joeq compiler system

Benjamin Livshits

CS 243

Plan for Today

1. Joeq System Overview
2. Lifecycle of Analyzed Code
3. Source Code Representation
4. Writing and Running a Pass
5. Assignment: Dataflow Framework

1. Background on joeq

- A compiler system for analyzing Java code
  - Developed by John Whaley and others
  - Used on a daily basis by the SUIF compiler group
  - An infrastructure for many research projects: 10+ papers rely on joeq implementations
- Visit http://joeq.sourceforge.net for more...
- Or read http://www.stanford.edu/~jwhaley/papers/ivme03.pdf

joeq Design Choices

- Most of the system is implemented in pure Java
- Thus, analysis framework and bytecode processors work everywhere
- For the purpose of programming assignment, we treat joeq as a front- and middle end
- But it can be used as a VM as well
  - System-specific code is patched in when the joeq system compiles itself or its own runtime
  - These are ordinary C routines
  - Systems supported by full version: Linux and Windows under x86

joeq Components

- Full system is very large: ~100,000 lines of code
- Allocator
- Bootstrapper
- Classfile structure
- Compiler (Quad)
- Garbage Collector
- Quad Interpreters
- Memory Access
- Safe/Unsafe barriers
- Synchronization
- Assembler
- Class Library
- Compiler (Bytecode)
- Debugger
- Bytecode Interpreters
- Linkers
- Reflection support
- Scheduling
- UTF-8 Support

We restrict ourselves to only the compiler and classfile routines, which is closer to 40,000 lines of code

Starting at the Source

Lifecycle of Analyzed Code

- Everything begins as source code
- A very “rich” representation
  - Good for reading
  - Hard to analyze
- Lots of high-level concepts here with (probably) no counterparts in the hardware
  - Virtual function calls
  - Direct use of monitors and condition variables
  - Exceptions
  - Reflection
  - Anonymous classes
  - Threads

Source to Bytecode

- javac or jikes compiles source into a machine-independent bytecode format
- This still keeps the coarse structure of the program
  - Each class is a file
  - Split up into methods and fields
  - The bytecodes themselves are stored as a member attribute in methods that have them
  - Bytecoded instructions are themselves high level:
    - invokevirtual
    - monitorenter
    - arraylength

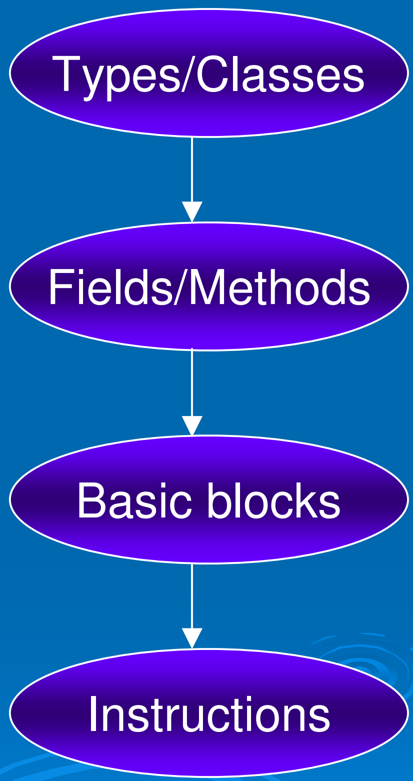
Analysis and Source Code

- Because so much of the code structure stays in the classfile format, there's no need for Java analyzers to bother with source code at all
- Moreover, bytecode is indifferent to language changes
- Reading in code:
  1. joeq searches through the ordinary classpath to find and load requested files
  2. Each source component in the classfile has a corresponding object representing it:
     - jq_Class
     - jq_Method
     - etc.
- Method bodies are transformed from bytecode arrays to more convenient representations:
  - more on this later

How Source Code is Represented within joeq

Source Code Representation

- joeq is designed primarily to work with Java
  - Operates at all levels of abstraction
  - Has classes corresponding to each language component
- Relevant packages in joeq
  - joeq.Class package: classes that represent Java source components (classes, fields, methods, etc.) reside in joeq's
  - Compil3r.BytecodeAnalysis package: analysis of Java bytecode
  - Compil3r.Quad package: Classes relevant to joeq's internal “quad” format
- Be careful with your imports:
  - avoid name conflicts with java.lang.Class and java.lang.Compiler classes



joeq.Class: Types and Classes

- jq_Type: Corresponds to any Java type
- jq_Primitive: subclass of jq_Type. Its elements (all static final fields with names like jq_Primitive.INT) represent the primitive types
- jq_Array: array types. Multidimensional arrays have a component type that is itself a jq_Array
- jq_Class: A defined class
- … all located in package

joeq.Class: Fields and Methods

- Subclasses of jq_Field and jq_Method, respectively
  - Class hierarchy distinguishes between instance and class (static) members, but this detail is generally hidden from higher analyses
- These classes know about their relevant types: who declares them, parameter/return types, etc.
- Names of members are stored as UTF.Utf8 objects, so you'll need to convert them with toString() to get any use out of them!

Analyzing Bytecode

- The Java Virtual Machine stores program code as *bytecodes* that serve as instructions to a stack machine of sorts
- Raw material for all analysis of Java code
- Preserves vast amounts of source information:
  - Java decompilers can almost perfectly reconstruct source, down to variable names and line numbers

Example of Java Bytecode

```java
class ExprTest {
    int test(int a){
      int b, c, d, e, f;
      c = a + 10;
      f = a + c;

      if(f > 2){
        f = f - c;
      }

      return f;
    }
}
```

```
int test(int);
  Code:
   0:   iload_1
   1:   bipush  10
   3:   iadd
   4:   istore_3
   5:   iload_1
   6:   iload_3
   7:   iadd
   8:   istore  6
   10:  iload   6
   12:  iconst_2
   13:  if_icmple       22
   16:  iload   6
   18:  iload_3
   19:  isub
   20:  istore  6
   22:  iload   6
   24:  ireturn
```

- javac test.java
- javap -c ExprTest

Bytecode Details

- The implied running model of the Java Virtual Machine is that of a stack machine - there are local variables that correspond to registers, and a stack where all computation occurs.
    - This is hard to analyze!
- Fortunately, the JVM requires that bytecode pass strict typechecking and stack consistency checking
- **Gosling Property:** At each instruction, the types of every element on the stack, and every local variable, are all well defined
- By extension, the stack must have a specific height at each program point

Converting Bytecodes to Quads

- joeq thus converts bytecodes to something closer to standard three-address code, called "Quads"
- The highly abstract bytecode instructions for the most part have direct counterparts in the Quad representation
- One operator, up to four operands

OPERATOR	OP1	OP2	OP3	OP4

- Approximately 100 operators, all told (filed into a dozen or so rough categories), about 15 varieties of operands
- Full details on these and the methods appropriate to them on the course website's joeq documentation:
    - http://suif.stanford.edu/~courses/cs243/joeq/

Operators

- Types of operators
  - Primitive operations: Moves, Adds, Bitwise AND, etc.
  - Memory access: Getfields and Getstatic
  - Control flow: Compares and conditional jumps, JSRs
  - Method invocation: OO and traditional
- Operators have suffixes indicating return type:
  - ADD_I adds two integers.
  - L, F, D, A, and V refer to longs, floats, doubles, references, and voids respectively
  - Operators may have _DYNLINK (or %) appended, which means that a new class may need loading at that point

Operands

- Operands are split into 15 types
  - The ConstOperand classes (I, F, A, etc.) indicate constant values of the relevant type
  - RegisterOperands name pseudo-registers
  - MethodOperands and ParamListOperands are used to identify method targets
  - TypeOperands are passed to type-checking operators, or to "new" operators
  - TargetOperands indicate the target of a branch

Converting a Method to Quads

```
BB0 (ENTRY) (in: <none>, out: BB2)
BB2          (in: BB0 (ENTRY), out: BB3, BB4)
1   ADD_I          T0 int,  R1 int, IConst: 10
2   MOVE_I         R3 int,  T0 int
3   ADD_I          T0 int,  R1 int, R3 int
4   MOVE_I         R6 int,  T0 int
5   IFCMP_I        R6 int,  IConst: 2, LE, BB4
BB3     (in: BB2, out: BB4)
6   SUB_I          T0 int,  R6 int, R3 int
7   MOVE_I         R6 int,  T0 int
BB4     (in: BB2, BB3, out: BB1 (EXIT))
8   RETURN_I       R6 int
BB1 (EXIT) (in: BB4, out: <none>)

Exception handlers: []
Register factory: Local: (I=7, F=7, L=7, D=7, A=7)
                    Stack: (I=2, F=2, L=2, D=2, A=2)
```

Control Flow and CFGs

- The class Compil3r.Quad.ControlFlowGraph encapsulates most of the information we'll ever need for our analyses
  - There's a a ControlFlowGraph in Compil3r.BytecodeAnalysis too, so be careful about your imports
- These are generated from jq_Methods by the underlying system's machinery (the CodeCache class) -- we use them to make QuadIterators
- (which we'll get to later)

Basic Blocks

- Raw components of Control Flow Graphs
- These know about their predecessors, successors, a list of Quads they contain, and information about exception handlers
  - Which ones protect this basic block
  - Which blocks this one protects
- Traditional BB semantics are violated by exceptions:
  - if an exception occurs, there is a jump from the middle of a basic block
  - We will ignore this subtlety

Safety Checks

- Java's safety checks are *implicit:* various instructions that do computation can also throw exceptions
- Joeq's safety checks are *explicit:* arguments have their values tested by various operators like NullCheck and BoundsCheck
  - Exceptions are thrown if checks fail
- When converting from bytecodes to quads, all necessary checks are automatically inserted

Iterating Over the Quads: QuadIterator

- Dealing with control flow graphs or basic blocks directly becomes tedious quickly
- Dealing with individual quads tends to miss the forest for the trees
- Simple interface to iterate through all the quads in reverse post-order, and provides immediate predecessor/successor data on each quad

```
jq_Method m = ...
ControlFlowGraph cfg = CodeCache.getCode(m);
QuadIterator iter = new QuadIterator(cfg)
while(iter.hasNext()) {
        Quad quad = (Quad)iter.next();

        if(quad.getOperator() instanceof Operator.Invoke) {
            processCall(cfg.getMethod(), quad);
        }
}
```

Developing a joeq Compiler Pass

4. Writing and Running a Pass

- Passes themselves are written in Java, implementing various interfaces Joeq provides
- Passes are invoked through library routines in the Main.Helper class
- Useful classes to import: Clazz.*, Compil3r.Quad.*, Main.Helper, and possibly Compil3r.Quad.Operator.* and Compil3r.Quad.Operand.*

The Main.Helper Class

- Main.Helper provides a clean interface to the complexities of the joeq system
  - load(String) takes the name of a class provides the corresponding jq_Class
  - runPass(target, pass) lets you apply any pass to a target that's at least that big
- So, how do we write a pass?

Visitors in joeq

- joeq makes heavy use of the visitor design pattern
- The visitor for a level of the code hierarchy has methods visitFoo(code object) for each type of object that level can take
- For some cases, you may have overlapping types (e.g., visitStore and visitQuad) -- the methods will be called from most-general to least-general
- Visitor interfaces with more than one method have internal abstract classes called "EmptyVisitor"
- Visitors are described in detail in “Design Patterns” by Gamma et al.

Visitors: Some Examples

```
public class QuadCounter extends QuadVisitor.EmptyVisitor {
    public int count = 0;
    public void visitQuad(Quad q){
        count++;
    }
}

public class LoadStoreCounter extends QuadVisitor.EmptyVisitor {
    public int loadCount = 0, storeCount = 0;
    public void visitLoad(Quad q){ loadCount++;}
    public void visitStore(Quad q){ storeCount++;}
}
```

Running a Pass

```
public class RunQuadCounter {
     public static void main(String[] args){
         jq_Class[] c = new jq_Class[args.length];
         for(int i = 0; i < args.length; i++){
                 c[i] = Helper.load(args[i]);
         }
         QuadCounter qc = new QuadCounter();
         for(int i = 0; i < args.length; i++){
                 qc.count = 0;
                 Helper.runPass(c[i], qc);
                 System.out.println(
                         c[i].getName() + " has " +
                         qc.count + " Quads.");
         }
     }
}
```

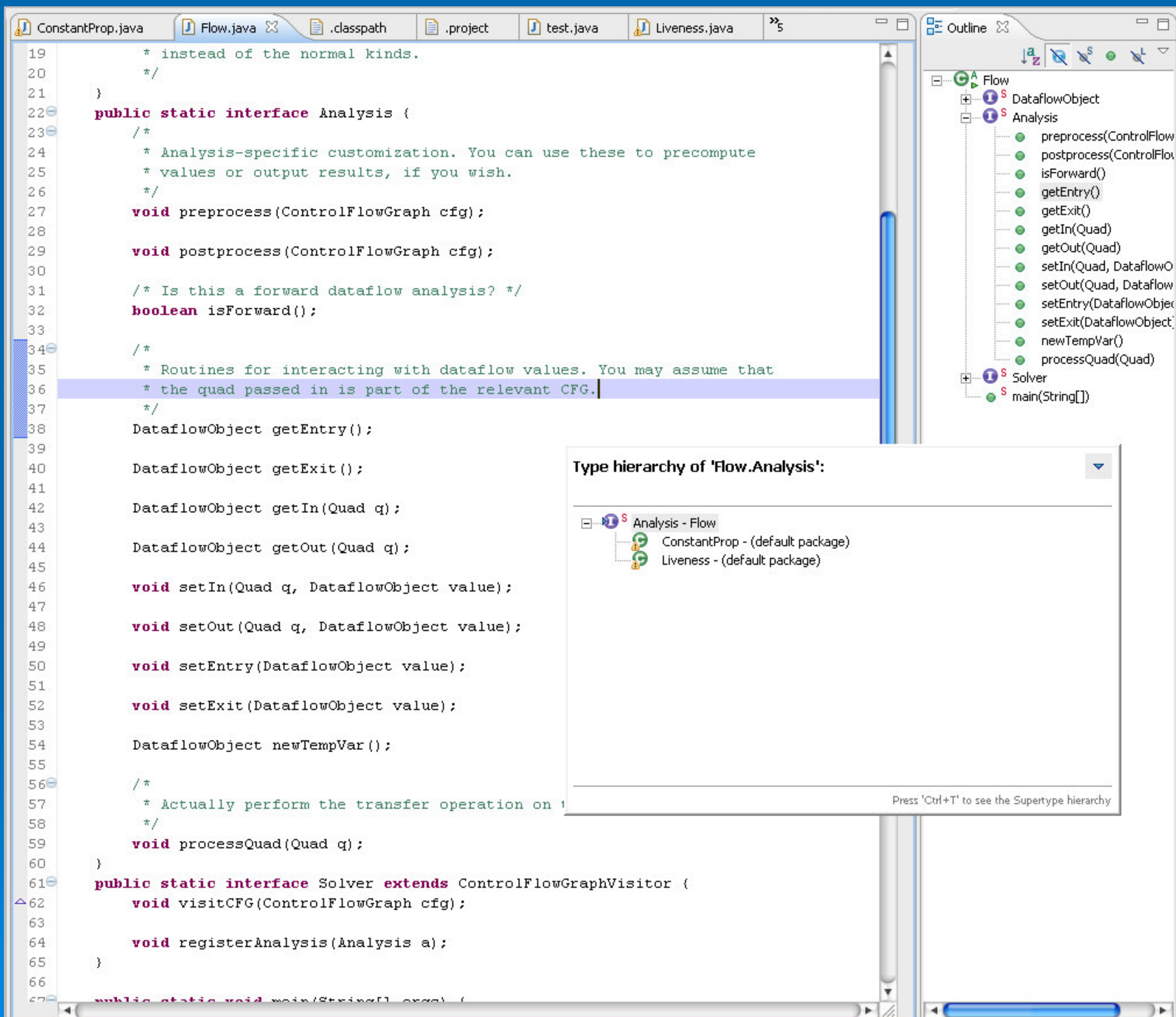
Summary

- We're using the Joeq compiler system
- Review of Java VM's code hierarchy
- Review of Joeq's code hierarchy
- QuadIterators
- Main.Helper
- Visitor pattern
- Defining and running passes

Programming Assignment 1

- Your assignment is to implement a **basic dataflow framework** using joeq
- We will provide the interfaces that your framework must support
- You will write the iterative algorithm for any analysis matching these interfaces, and also phrase Reaching Definitions in terms that any implementation of the solver can understand
  - A skeleton and sample analysis are available in /usr/class/cs243/dataflow
    - Flow.java contains the interfaces and the main program
    - ConstantProp.java contains classes that define a limited constant propagation algorithm

Flow.Analysis Interface



- You implement
  1. the solver and
  2. reaching definitions
- Test it first on the provided input
- Compare the output with the canonical one
- Be careful when writing your code
- We will throw more test cases at it