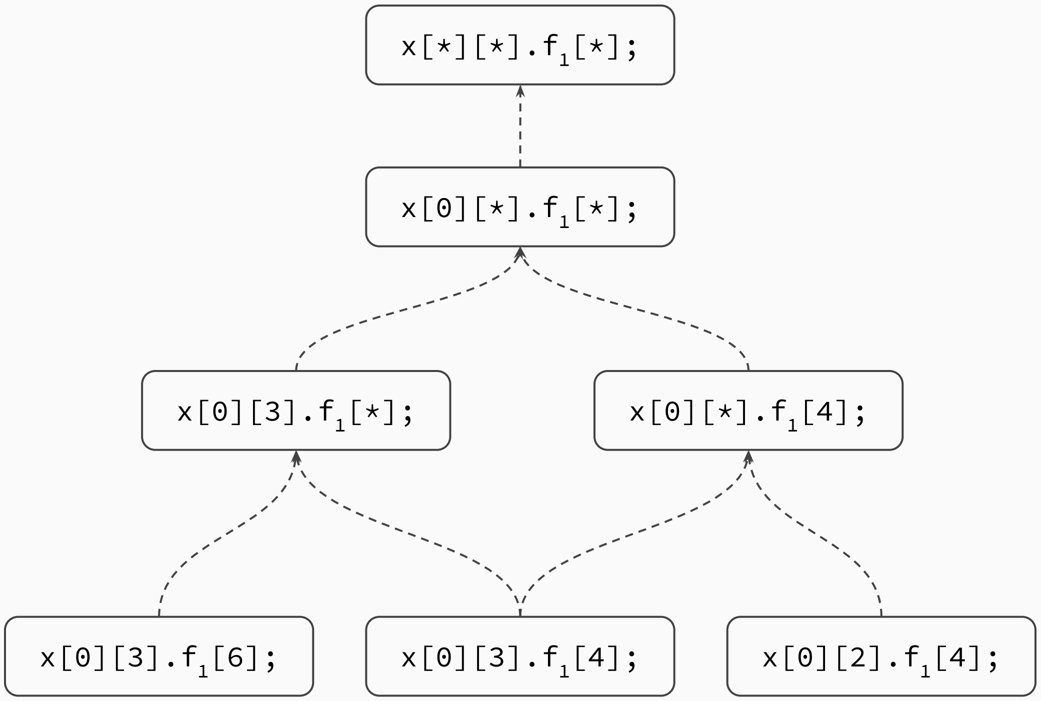
```
int* @gv = global int 0;


void func() {
  %x = alloca [100 x int*];
  %i = ...
  %1 = getelementptr %x, 0, %i; // &(x
[i])
  store @gv, %1;
}
```



Array Sensitivity

- Define partial order
- (n_1, n_2) when n_1 can be turned to n_2 by substituting constant indices with '*'
- points-to set of a node is a *superset* of the points-to set of its parent
- At *load* instructions, merge with the points-to sets of *all* children nodes



Abstract allocation sub-objects

Type-driven approach

- Create subobject as soon as type of base is determined
  - Create all field subobjects, for struct types
  - Consider only indices that appear on source code, for array/pointer types
- Allocation types will be mostly available at the allocation site
- If not, track *cast instructions* and create a *new* abstract object per possible type

Analyzing C++ code

compiled to LLVM IR

Challenges

- LLVM bitcode is a representation that is well-suited for C code
- Too low-level for C++
- C++ features like classes, v-tables, references, and so on are translated to low-level constructs

Dynamic Dispatch Example

LLVM Bitcode

```
%class.B = type { int (...)**, ...}

void func() {
  %b = alloca [%class.B];
  ...
  %1 = bitcast %b to int (%class.B*)***
  %2 = load int (%class.B*)** %1
  %3 = getelementptr int (%class.B*)** %2, 1
  %4 = load int (%class.B*)* %3
  call int %4 (%class.B* %b)
}
```