Enabling Optimizations through Demodularization

Blake Johnson

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Eric Mercer, Chair
Christophe Giraud-Carrier
Quinn Snell

Department of Computer Science

Brigham Young University

December 2015

ABSTRACT

Enabling Optimizations through Demodularization

Blake Johnson
Department of Computer Science, BYU
Master of Science

Programmers want to write modular programs to increase maintainability and create abstractions, but modularity hampers optimizations, especially when modules are compiled separately or written in different languages. In languages with syntactic extension capabilities, each module in a program can be written in a separate language, and the module system must ensure that the modules interoperate correctly. In Racket, the module system ensures this by separating module code into phases for runtime and compile-time and allowing phased imports and exports inside modules. We present an algorithm, called demodularization, that combines all executable code from a phased modular program into a single module that can then be optimized as a whole program. The demodularized programs have the same behavior as their modular counterparts but are easier to optimize. We show that programs maintain their meaning through an operational semantics of the demodularization process and verify that performance increases by running the algorithm and existing optimizations on Racket programs.

Keywords: macros, Racket, modules, optimization

ACKNOWLEDGMENTS

Jay McCarthy, Matthew Flatt, Eric Mercer

Contents

List of Figures

List of Tables

List of Listings

Chapter 1

Introduction

Programmers should not have to sacrifice the software engineering goals of modular design and good abstractions for performance. Instead, their tools should make running a well-designed program as efficient as possible.

Many languages provide features for creating modular programs which enable separate compilation and module reuse. Some languages provide expressive macro systems, which enable programmers to extend the compiler in arbitrary ways. Combining module systems with expressive macro systems allow programmers to write modular programs with each module written in its own domain-specific language. A compiler for such a language must ensure that modular programs have the same meaning independent of the order in which the modules are compiled. A phased module system, like the one described by Flatt [5] for Racket, is a way to allow both separately compiled modules and expressive macros in a language.

Modular programs are difficult to optimize because the compiler has little to no information about values that come from other modules when compiling a single module. Existing optimizations have even less information when modules can extend the compiler. Good abstractions are meant to obscure internal implementations so that it is easier for programmers to reason about their programs, but this obscurity also limits information available for optimizations. In contrast, non-modular programs are simpler to optimize because the compiler has information about every value in the program.

Some languages avoid the problem of optimizing modular programs by not allowing

modules, while others do optimizations at link time, and others use inlining. Not allowing modules defeats the benefits of modular design. Link time optimizations can be too low level to do useful optimizations. Inlining must be heuristic-based, and good heuristics are hard to develop.

Our solution for optimizing modular programs, called demodularization, is to transform a modular program into a non-modular program by combining all runtime code and data in the program into a single module. In a phased module system, finding all of the runtime values is not trivial. Phased module systems allow programmers to refer to the same module while writing compiler extensions and while writing normal programs. A demodularized program does not need to include modules that are only needed during compile-time, but whether or not the module is needed only at compile-time is not obvious from just examining the module in isolation.

A program with a single module is effectively a non-modular program. After demodularization, a program becomes a single module, so existing optimizers have more information. Also, demodularization enables new optimizations that need whole program information.

We provide an operational semantics for a simple language with a phased module system, and argue that the demodularization process preserves program meaning. We also provide an implementation of demodularization for the Racket programming language, and verify experimentally that programs perform better after demodularization.

We explain demodularization at a high level with a detailed example (Chapter 2). Next, we use the operational semantics model of the demodularization process to explain why demodularization is correct (Chapter 3), then describe an actual implementation for Racket (Chapter 4), followed by experimental results of demodularizing and optimizing real-world Racket programs (Chapter 5). The operational semantics model removes the unnecessary details of the full implementation so the demodularization process is easier to understand and verify. The actual implementation presents interesting difficulties that the model does not. The experimental results show that demodularization improves performance, especially

when a program is highly modular.

Chapter 2

The Racket Module System

Modules are the basic building block of programs written in Racket. A Racket module specifies the language it is written in, definitions, expressions (run for their effect), and imports and exports. Modules usually correspond to files, but multiple modules can be defined in a single file through the use of submodules. A module's language controls all aspects of what the module means, from how it is parsed to what it means to do function application. Usually the language definition is just another Racket module. Racket definitions can be normal runtime definitions or compile-time definitions. The Racket compiler separates compile-time and runtime definitions into different phases, with compile-time definitions at phase 1 and runtime definitions at phase 0. Macros are special definitions with phase 0 bindings, but with phase 1 values. Macros allow programmers to add new features to their programs that can't be added through normal function definitions. Racket also allows for definitions directly at phase 1 so that programmers can write helper code for macros through `begin-for-syntax`.

Imports in Racket use the `require` form, which allows for importing definitions from other modules in a variety of ways. `require` lets the programmer specify which identifiers to import and any renamings to use. Also, the programmer can specify at what phase to import another module so they can use other modules to help write macros. Exports in Racket use the `provide` form, which has similar features as `require`, such as control over which identifiers are exported, renaming capabilities, and phase shifting capabilities.

The following example illustrates some of the features of the racket module system.

TODO: racket module example

The `obj.rkt` file contains the `obj` module. The `obj` module contains a variable definition, and an accessor and mutator function which are exported. The `while-test.rkt` file contains a module written in the `while-lang` language, which adds a `while` loop form to Racket (Racket has many sophisticated ways to write loops, but none of them use the keyword `while`). The `while-lang` language is just another Racket module, contained in `while-lang.rkt`. The `while-lang` module defines a `while` loop as a macro that transforms into a combination of a `loop` and an `if`. Also, the module defines a special `#%module-begin` macro which is a hook that will run for every module written in the `while-lang` language. Next, the module includes a `begin-for-syntax` expression. All of the code written inside a `begin-for-syntax` expression is shifted to phase 1 (compile-time). Therefore, the `(update-val)` and `(printf)` expressions will run when the module is compiled.

Chapter 3

Implementation

The demodularization algorithm for the Racket module system operates on Racket bytecode. Racket's bytecode format is one step removed from the fully-expanded kernel language: instead of identifiers for bindings, it uses locations. For toplevel bindings, these locations point to memory allocated for each module known as the module's prefix. So, in `long-queue.rkt`, `make-long-queue` would be in prefix location 0 and `long-enqueue` would be in prefix location 1, and all the references to `make-long-queue` and `long-enqueue` are replaced with references to 0 and 1. Like in the model, the algorithm combines all phase 0 code into a single module, but since the references are locations instead of identifiers, the locations of different modules overlap. We solve this by extending the prefix of the main module to have locations for the required module's toplevel identifiers, and then adjusting the toplevel references in the required module to those new locations.

After combining all the code for a program into a single module, we want to optimize it. The existing optimizations for Racket operate on an intermediate form that is part way between fully-expanded code and bytecode. Therefore, to hook into the existing optimizations, we decompile the bytecode of the demodularized program into the intermediate form and then run it through the optimizer to produce bytecode once more.

Racket provides features that treat modules as first-class objects during runtime. For example, programs can load and evaluate modules at runtime through `dynamic-require`. These features can work with demodularization, but the onus is on the programmer to make sure to use the features in particular ways. The main restriction is that the program cannot

share a module that is part of the demodularized program and also part of a dynamically required module. This restriction may seem easy to follow in theory, but in practice it is hard because most modules rely on built-in Racket libraries that will be in both the static and dynamic parts of the program.

Chapter 4

Evaluation

We tested our implementation of demodularization by selecting existing Racket programs and measuring their execution time before and after demodularization. We also measured the memory usage and compiled bytecode size of the programs. We ran the benchmarks on an Intel Core 2 Quad machine running Ubuntu and ran each program X times. We expect programs to perform better based on how modular the program is, which we measure by counting the number of modules in a program's require graph and how many cross module references occur in the program.

Figure XXX shows the results of running this experiment on XXX Racket programs. On one end of the spectrum, there are programs like XXX which are already basically single module programs, so demodularization does little besides rerun the optimizer on the program. Running the optimizer again may have positive or negative effects on performance, it may unroll loops and inline definitions more aggressively the second time, but some of these "optimizations" may hurt performance. On the other end of the spectrum, highly modular programs like XXX perform much better after demodularization. We expect performance to increase at a linear or even superlinear pace as modularity increases because of the extra information available to the optimizer.

This experiment uses only the existing Racket optimizations, which are intra-module optimizations. Certain optimizations that are not worthwhile to do at the intra-module level have larger payoffs when applied to whole programs. With demodularization, we anticipate that new whole-program optimizations enabled by demodularization will increase

performance even more.

Chapter 5

Related Work

Prior work on whole-program optimization has come in two flavors, depending on how much access to the source code the optimizer has. The first approach assumes full access to the source code and is based on inlining. The second approach only has access to compiled modules and is based on combining modules.

The first approach is based on selectively inlining code across module boundaries because it has full access to the source code of the program [1, 2]. Most of the focus of this approach is finding appropriate heuristics to inline certain functions without ballooning the size of the program and making sure the program still produces the same results. Resulting programs are not completely demodularized; they still have some calls to other modules. Specifically, Chambers et al. [2] show how this approach applies to object-oriented languages like C++ and Java, where they are able to exploit properties of the class systems to choose what to inline. Blume and Appel [1] showed how to deal with inlining in the presence of higher order functions, to make sure the semantics of the program didn't change due to inlining. Their approach led to performance increases of around 8%.

The second approach is taking already compiled modules, combining them into a single module, and optimizing the single module at link time [3, 4]. Most of the work done with this approach optimized at the assembly code level, but because they were able to view the whole program, the performance increases were still valuable. The link-time optimization system by Sutter et al. [3] achieves a 19% speedup on C programs. One of the reasons for starting with compiled modules is so that programs using multiple languages

can be optimized in a common language, like the work done by Debray et al. [4] to combine a program written in both Scheme and Fortran. The main problem with this approach is that the common language has less information for optimization than the source code had. These approaches are similar to demodularization, but the operate at a lower level and work on languages without phased module systems.

Chapter 6

Conclusion

Demodularization is a useful optimization for deploying modular programs. A programmer can write a modular program and get the benefits of separate compilation while devloping the program, and then get additional speedups by running the demodularizer on the completed program. Demodularization also enables new optimizations that are not feasible to implement for modular programs. Without module boundaries, inter-procedural analysis is much easier and worthwhile. Also, dead code elmination works much better because the whole program is visible, while in a modular program, only dead code that is private to the module can be eliminated.

In the future, we would like to implement an aggressive dead code elimination algorithm for Racket. We implemented a naive one that does not respect side effects, but shows the potential gains from this optimization; it is able to shrink Racket binaries down from about 2MB to about 100KB. This promising result implies that other low-hanging optimizations should be possible on demodularized programs that can increase performance.

References

[1] Matthias Blume and Andrew W. Appel. Lambda-splitting: a higher-order approach to cross-module optimizations. In *Proceedings of the second ACM SIGPLAN international conference on Functional programming*, ICFP '97, pages 112–124, New York, NY, USA, 1997. ACM.

[2] Craig Chambers, Jeffrey Dean, and David Grove. Whole-program optimization of object-oriented languages. Technical report, 1996.

[3] Bjorn De Sutter, Bruno De Bus, and Koen De Bosschere. Link-time binary rewriting techniques for program compaction. *ACM Trans. Program. Lang. Syst.*, 27(5):882–945, September 2005.

[4] Saumya K. Debray, Robert Muth, and Scott A. Watterson. Link-time improvement of scheme programs. In *Proceedings of the 8th International Conference on Compiler Construction, Held as Part of the European Joint Conferences on the Theory and Practice of Software, ETAPS'99*, CC '99, pages 76–90, London, UK, UK, 1999. Springer-Verlag.

[5] Matthew Flatt. Composable and compilable macros:: you want it when? In *Proceedings of the seventh ACM SIGPLAN international conference on Functional programming*, ICFP '02, pages 72–83, New York, NY, USA, 2002. ACM.