Global Value Numbering in Factor

A Thesis

Presented to the

Faculty of

California State Polytechnic University, Pomona

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

In

Computer Science

By

Alex Vondrak

2011

Signature Page

Thesis:	Global Value Numbering in Factor
Author:	Alex Vondrak
Date Submitted:	Summer 2011
	Computer Science

Dr. Craig Rich
Thesis Committee Chair
Computer Science

____

Dr. Daisy Sang
Computer Science

____

Dr. Amar Raheja
Computer Science

____

Acknowledgments

To Lindsay—he is my rock

Abstract

Compilers translate code in one programming language into semantically equivalent code in another language—canonically from a high-level language to low-level machine primitives. Generally, the further removed a language's abstractions get from those of a computer, the harder it gets to compile code into an efficient representation. What isn't redundant in the source language may map to repetitive target instructions that waste time recomputing results. To combat this, compilers try to optimize away redundancies by looking for values that are provably equivalent when the program is run.

This thesis explores the theory and implementation of a particularly aggressive analysis called global value numbering in a particularly high-level language called Factor. Factor is a stack-based, dynamically-typed, object-oriented language born in late 2003. A baby among languages (now at version 0.94), its compiler craves all the optimizations it can get. By altering the existing local value numbering pass, redundancies can be identified and eliminated across entire programs, rather than isolated regions of code. This induces speedups as high as 45% across the majority of benchmarks. The results from these comparatively simple changes hold much promise for future improvements in making Factor programs more efficient.

Table of Contents

List of Figures

1 Introduction

Compilers translate programs written in a source language (e.g., Java) into semantically equivalent programs in some target language (e.g., assembly code). They let us make our source language arbitrarily abstract so we can write programs in ways that humans understand while letting the computer execute programs in ways that machines understand. In a perfect world, such translation would be straightforward. Reality, however, is unforgiving. Straightforward compilation results in clunky target code that performs a lot of redundant computations. To produce efficient code, we must rely on less-than-straightforward methods. Typical compilers go through a stage of *optimization*, whereby a number of semantics-preserving transformations are applied to an *intermediate representation* of the source code. These then (hopefully) produce a more efficient version of said representation. Optimizers tend to work in *phases*, applying specific transformations during any given phase.

Global value numbering (GVN) is such a phase performed by many highly-optimizing compilers. Its roots run deep through both the theoretical and the practical. Using the results of this analysis, the compiler can identify expressions in the source code that produce the same value—not just by lexical comparison (i.e., comparing variable names), but by proving equivalences between what's actually computed at runtime. These expressions can then be simplified by further algorithms for redundancy elimination. This is the very essence of most compiler optimizations: avoid redundant computation, giving us code that runs as quickly as possible while still following what the programmer originally wrote.

High-level, dynamic languages tend to suffer from efficiency issues. They're often interpreted rather than compiled, and perform no heavy optimization of the source code. However, the Factor language (`http://factorcode.org`) fills an intriguing design niche, as it's very high-level yet still fully compiled. It's still young, though, so its compiler craves all the improvements it can get. In particular, while the current Factor version (as of this writing, 0.94) has a *local* value numbering analysis, it is inferior to GVN in several significant ways.

In this thesis, we explore the implementation and use of GVN in improving the strength of optimizations in Factor. Because Factor is a young and relatively unknown language, Chapter 2 provides a short tutorial, laying a foundation for understanding the changes. **??** describes the overall architecture of the Factor compiler, highlighting where the exact contributions of this thesis fit in. Finally, **??** goes into detail about the existing and new value numbering passes, closing with a look at the results achieved and directions for future work.

In the unlikely event that you want to cite this thesis, you may use the following BibTeX entry:

```
@mastersthesis{vondrak:11,
  author = {Alex Vondrak},
  title  = {Global Value Numbering in Factor},
  school = {California Polytechnic State University, Pomona},
  month  = sep,
  year   = {2011},
}
```

2 Language Primer

Factor is a rather young language created by Slava Pestov in September 2003 [*Factor* 2010]. Its first incarnation was an embedded scripting language for a game that targeted the Java Virtual Machine (JVM). As such, its feature set was minimal. Factor has since evolved into a general-purpose programming language, gaining new features and redesigning old ones as necessary for larger programs. Today's implementation sports an extensive standard library and has moved away from the JVM in favor of native code generation. In this chapter, we cover the basic syntax and semantics of Factor for those unfamiliar with the language. This should be just enough to understand the later material in this thesis. More thorough documentation can be found via Factor's website, `http://factorcode.org`.

2.1 Stack-Based Languages

Like Reverse Polish Notation (RPN) calculators, Factor's evaluation model uses a global stack upon which operands are pushed before operators are called. This naturally facilitates postfix notation, in which operators are written after their operands. For example, instead of `1 + 2`, we write `1 2 +`. Figure 1 on the following page shows how `1 2 +` works conceptually:

- `1` is pushed onto the stack
- `2` is pushed onto the stack
- `+` is called, so two values are popped from the stack, added, and the result (`3`) is pushed back onto the stack

Other stack-based programming languages include Forth [American National Standards Institute and Computer and Business Equipment Manufacturers Association 1994], Cat [Diggins 2007], and PostScript [Adobe Systems Incorporated 1999].

The strength of this model is its simplicity. Evaluation essentially goes left to right: literals (like `1` and `2`) are pushed onto the stack, and operators (like `+`) perform some computation using values currently on the stack. This "flatness" makes parsing easier,

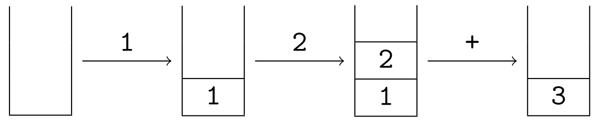

Figure 1: Visualizing stack-based calculation

since we don't need complex grammars with subtle ambiguities and precedence issues. Rather, we basically just scan left-to-right for tokens separated by whitespace. In the Forth tradition, functions are called *words* since they're made up of any contiguous non-whitespace characters. This also lends to the term *vocabulary* instead of "module" or "library". In Factor, the parser works as follows:

- If the current character is a double-quote (`"`), try to parse ahead for a string literal.
- Otherwise, scan ahead for a single token.
  - If the token is the name of a *parsing word*, that word is invoked with the parser's current state.
  - If the token is the name of an ordinary (i.e., non-parsing) word, that word is added to the parse tree.
  - Otherwise, try to parse the token as a numeric literal.

Parsing words serve as hooks into the parser, letting Factor users extend the syntax dynamically. For instance, instead of having special knowledge of comments built into the parser, the parsing word `!` scans forward for a newline and discards any characters read (adding nothing to the parse tree).

Similarly, there are parsing words for what might otherwise be hard-coded syntax for data structure literals. Many act as sided delimeters: the parsing word for the left-delimiter will parse ahead until it reaches the right-delimiter, using whatever was read in between to add objects to the data structure. For example, `{ 1 2 3 }` denotes an array of three numbers. Note the deliberate spaces in between the tokens, so that the delimeters are themselves distinct words. In `{␣1␣2␣3␣}` (with spaces as marked), the parsing word `{` parses objects until it reaches `}`, collecting the results into an array. The `{` word would not

```
V{ 1 2 3 }                          ! vector
B{ 1 2 3 }                          ! byte array
BV{ 1 2 3 }                         ! byte vector
HS{ 1 2 3 }                         ! hash set
H{ { key1 val1 } { key2 val2 } }    ! hash table
```

Figure 2: Data structure literals in Factor

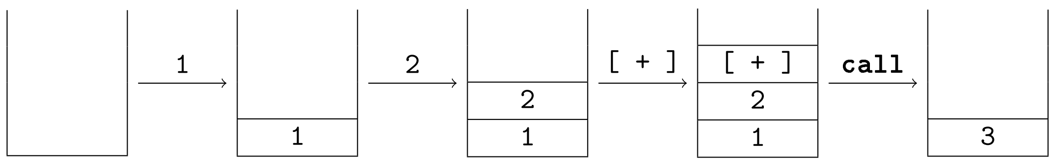

Figure 3: Quotations

be called if not for that space, whereas `{1␣2␣3}` parses as the word `{1`, the number `2`, and the word `3}`—not an array. Further, since the left-delimeter words parse recursively, such literals can be nested, contain comments, etc. Other literals include those in Figure 2.

A particularly important set of parsing words in Factor are the square brackets, `[` and `]`. Any code in between such brackets is collected up into a special sequence called a *quotation*. Essentially, it's a snippet of code whose execution is suppressed. The code inside a quotation can then be run with the **`call`** word. Quotations are like anonymous functions in other languages, but the stack model makes them conceptually simpler, since we don't have to worry about variable binding and the like. Consider a small example like

```
1 2 [ + ] call
```

You can think of **`call`** working by "erasing" the brackets around a quotation, so this example behaves just like `1 2 +`. Figure 3 shows its evaluation: instead of adding the numbers immediately, `+` is placed in a quotation, which is pushed to the stack. The quotation is then invoked by **`call`**, so `+` pops and adds the two numbers and pushes the result onto the stack. We'll show how quotations are used in **??** on page ??.

References

Adobe Systems Incorporated. 1999. *PostScript Language Reference*. Third edition. Addison-Wesley. ISBN: 0-201-37922-8. URL: `http://partners.adobe.com/public/developer/en/ps/PLRM.pdf` (visited on August 15, 2011).

American National Standards Institute and Computer and Business Equipment Manufacturers Association. 1994. *American National Standard for information systems: programming languages: Forth: ANSI/X3.215-1994*. American National Standards Institute.

Diggins, C. 2007. *The Cat Programming Language*. URL: `http://cat-language.com/` (visited on August 15, 2011).

Factor. 2010. From Factor's documentation wiki. URL: `http://concatenative.org/wiki/view/Factor` (visited on August 15, 2011).