Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2015-012 April 14, 2015

Horizontal Code Transfer via Program Fracture and Recombination

Stelios Sidiroglou-Douskos, Eli Davis, and Martin Rinard

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Horizontal Code Transfer via Program Fracture and Recombination

Stelios Sidiroglou-Douskos, Eli Davis, and Martin Rinard

April 14, 2015

Abstract

We present a new horizontal code transfer technique, program fracture and recombination, for automatically replacing, deleting, and/or combining code from multiple applications. Benefits include automatic generation of new applications incorporating the best or most desirable functionality developed anywhere, the automatic elimination of security vulnerabilities, effective software rejuvenation, the automatic elimination of obsolete or undesirable functionality, and improved performance, simplicity, analyzability, and clarity.

1 Introduction

Horizontal gene transfer enables organisms to acquire useful functionality evolved and refined in other organisms. Examples include plasmid transfer (which plays a major role in acquired antibiotic resistance [2]), virally-mediated gene therapy [6], and the transfer of genes that code for insect toxins from bacteria to fungi to provide insect resistance in symbionts [1]. Horizontal gene transfer is recognized as significant factor in the development of many forms of life [7].

Like biological organisms, software developers also leverage previous development and refinement efforts, in this case via code reuse. While code reuse can significantly reduce development effort, in its current form it requires significant manual effort as developers work to locate, extract, and integrate previously developed code from other programs into the program they are developing. CodePhage, which automatically locates and transfers code that implements security checks from donors to recipients (endowing the recipient with immunity against previously effective attacks) highlights the tremendous potential of automatic horizontal code transfer techniques [13, 12].

We present a new horizontal code transfer technique, *program fracture and recombination*, for automatically locating and transferring computations between multiple applications. This technique promises to significantly advance our ability to more productively leverage the enormous amount of software that already exists but has not been packaged into easily reusable components.

Even more, program fracture and recombination holds out the promise of automatic program improvement and evolution without the need for any developer or potentially even any human involvement. Starting with multiple programs, horizontal program fracture and recombination operates as follows:

- **Fracture:** Fracture the programs into *shards* — each shard is a piece or pieces of the program that implements a computation or functionality. The granularity of the fracture determines the size of the shards. Potentially useful granularities include functions, procedures, classes, abstract data types, modules, loops, and program slices. Program fracture typically includes the encapsulation of each shard into its own separately invocable program for testing, analysis, and exploration.

- **Characterization:** Characterize the behavior and characteristics of each shard. Examples of such characterization include running the encapsulated shard on automatically generated inputs to obtain example input/output pairs for the shard, recording input/output pairs for the shard as invoked in context by sample executions of the program in which it was originally embedded, static analyses which partially or completely characterize the semantics of the shard, abstractions of the shard semantics obtained by generalizing the recorded input/output pairs, and specifications, either inferred or provided by the developer.

- **Shard Matching and Replacement:** One potential application of program fracture and recombination is to replace an original shard from one program with a better replacement shard from another program. There are many axes along which the replacement shard may be better — it may be more efficient, simpler to understand, endowed with additional capabilities such as the ability to execute successfully in parallel or distributed environments, it may have more error checking code, it may be more secure or better preserve privacy, or it may be missing undesirable or irrelevant functionality.

  Shard matching and replacement matches a shard from one program with a replacement shard from another program (or even potentially the same program). It then inserts the replacement shard into the place of the original shard.

- **Shard Insertion:** Another potential application of program fracture and recombination is to transfer functionality from one program to another. In this application a shard is taken from a donor program and inserted into a recipient. CodePhage, which automatically locates and transfers security checks across multiple applications, implements a form of shard insertion [13, 12].

- **Shard Removal:** As software evolves, previously desirable functionality can often become irrelevant. Potential drawbacks include difficulty analyzing the program and residual errors and vulnerabilities left over in the

now irrelevant code [11, 10]. Shard removal can automatically eliminate the now undesirable code and functionality. It can also eliminate functionality that should never have been introduced into the application at all.

Given the ease of obtaining sample inputs and outputs, either by automatically generating the input or by recording inputs presented to shards in context during executions of the enclosing programs, we expect input/output driven shard identification and transfer to play a prominent role. Input/output driven approaches can also promote the replacement of shards with other shards with different semantics — for example, the replacement of shards with errors or incomplete implementations with shards that have fewer errors or implement more cases. One straightforward approach to implementing this kind of shard replacement with analysis/semantics-based approaches would involve a concept of specification ordering and desirability.

1.1 Efficient Implementations

Many computations have straightforward basic implementations but very complex maximally efficient implementations. Particularly prominent examples include sorting, linear algebra, and linear programming. The availability of powerful but difficult to program hardware such as graphics accelerators can increase the distance between basic and maximally efficient implementations.

Program fracture can enable developers to write a basic implementation, in some cases by simply copying several lines from a textbook. Fracturing their program can expose the basic implementation as one of the shards. Fracturing other programs can expose replacement shards that implement the same functionality (or even approximate versions of the functionality). Replacing the original basic shard with the more efficient shard can automatically improve the performance without the need for the developer to manually investigate the performance, find substitutes, or refine the implementation. One way to estimate the efficiency is to simply run and time the shards on sample inputs.

1.2 Software Transparency and Simplicity

One drawback of optimized code is its complexity and opacity — an efficient autotuned implementation of a basic operation such as FFT can include dynamic code generation and compilation, exploration of multiple alternatives, and multilingual implementations [4]. The opacity can make it difficult to understand the semantics of the code, how it operates, and the role that it plays in the computation. The complexity can induce extraneous dependences that impair portability and the ability to operate on specialized platforms. Systems that dynamically generate and compile different versions of the code (autotuners are a prominent example) may require the presence of a compiler, a significant requirement that not all computing platforms can satisfy. Appropriate fractures could automatically find a simpler, more transparent replacement with fewer

dependences. Enhanced code transparency and portability are but two of the potential benefits.

1.3 Software Aging and Shard Rejuvenation

Of course, optimization can also become counterproductive over time — as the characteristics of the computational platform change, optimizations targeted at previous platforms often persist in the code even though they have no remaining purpose. Too often the result is a complex block of code optimized for an obsolete hardware platform. Another source of counterproductive software aging is obsolete functionality that remains (potentially disabled) in the current version of the software. Like obsolete optimizations, obsolete functionality can add complexity and obscure the purpose and operation of the software. Both obsolete optimization and obsolete functionality are simply examples of how undesirable software changes accrete over time, increasing software complexity and making it difficult to understand or modify the software.

This accretion has known negative consequences — as software systems age, they are known to become increasingly difficult for human developers to maintain [3]. At some point, the code becomes so difficult to change that the probability that a given change will introduce new unacceptable behavior exceeds the probability that the change will accomplish its desired goal.

Program fracture and replacement can help promote software rejuvenation. Identifying more modern, transparent, and relevant replacement implementations for specific shards, then replacing the aging original shard with the younger replacement shard can help extend the useful lifetime of the program.

1.4 Error Handling and Corner Case Code

Error handling and corner case code can be notoriously difficult to develop and get right, in part because the situations in which the code is relevant can be difficult to envision. Program fracture and recombination supports a model of development in which programmers write code that is only designed to handle the common case and elides the error handling and corner case code that can obscure the structure, purpose, and operation of the software. The system can then automatically discover and replace the original shard with a more elaborately developed shard that includes error handling and corner case code. It is also possible to leave the core computation itself in place, then enhance the computation with automatically located and transferred checks from other programs [13, 12].

Note that this approach can support quickly engineered prototype components that are designed to work only in limited contexts or for limited use cases. A linear program solver taken directly from a textbook, for example, may work only for very small linear programs, with phenomena such as numerical instability crippling the solver for larger problems. But a textbook linear program solver can define the desired behavior for small problems, making it possible to identify and use a more intensively engineered replacement linear program

solver. Given that developing a robust solver can take years or even decades [9], the potential productivity enhancements are substantial.

It is even possible to work with shards that implement only a few hard-coded cases or rely on humans to provide the desired functionality.

1.5 Multiple Code Views

One drawback of including error handling and corner case code is that it obscures the structure and purpose of the software, making the software more difficult to understand. One way to attack this problem is to replace the original complex code with a simpler replacement shard that focuses on the common case. The concept is to automatically generate simpler, easier to understand code views with multiple perspectives designed to support multiple different purposes. Instead of a single monolithic code base that includes all of the code and functionality, whether relevant or not, this perspective promotes a fluid set of views designed to satisfy different needs and goals. One particularly noteworthy aspect is the fluid view of program semantics and the recognition that technically incorrect or incomplete programs are, for many purposes, more useful than more correct or complete programs. We have focused here on the elimination of potentially confusing error handling or corner case code as a way to enhance the ability of a developer to identify and work with the important core functionality of the program. But excess error handling and corner case code can also impair performance or introduce undesirable dependences on other software or system components.

1.6 Shard Analyzability

Complex code (whether resulting from optimization, aging, the inclusion of error handling and/or corner case functionality) can significantly impair the static analyzability of the program. Replacing this complex code with simpler shards can improve the analyzability of the program and promote the automatic extraction of useful information that can provide insight into the software and its properties. The success of accurate analysis stubs in enabling the successful information flow analysis of Android applications provides some indication of the potential for improved static analyzability that shard replacement can deliver [5].

One particularly intriguing aspect of program fracture and recombination is the potential it offers to simplify not just the code, but also the analysis results. Simpler code can be easier to analyze, enhancing the precision and scalability of the overall analysis. Sound static analyses reflect the complete semantics of the program. Substituting shards with simpler semantics can simplify the analysis results and enhance compositionality and scalability. Devising static analyses that analyze the program along several different axes (for example, effects on different parts of the program state and different aspects of the semantics) and using the axes to place the resulting analysis result within a projectable space of analysis results can promote the effective elimination of irrelevant analysis

components. In this way the system can deploy even sound program analyses and use the analysis results to drive shard replacement algorithms that work with compatible but not semantically equivalent shards.

1.7 Enhanced Capabilities

Like optimization, capabilities such as distribution or the ability to operate safely in parallel contexts can require significant development effort. Program fracture and recombination can enable developers to write simple, less capable versions of core functionality, then automatically replace this functionality with more extensively engineered versions that can operate successfully in new contexts or exploit resources available in specialized contexts.

1.8 Functionality Elimination via Shard Removal

Programs may often contain components that have undesirable functionality. Security vulnerabilities [11], the ability to process a larger than desired set of inputs, logging or error reporting code that interacts with an obsolete subsystem, or even certain features such as support for SNMP that may no longer be desirable [10] can all be seen as undesirable functionality. Replacing the shard that implements the undesirable functionality with the null shard that does nothing can improve the program by eliminating the undesirable functionality. It may also be possible, of course, to find other replacement shards that implement the desired functionality but not the undesired functionality.

1.9 Functionality Enhancement via Shard Insertion

It is also possible to transfer shards directly into programs to obtain software hybrids with the combined functionality of both programs. This insertion would take a shard or shards from a donor and insert them into a recipient at an appropriate insertion point. There would be no replacement shard — the new shard or shards would enhance the recipient with new functionality not present before the insertion of the shard. CodePhage is one system that implements this shard transfer capability [13, 12].

1.10 Shards as Specifications and Code Search

Program fracture and recombination promotes the view of shards as specifications. It is often straightforward to code up enough of a computation to enable the program fracture and replacement system to find an appropriate replacement shard. In this way shards can serve as specifications that enable effective code search.

1.11 Dependence Elimination

One of the primary goals of program fracture is to obtain shards that are freed from dependences they may have had on parts of the original program that

should not be transferred with the shard. One way to realize this goal is to break the program into shards in such a way that each shard omits these dependences. In some cases this may require an intelligent fracture process that explicitly finds dependences on undesirable parts of the program, then eliminates the code that creates the dependence. The fracture may simply create a stub that breaks the dependence while still enabling the rest of the code with transitive dependences to execute, it may trace out and remove code that transitively depends on the excised dependence, or it may employ some combination of the two techniques.

2 Case Study

We have implemented a protoype program fracture and recombination system. This system fractures applications up into individual shards, with one shard for each procedure in the application. There is a program for each shard — the program reads in inputs for the parameters of the shard's procedure, invokes the procedure, then prints the result that the procedure returns and the (potentially updated) values of the parameters.

The current implementation is designed to work with multiple programs. It finds shards whose parameter types match, one from each program, then automatically generates sample parameter values, invokes each shard on the sample values, and records the execution times and resulting return and updated parameter values. It then uses this data to find set of semantically interchangable shards — i.e., sets of shards that have (if desired, approximately) the same observed input/output relation. It uses the recorded execution times to find the most efficient shard, then replaces every other interchangable shard with this shard.

We applied our prototype implementation to two programs. The first reads in a file of salary information, then prints the (anonymized) top ten salaries. Figure 1 presents the relevant source code.

```
void doSalaries() {
  int next;
  next = 0;

  for (next = 0; next < 1048576; next++) {
    if (scanf("%4095s", first) == EOF) break;
    if (scanf("%4095s", last) == EOF) break;
    if (scanf("%d", salaries+next) == EOF) break;
  }
  quicksort(salaries, next);
  for (int i = next - 1; (i >= 0) && (next-i < 10); i--) {
    printf("%d\n", salaries[i]);
  }
}
```

Figure 1: Salary Program

The second program is designed to compact a set of objects stored persistently in a file. Over time the file may become fragmented as objects are allocated and deallocated. The program reads in an index. The first line of the index specifies the file containing the objects and the file to write that contains the compacted objects. The remaining lines of the index contain pairs of numbers. Each pair specifies the starting and ending offsets of the object within the file. The program reads in the offsets, sorts them, then uses the sorted offsets to read the objects in and write out the compacted file. Figure 2 presents the relevant source code.

In this case, the `quicksort` shard from the salary application exhibits superior performance to the `sort` shard from the object compaction application (which uses bubble sort). So our implementation automatically replaces the `sort` in the object compaction implementation with the quick sort from the salary application.

3 Compatible Shards and Adapters

Our current implementation works with compatible shards that have the same type signatures. There are several ways to generalize this technique to include more replacement shards. In general, we expect the automatic generation of adapters and shims that enable replacement shards that are not immediately precisely compatible. These techniques can also enable the automatic insertion of shards into contexts that require some adaptation before the shard can be automatically inserted [8].

Explicit Polymorphism: Some programming language constructs immediately support this kind of replacement. The use of explicitly polymorphic constructs such as C++ templates, for example, promotes the construction of specializable shards that can work in a variety of contexts. The use of parameter order canonicalization (by, for example, defining a total order on program types and working with adapters that expose the shard interface in that total order) can enable parameter redordering adapters that again enable the discovery and replacement of shards with otherwise incompatible parameter orders.

Data Structure Translators: The next step is data structure translation adapters. One example is adapters that perform array index translations to enable replacement shards to work with different array indexing schemes. One class of adapters changes the way the arrays are stored in memory, leaving the indexing code of the replacement shard in place. Another class leaves the storage scheme in place but statically analyzes and transforms the code of the replacement shard to work with the storage scheme of the program into which it is inserted.

It is also possible to leverage abstract data types — many abstract data types offer implementations of equivalent semantics with different performance or other tradeoffs. In some cases the developer may have used only a subset of the functionality of one abstract data type that could be implemented more efficiently by a less general but more efficient other abstract data type. In all of

```
void doCompact() {
  int ifd, ofd;
  int next;

  scanf("%4095s %4095s\n", ifn, ofn);

  for (next = 0; next < 1048576; next++) {
    if (scanf("%d", begin+next) == EOF) break;
    if (scanf("%d", end+next) == EOF) break;
  }

  sort(begin, next);
  sort(end, next);

  ifd = open(ifn, O_RDONLY);
  ofd = open(ofn, O_CREAT | O_TRUNC | O_WRONLY, S_IRUSR | S_IWUSR);
  int current = 0;

  if (ifd >= 0 && ofd >= 0) {
    printf("%s\n", ofn);
    for (int i = 0; i < next; i++) {
      if (lseek(ifd, begin[i], SEEK_SET) == begin[i]) {
        if (end[i]-begin[i] < 1048576) {
          int len = end[i]-begin[i];
          if (read(ifd, buf, len) == len) {
            printf("%d %d\n", current, current + len);
            if (write(ofd, buf, len) == len) {
              current = current + len;
            } else {
              fprintf(stderr, "compact failed\n");
              exit(1);
            }
          }
        }
      }
    }
  }
}
```

Figure 2: Object Compaction Program

these cases the recombination can use the data structure semantics to recognize compatible shards from different data structures, with each shard in this case consisting of subsets of the abstract data types. This example shows that shards do not need to consist of single procedures or methods — they can be modules, related groups of code, or even arbitrary otherwise disconnected pieces of code extracted from single or multiple applications.

Specification-Based Compatibility: It is also possible to base compatability on specifications, either provided by the developer as part of the software development process or automatically inferred by an analysis (either static or dynamic). Shards with the same specification are interchangable. Factoring the specifications along different axes, where each axis captures a different aspect of the behavior of the shard such as side effects on different parts of the system, input or output effects, logging effects, or various aspects of the semantic properties of the shard.

4 External Dependences

Shards may often have external dependences on functionality present in their original context but not present in their new context. Examples include invoked modules and externally visible state such as global variables. These dependences may require importing into the recipient and/or conflict resolution along with appropriate initialization to operate properly.

It is possible to use program analysis to automatically identify and extract accessed global variables. The global variable may be initialized to an appropriate value either by program analysis or by recording values that appear in executions of one of the programs. Note that the program fracture and recombination system may need to pull in data structures such as hash tables and trees that store auxiliary data upon which the shard depends. Another approach is to trace the code in the original program that initializes the global variables or other state upon which the shard depends, then transfer and insert that code as well. Note that the process may be recursive — the initialization code may access files or other external resources. In that case the program fracture and repair system can transfer the files or external resources, simulate the effect they have on the system via a spoofing or record/replay system, or even simply record the data structures that are constructed as a consequence of interacting with the external resources. In general, there is a chain of dependences from the resource interaction to the end result on the system, and it is possible for the program fracture and repair system to intercept the dependences anywhere along the chain and reconstruct (or alternatively potentially transitively excise) the effects.

The shard itself may also have a direct dependence on configuration files, normal files, or other external resources. It is possible to use the same techniques described above to deal with these dependences.

A more challenging situation occurs with dependences mediated via the operating system or some other opaque system. For example, some packages or

system calls require initialization to operate successfully, but the dependences are controlled by data structures stored in the operating system or some other opaque system. Here the program fracture and recombination system may need some external knowledge of the dependence relationship so that it can, for example, find and replay the appropriate initialization calls, synthesize an appropriate initialization sequence, or do a search to find the sequence in the original program.

The shard may also have dependences on specific hardware components unavailable in other contexts or other unmovable resources. In this case the program fracture and recombination system can simply excise the dependence. In general, such dependence excision can be incorporated into arbitrary parts of the system and not just for hardware dependences. One of the goals of program fracture and recombination is to free useful pieces of code from dependences on the original context in which it appeared. Dependence excision, potentially following the dependences to completely eliminate the effects when appropriate, is therefore a key element of program fracture and recombination.

5 Conclusion

We now have decades of investment in software systems, with desired functionality often implemented multiple times at varying levels of quality, performance, and capabilities. In principle, the vast majority of the functionality that most programs need has already been implemented. Our ability to profitably find and combine relevant pieces of functionality, as opposed to our ability to develop new software, may ideally become the limiting factor in software development. Program fracture and recombination promises to significantly enhance our potential in this critical area.

References

[1] Karen Ambrose, Albrecht Koppenhofer, and Faith Belanger. Horizontal gene transfer of a bacterial insect toxin gene into the epichloe fungal symbionts of grasses. *Scientific Reports*, 4, July 2014.

[2] Miriam Barlow. What Antimicrobial Resistance Has Taught Us About Horizontal Gene Transfer. *Methods in Molecular Biology*, 532:397–411, 2009.

[3] Laszlo A. Belady and M. M. Lehman. A model of large program development. *IBM Systems Journal*, 15(3):225–252, 1976.

[4] Matteo Frigo. A fast fourier transform compiler. In *Proceedings of the 1999 ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI), Atlanta, Georgia, USA, May 1-4, 1999*, pages 169–180, 1999.

[5] Michael I. Gordon, Deokhwan Kim, Jeff Perkins, Limei Gilham, Nguyen Nguyen, and Martin Rinard. Information-flow analysis of Android applications in DroidSafe. In *Proceedings of the 22nd Annual Network and Distributed System Security Symposium (NDSS'15)*, 2015.

[6] Mark A. Kay, Joseph C. Glorioso, and Luigi Naldini. Viral vectors for gene therapy: the art of turning infectious agents into vehicles of therapeutics. *Nat Med*, 7(1):33–40, January 2001.

[7] Patrick J. Keeling and Jeffrey D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8), 8 2008.

[8] Fan Long and Martin Rinard. Staged Program Repair in SPR. Technical Report MIT-CSAIL-TR-2015-008, 2015.

[9] Bruce Murtagh and Michael Saunders. MINOS 5.51 user's guide. Technical Report SOL-83-20R, 2003.

[10] Huu Hai Nguyen and Martin C. Rinard. Detecting and eliminating memory leaks using cyclic memory allocation. In *Proceedings of the 6th International Symposium on Memory Management, ISMM 2007, Montreal, Quebec, Canada, October 21-22, 2007*, pages 15–30, 2007.

[11] Martin Rinard. Manipulating program functionality to eliminate security vulnerabilities. *Advances in Information Security*, 54, July 2011.

[12] Stelios Sidiroglou-Douskos, Eric Lahtinen, Fan Long, and Martin Rinard. Automatic error elimination by multi-application code transfer. Technical Report MIT-CSAIL-TR-2014-024, August 2014.

[13] Stelios Sidiroglou-Douskos, Eric Lahtinen, Fan Long, and Martin Rinard. Automatic error elimination by horizontal code transfer across multiple applications. In *Proceedings of the 2015 ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI)*, June 2015.