

# Reducing Code Size with Function Merging

*Rodrigo Caetano de Oliveira Rocha*



Doctor of Philosophy  
Institute of Computing Systems Architecture  
School of Informatics  
University of Edinburgh  
2020



# Abstract

# Acknowledgements

TODO.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. Some of the material used in this thesis has been published in the following papers:

- Rodrigo Rocha, Pavlos Petoumenos, Zheng Wang, Murray Cole, and Hugh Leather. “Function merging by sequence alignment.” In IEEE/ACM International Symposium on Code Generation and Optimization (CGO), pp. 149-163. Best Paper Award. 2019.
- Rodrigo Rocha, Pavlos Petoumenos, Zheng Wang, Murray Cole, and Hugh Leather. “Effective function merging in the SSA form.” In ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI), pp. - . 2020.

*(Rodrigo Caetano de Oliveira Rocha)*



# Table of Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Sources of Similarities . . . . .	1
<b>2</b>	<b>Background</b>	<b>3</b>
2.1	Compiler Infrastructure . . . . .	3
2.1.1	Link-Time Optimisations . . . . .	4
2.2	Optimisation Scope . . . . .	4
2.2.1	Interprocedural Optimisations . . . . .	4
2.3	Sequence Alignment . . . . .	4
<b>3</b>	<b>Related Work</b>	<b>7</b>
3.1	Merging Identical Object Code . . . . .	7
3.2	Identical Function Merging . . . . .	8
3.3	Merging Beyond Identical Functions . . . . .	11
3.4	Code Factoring . . . . .	12
3.5	Code Similarity . . . . .	12
<b>4</b>	<b>Merging Pairs of Functions</b>	<b>15</b>
4.0.1	Linearisation . . . . .	16
4.0.2	Sequence Alignment . . . . .	17
4.0.3	Equivalence Relation . . . . .	17
4.0.4	Control-Flow Graph Generation . . . . .	20
4.0.5	Operand Assignment . . . . .	22
4.0.6	Preserving the Dominance Property . . . . .	24
4.0.7	Phi-Node Coalescing . . . . .	25
<b>5</b>	<b>Optimisation Strategy</b>	<b>29</b>
5.1	Profitability Cost Model . . . . .	29

5.2	Exhaustive Search . . . . .	30
5.3	Focusing on Profitable Functions . . . . .	30
5.3.1	Link-Time Optimisation . . . . .	33
<b>6</b>	<b>Reducing Runtime Overhead</b>	<b>35</b>
6.1	Conservative Function Merging . . . . .	35
6.2	Profile-Guided Function Merging . . . . .	35
<b>7</b>	<b>Conclusion</b>	<b>37</b>
	<b>Bibliography</b>	<b>39</b>



# Chapter 1

## Introduction

### 1.1 Sources of Similarities

Note that functions that are identical at the IR or machine level are not necessarily identical at the source level. Figure 3.2a shows two real functions extract from the 447.dealII program in the SPEC CPU2006 [12] benchmark suite. Although these two functions are not identical at the source level, they become identical after a template specialization and some optimizations are applied, in particular, constant propagation, constant folding, and dead-code elimination. Specializing `dim` to 1 enables to completely remove the loop in the function `PolynomialSpace`. Similarly, specializing `dim` to 1 results in only the first iteration of the loop in the function `TensorProductPolynomials` being executed. The compiler is able to statically analyze and simplify the loops in both functions, resulting in the identical functions shown at the bottom of Figure 3.2a.

```
template <int dim>
unsigned int PolynomialSpace<dim>::
compute_n_pols (const unsigned int n) {
    unsigned int n_pols = n;
    for (unsigned int i=1; i<dim; ++i) {
        n_pols *= (n+i);
        n_pols /= (i+1);
    }
    return n_pols;
}

----- After template specialization and applying optimizations:-----
unsigned int PolynomialSpace<1>::
compute_n_pols(const unsigned int n) {
    return n;
}

template <int dim> inline
unsigned int TensorProductPolynomials<dim>::
x_to_the_dim (const unsigned int x) {
    unsigned int y = 1;
    for (unsigned int d=0; d<dim; ++d) {
        y *= x;
    }
    return y;
}

unsigned int TensorProductPolynomials<1>::
x_to_the_dim(const unsigned int x) {
    return x;
}
```

Figure 1.1: Two function extracted from the 447.dealII benchmark that are not identical at the source level, but after applying template specialization and optimizations they become identical at the IR level.

Identical code is particularly common in C++ programs with heavy use of *parametric polymorphism*, via template or *auto* type deduction.

# Chapter 2

## Background

### 2.1 Compiler Infrastructure

Compilers are programming tools responsible for translating programs in a given source language to a lower-level target language. Compilers must preserve the program semantics, i.e., both This compilation process must preserve the program semantics, as specified by the source language, while also

Compilers are usually organised in *three-phases*: frontend, optimiser, and backend. The frontend is responsible for parsing, validating and diagnosing errors in the source code. This parsed source code is then translated into an intermediate representation, which is the LLVM IR in this case. The optimiser is responsible for doing a broad variety of transformations, that are usually independent of language and target machine, to improve the code's performance. The backend, also known as the code generator, then translates the code from the intermediate representation onto the target instruction set. It is common for the backend to also perform some low-level optimisations that take advantage of unusual features of the supported architecture.



Figure 2.1: Overview of the three-phase compiler infrastructure.

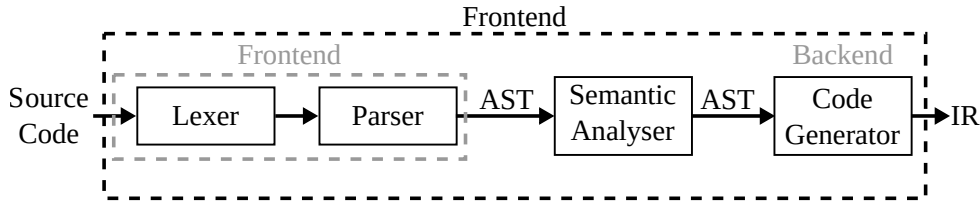


Figure 2.2: Overview of the three-phase compiler infrastructure.

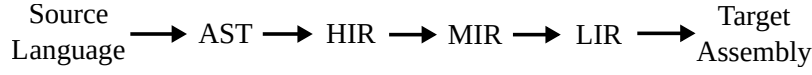


Figure 2.3: Overview of the three-phase compiler infrastructure.

### 2.1.1 Link-Time Optimisations

## 2.2 Optimisation Scope

### 2.2.1 Interprocedural Optimisations

## 2.3 Sequence Alignment

Sequence alignment is important in many scientific areas, most notably in molecular biology [3, 9, 11, 15] where it is used for identifying homologous subsequences of amino acid in proteins. Essentially, sequence alignment algorithms insert blank characters in both input sequences so that the final sequences end up having the same size, where equivalent segments are aligned with their matching segments from the other sequence and non-equivalent segments are paired with blank characters.

Formally, sequence alignment can be defined as follows: For a given alphabet  $\alpha$ , a sequence  $S$  of  $k$  characters is an element of  $\alpha^k$ , i.e.,  $S = (a_1, \dots, a_k)$ . Let  $S_1, \dots, S_m$  be a set of sequences, possibly of different lengths but all derived from the same alphabet  $\alpha$ , where  $S_i = (a_1^{(i)}, \dots, a_{k_i}^{(i)})$ , for all  $i \in \{1, \dots, m\}$ . Consider an extended alphabet that includes the *blank* character “–”, i.e.,  $\beta = \alpha \cup \{-\}$ . An alignment of the  $m$  sequences,  $S_1, \dots, S_m$ , is another set of sequences,  $\bar{S}_1, \dots, \bar{S}_m$ , such that each sequence  $\bar{S}_i$  is obtained from  $S_i$  by inserting blanks in positions where some of the other sequences have non-blank and possibly equivalent characters, for a given equivalence relation. All sequences  $\bar{S}_i$  in the alignment set have the same length  $l$ , where  $\max\{k_1, \dots, k_m\} \leq l \leq k_1 + \dots + k_m$ . Moreover,  $\forall i \in \{1, \dots, m\}$ ,  $\bar{S}_i = (b_1^{(i)}, \dots, b_l^{(i)})$ , there are increasing functions  $v_i : \{1, \dots, k_i\} \rightarrow \{1, \dots, l\}$ , such that:

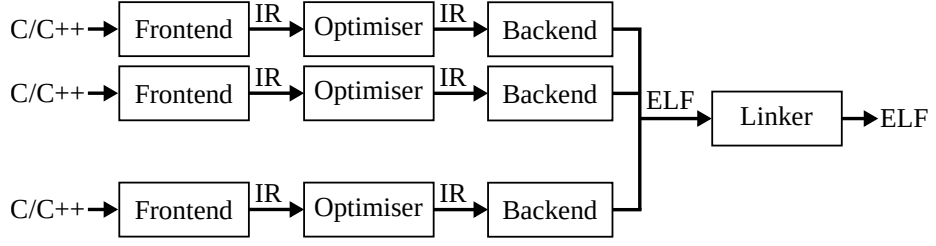


Figure 2.4: Overview of the three-phase compiler infrastructure.

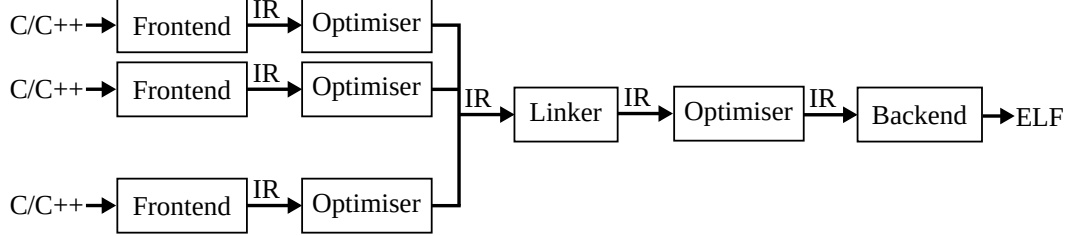


Figure 2.5: Overview of the three-phase compiler infrastructure.

- $b_{v_i(j)}^{(i)} = a_j^{(i)}$ , for every  $j \in \{1, \dots, k_i\}$ ;
- any position  $j$  not mapped by the function  $v_i$ , i.e., for all  $j \in \{1, \dots, l\} \setminus \text{Im } v_i$ , then  $b_j^{(i)}$  is a blank character.

Finally, for all  $j \in \{1, \dots, l\}$ , there is at least one value of  $i$  for which  $b_j^{(i)}$  is not a blank character. Note that two aligned sequences may contain both non-blank and non-equivalent characters at any given position, in which case there is a mismatch.

Specifically for function-merging, we are concerned with the alphabet consisting of all possible typed instructions and labels. Every linearized function represents a sequence derived from this alphabet. We explain the equivalence relation used for this alphabet in the next section. Although we only consider pair-wise alignments, the technique would also work for multi-sequences.

Our work uses the Needleman-Wunsch algorithm [9] to perform sequence alignment. This algorithm gives an alignment that is guaranteed to be optimal for a given scoring scheme [6], however, other algorithms could also be used with different performance and memory usage trade-offs [3, 5, 9, 11]. Different alignments would produce different but valid merged functions.

The Needleman-Wunsch algorithm [9] is based on dynamic programming and consists of two main steps. First, it builds a *similarity matrix*, based on a scoring scheme, which assigns weights for matches, mismatches, and *gaps* (blank characters). Afterwards, a backward traversal is performed on the similarity matrix, in order to recon-

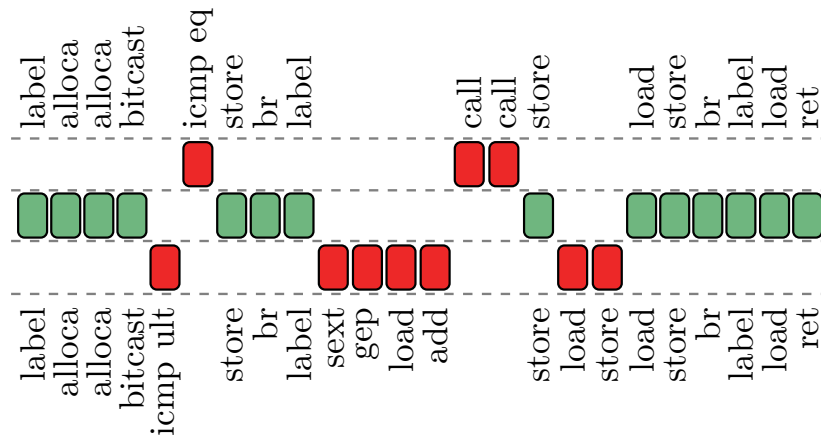


Figure 2.6: The sequence alignment between two functions, identifying the equivalent segments of code (green in the center) and the non-equivalent ones (red at the sides).

struct the final alignment by maximizing the total score. We use a standard scoring scheme for the Needleman-Wunsch algorithm that rewards matches and equally penalizes mismatches and gaps.

# Chapter 3

## Related Work

### 3.1 Merging Identical Object Code

The simplest way of merging identical functions is by looking at their object code, during link time. *Identical code folding* (ICF) is an optimisation that identifies and merges two or more read-only sections, typically functions, that have identical contents. This optimisation is commonly found in major linkers, such as *gold* [7, 13], LLVM’s *lld*, and the MSVC linker [1].

Figure 3.2 shows an example, adapted from Tallam et al. [13], of how generic programming in C++ can lead to identical functions in the object file. The C++ code in in Figure ?? presents a simple *template class* and its member function being instantiated multiple times with different pointer types. Figure ?? shows the object code targeting the Intel x86 architecture. For each instantiation of *Foo*, a replica of its member function *getElement* is created. Because the size of the different pointer types is the same, all replicas of *getElement* are identical in the object file, which can be easily confirmed by comparing their binary representation. as shown in Figure ??.

Figure 3.1: Example showing how a member function of a *template class* can produce code replication susceptible to *identical code folding*. For each instantiation of *Foo*, a replica of the function *getElement* is created for the template instance. When instantiated with pointer types, the object code of these functions will be identical.

Most object file formats, such as the *Executable and Linkable Format* (ELF) [7, 13], are structured as separate sections of content, each section containing a certain type of content. The main types are code segment, different types of data segment,

and relocation information. Relocation information describes how to modify other sections, connecting symbolic references to their definition. In other words, it assigns actual addresses for position-dependent code and data. For example, when a program calls a function, the associated call instruction must transfer control to the proper destination address at execution.

It is common practice for linkers to place functions in separate sections, as exemplified in Figure ???. Therefore, merging identical functions can be generalised to the problem of merging identical sections. Two sections are considered identical if they have the identical section flags, data, code, and relocations. Two relocations are considered the identical if they have the same relocation types, values, and if they point to the same or identical sections.

Since this equality has a cyclic definition, ICF is defined as a fixed-point computation, i.e., it is applied repeatedly until a convergence is obtained. There are two approaches with distinct trade-offs. (i) The pessimistic approach starts with all sections marked as being different and then repeatedly compare them trying to prove their equality, grouping those found to be identical, including their relocations. This approach is implemented in the widely used *gold* linker. (ii) The optimistic approach starts with all functions marked as potentially identical and then repeatedly compare trying to disprove their equality, partitioning those found to be different. This approach is implemented in LLVM's linker, *lld*.

## 3.2 Identical Function Merging

A similar optimisation for merging identical functions, but instead at the intermediate representation (IR) level, is also offered by both GCC and LLVM [2, 8]. This optimisation is only flexible enough to accommodate simple type mismatches provided they can be bitcasted in a losslessly way.

A very strict function comparator is used to identify if two functions are semantically equivalent. First it compares the signature and other general attributes of the two functions. The functions must have identical signature, i.e., the same return type, the same number of arguments, and exactly the same list of argument types. Then this function comparator performs a simultaneous walk, in depth first order, in the functions' control-flow graphs. This walk starts at the entry block for both functions, then takes each block from each terminator in order. As an artifact, this also means that unreachable blocks are ignored. Finally, it iterates through each instruction in each



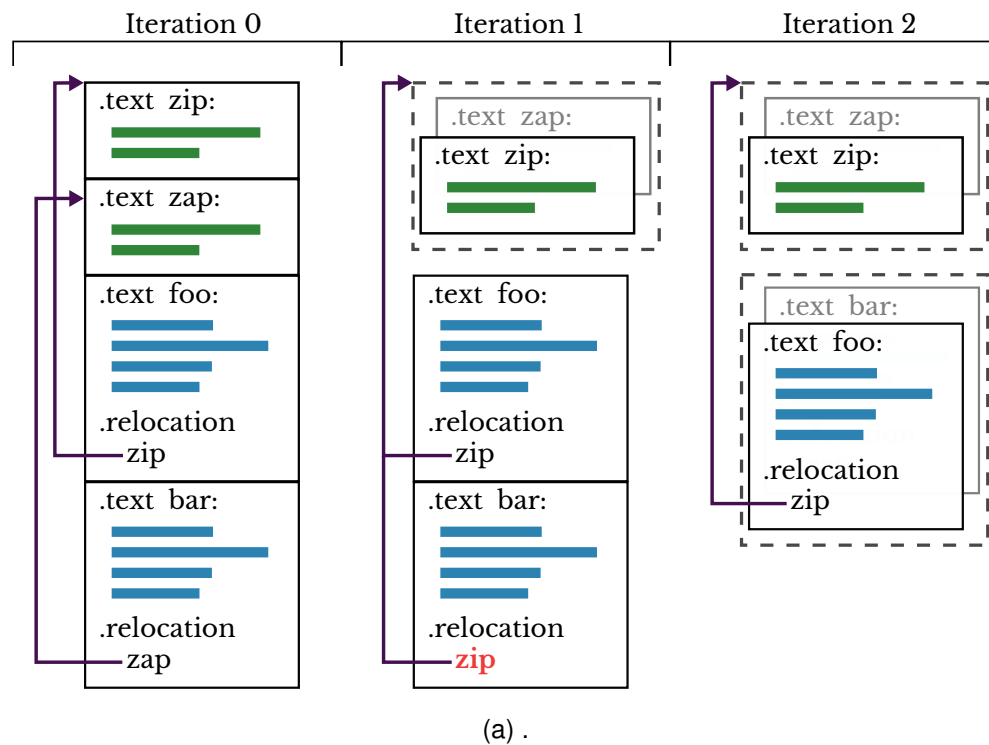


Figure 3.2: Example showing how a member function of a *template class* can produce code replication susceptible to *identical code folding*. For each instantiation of *Foo*, a replica of the function *getElement* is created for the template instance. When instantiated with pointer types, the object code of these functions will be identical.

basic block. Two blocks are equivalent if they have equivalent instructions in exactly the same order, without excess. The comparator always fails conservatively, erring on the side of claiming that two functions are different.

When a pair of equivalent functions is identified, we can create either an alias or a *thunk*. Aliasing entails eliminating one of the functions and replacing all its callsites to the other function. Thunks must be created when neither of the equivalent functions can be eliminated by aliasing. In such case, a thunk is created for either one of the functions, replacing its body by a call to the other function, which allows all callsites and name references to both functions to be preserved. Aliasing is preferred since it is cheaper and adds no runtime overhead. The appropriate merging is applied according to following rules:

- If the address of at least one function is not taken, alias can be used.
- But if the function is part of COMDAT section that can be replaced, we must use thunk.

- If we create a thunk and none of functions is writeable, we can redirect calls instead.

Although very restrictive, this optimisation guarantees that any pair of mergeable functions will result in code size reduction with no performance overhead.

Its simplicity also allows for an efficient exploration approach based on computing a hash of the functions and then using a binary tree to identify equivalent functions. Since hashing is cheap to compute, it allows us to efficiently group possibly equivalent functions and filter out functions that are obviously unique. This hash must have the property that if function  $F = G$  according to the comparison function, then  $hash(F) = hash(G)$ . Therefore, as an optimisation, two functions are only compared if they have the same hash. This consistency property is critical to ensuring all possible merging opportunities are exploited. Collisions in the hash affect the speed of the pass but not the correctness or determinism of the resulting transformation.

A function hash is calculated by considering only the number of arguments and whether a function is varargs, the order of basic blocks (given by the successors of each basic block in depth first order), and the order of opcodes of each instruction within each of these basic blocks. This mirrors the strategy of `compare()` uses to compare functions by walking the BBs in depth first order and comparing each instruction in sequence. Because this hash does not look at the operands, it is insensitive to things such as the target of calls and the constants used in the function, which makes it useful when possibly merging functions which are the same modulo constants and call targets.

All functions can be sorted based on their hash value, which ends up grouping possibly equivalent functions together. If the hash value of a given function matches any of its adjacent values in the sorted list, this function must be considered for merging. Functions with a unique hash value can be easily ignored since no other function will be found equivalent.

The functions that remain are inserted into a binary tree, where functions are the node values themselves. An order relation is defined over the set of functions. We need total-ordering, so we need to maintain four properties on the functions set:

- $a \leq a$  (reflexivity);
- if  $a \leq b$  and  $b \leq a$  then  $a = b$  (antisymmetry);
- if  $a \leq b$  and  $b \leq c$  then  $a \leq c$  (transitivity);
- for all  $a$  and  $b$ ,  $a \leq b$  or  $b \leq a$  (totality).

This total-ordering was made through special function comparison procedure that returns:

- 0 when functions are semantically equal,
- -1 when Left function is less than right function, and
- 1 for opposite case.

Functions are kept on binary tree. For each new function  $F$  we perform lookup in binary tree.

### 3.3 Merging Beyond Identical Functions

The state-of-the-art function-merging technique exploits structural similarity among functions [4]. Their optimization is able to merge similar functions that are not necessarily identical. Two functions are structurally similar if both their function types are equivalent and their CFGs isomorphic. Two function types are equivalent if they agree in the number, order, and types of their parameters as well as their return types, linkage type, and other compiler-specific properties. In addition to the structural similarity of the functions, their technique also requires that corresponding basic blocks have exactly the same number of instructions and that corresponding instructions must have equivalent resulting types. Mergeable functions are only allowed to differ in corresponding instructions, where they can differ in their opcodes or in the number and type of their input operands.

Because the state-of-the-art is limited to functions with identical CFGs and function types, once it merges a pair of functions, a third *similar* function cannot be merged into the resulting merged function since they will differ in both CFGs and their lists of parameters. Due to this limiting factor, the state-of-the-art has to first collect all mergeable functions and merge them simultaneously.

The state-of-the-art algorithm iterates simultaneously over corresponding basic blocks in the set functions being merged, as they have isomorphic CFGs. For every basic block, if their corresponding instructions have different opcodes, they split the basic block and insert a switch branch to select which instruction to execute depending on a function identifier. Because these instructions have equivalent resulting types, their results can be merged using a phi-operator, which can then be used transparently as operands by other instructions.

Although the state-of-the-art technique improves over LLVM’s identical function merging, it is still unnecessarily limited. In Section ??, we showed examples of very similar real functions where the state-of-the-art fails to merge. Our approach addresses such limitations improving on the state-of-the-art across the board.

### 3.4 Code Factoring

Code factoring is a related technique that addresses the same fundamental problem of duplicated code in a different way. Code factoring can be applied at different levels of the program [? ]. Local factoring, also known as local code motion, moves identical instructions from multiple basic blocks to either their common predecessor or successor, whenever valid [? ? ? ]. Procedural abstraction finds identical code that can be extracted into a separate function, replacing all replicated occurrences with a function call [? ? ].

Procedural abstraction differs from function merging as it usually works on single basic blocks or single-entry single-exit regions. Moreover, it only works for identical segments of code, and every identical segment of code is extracted into a separate new function. Function merging, on the other hand, works on whole functions, which can be identical or just partially similar, producing a single merged function.

However, all these techniques are orthogonal to the proposed optimization and could complement each other at different stages of the compilation pipeline.

### 3.5 Code Similarity

Code similarity has also been used in other compiler optimizations or tools for software development and maintenance. In this section, we describe some of these applications.

Coutinho et al. [? ] proposed an optimization that uses instruction alignment to reduce divergent code for GPU. They are able to fuse divergent branches that contain single basic blocks, improving GPU utilization.

Similarly, analogous algorithms have also been suggested to identify the differences between two programs, helping developers with source-code management and maintenance [? ? ]. These techniques are applied in tools for source-code management, such as the *diff* command [? ].

Similar techniques have also been applied to code editors and IDEs [? ? ]. For example, SourcererCC [? ] detects possible clones, at the source level, by dividing

the programs into a set of code blocks where each code block is itself represented by a bag-of-tokens, i.e., a set of tokens and their frequencies. Tokens are keywords, literals, and identifiers, but not operators. Code blocks are considered clones if their degree of similarity is higher than a given threshold. In order to reduce the number of blocks compared, candidate blocks are filtered based on a few of their tokens where at least one must match.

Our ranking mechanism uses an approach similar to SourcererCC, where we use opcode frequencies and type frequencies to determine if two functions are likely to have similar code. However, we need a precise and effective analysis of code similarity when performing the actual merge. To this end, we use a sequence alignment technique.



# Chapter 4

## Merging Pairs of Functions

In this section, we describe the proposed function-merging technique. Contrary to the state-of-the-art, our technique is able to merge any two functions. If the two functions are equivalent, i.e., identical, then the two functions can be completely merged into a single identical function. However, if the two functions differ at any point, an extra parameter is required, so that the caller is able to distinguish between the functions. The two functions can differ in any possible way, including their list of parameters or return types. If the lists of parameters are different, we can merge them so that we are able to uniquely represent all parameters from both functions. If the return types are different, we can use an aggregate type to return values of both types or return just the non-void type if the other one is void.

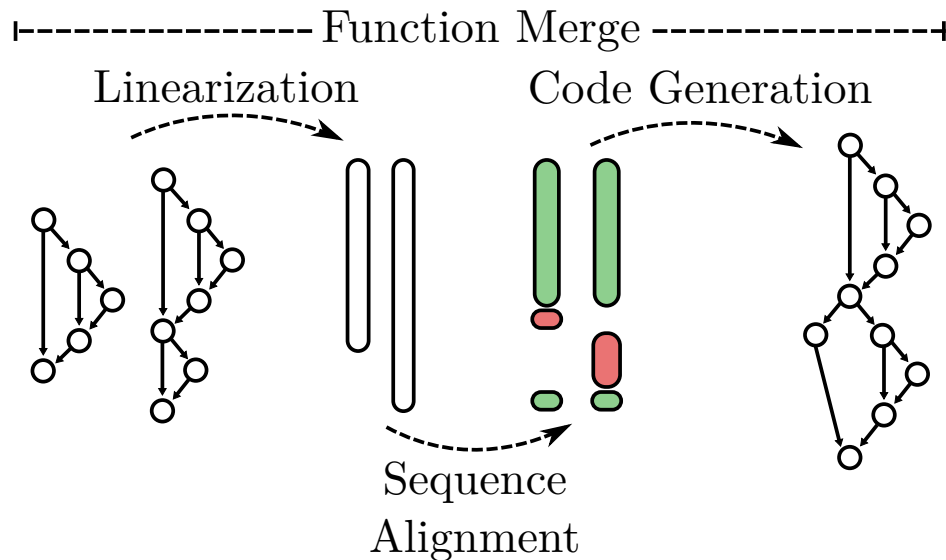


Figure 4.1: Overview of our function-merging technique.

The proposed technique consists of three major steps, as depicted in Figure 4.1.

First, we linearise each function, representing the CFG as a sequence of labels and instructions. The second step consists in applying a sequence alignment algorithm, borrowed from bioinformatics, which identifies regions of similarity between sequences. The sequence alignment algorithm allows us to arrange two linearised functions into segments that are equivalent between the two functions and segments where they differ from one another. The final step performs the code generation, actually merging the two functions into a single function based on the aligned sequences. Aligned segments with equivalent code are merged, avoiding redundancy, and the remaining segments where the two functions differ have their code guarded by a function identifier. At this point, we also create a merged list of parameters where parameters of the same type are shared between the functions, without necessarily keeping their original order. This new function can then be used to replace the original functions, as they are semantically equivalent, given the appropriate function-identifier parameter.

#### 4.0.1 Linearisation

The *linearisation*<sup>1</sup> of a function consists in specifying an ordering of the basic blocks based on a traversal of the CFG and then producing a sequence of basic block labels and instructions, similar to a textual representation of the function. Although this operation is trivial, the specific ordering of the basic blocks chosen can have an impact on the merging operation.

In our implementation, the linearisation uses a reverse post-order (RPO) of the basic blocks, following a canonical ordering of the successors. The RPO guarantees that the linearisation starts with the entry basic block and then proceeds favoring definitions before uses, except in the presence of loops. Although the specific ordering produced by the canonical linearisation may not be optimal, it is common practice for compilers to rely on prior canonicalisations, e.g., canonical loops, canonical induction variables, canonical reassociation, etc. For contrast, if, instead, we use an RPO linearisation with a uniformly randomised ordering of the successor basic blocks, the final code-size reduction of the function-merging optimisation can drop up to 10% for individual benchmarks. Note that our decision for using the canonical RPO is purely pragmatic and other orderings of the basic blocks could also be used, as long as it produces a sequence of labels followed by instructions.

---

<sup>1</sup>Although linearisation of CFGs usually refers to a predicated representation, in this paper, we refer to a simpler definition.



## 4.0.2 Sequence Alignment

When merging two functions, the goal is to identify which pairs of instructions and labels that can be merged and which ones need to be selected based on the actual function being executed. To avoid breaking the semantics of the original program, we also need to maintain the same order of execution of the instructions for each one of the functions.

To this end, after linearisation, we reduce the problem of merging functions to the problem of *sequence alignment*. Figure 4.2 shows an example of the sequence alignment between two linearised functions extracted from the `400.perlbench` benchmark in SPEC CPU2006 [12].

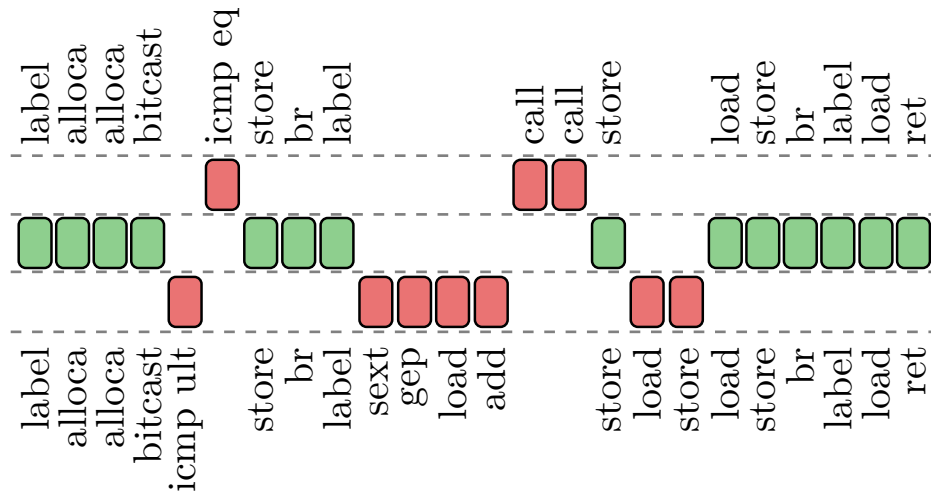


Figure 4.2: The sequence alignment between two functions.

Particularly for the function-merging, we are concerned with the alphabet consisting of all possible typed instructions and labels. Every linearised function represents a sequence derived from this alphabet. We explain the equivalence relation used for this alphabet in the next section.

## 4.0.3 Equivalence Relation

We describe the equivalence relation between values in two separate cases, namely, the equivalence between instructions and the equivalence between labels.

Labels can represent both normal basic blocks and landing blocks, which are used in exception handling code. Labels of normal basic blocks are always considered equivalent but landing blocks must have exactly the same landingpad instructions.

Two instructions are equivalent if: (1) their opcode are semantically equivalent, but not necessarily the same; (2) they both have equivalent types; and (3) they have pairwise operands with equivalent types. Types are considered equivalent if they can be bitcasted in a losslessly way from one to the other. It is also important to make sure that there is no conflict regarding memory alignment when handling pointers. No additional restriction is imposed on the operands of the two instructions being compared for equivalence. Whenever two operands cannot be statically proved to represent the same value, a select instruction is used to distinguish between the execution of two functions being merged. For function calls, the type equivalence requires that both instructions have identical function types, i.e., both called functions must have an identical return type and an identical list of parameter types.

#### 4.0.3.1 Handling Exception Handling Code

Most modern compilers implement the zero-cost Itanium ABI for exception handling [?], including GCC and LLVM, sometimes called the *landing-pad* model. In this section, we describe restrictions imposed by exception handling code and their equivalence relation.

The invoke instruction co-operates tightly with its landing block, i.e., the basic block pointed by the exception branch of an invoke instruction. The landing block must land on a landingpad instruction as its first non- $\phi$  instruction. Given this restriction, two equivalent invoke instructions must also have landing blocks with equivalent landingpad instructions. This is easy to check since the landingpad instruction is always the first instruction in a landing block.

Landing blocks are responsible for handling all catch clauses of the higher-level programming language covering the particular callsite. All clauses are defined by the landingpad instruction, which encodes the list of all exception and cleanup handlers. Landingpad instructions are equivalent if they have the exactly same type and also encode an identical list of exception and cleanup handlers. The type of equivalent landingpad instructions must be identical as its value is crucial in deciding what action to take when the landing block is entered, and corresponds to the return value of the personality function, which must also be identical for the two functions being merged.

Properly handling *phi-nodes* requires a radical redesign in the code generator. The existing code generator produces code directly from the aligned sequence, with each instruction pair treated almost in isolation without considering any control flow context. Merging *phi-nodes* cannot work with this approach because *phi-nodes* are only

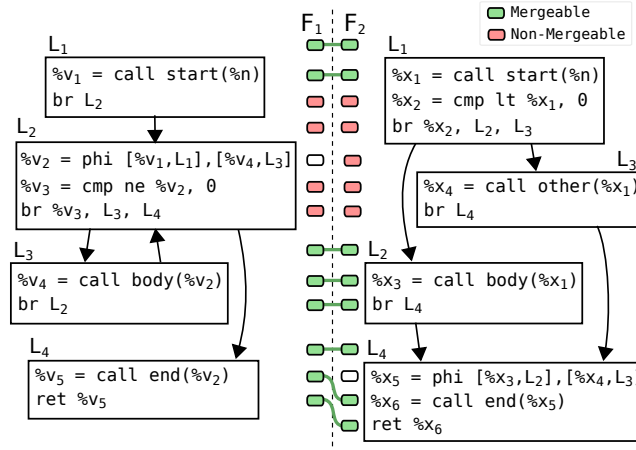


Figure 4.3: Example of functions aligned without register demotion. *Phi-nodes* are excluded from alignment.

understood in their control flow context.

**Road map** In the rest of this section, we describe SalSSA, our novel approach for merging functions through sequence alignment with full support for the SSA form. By removing the need for preprocessing the input functions and performing register demoting, our approach is able to merge functions better and faster. Instead of translating the aligned functions directly to merged code, the SalSSA follows a top-down approach centered on the CFGs of the input functions. It iterates over the input CFGs, constructing the CFG of the merged function, interweaving matching and non-matching instructions (Section 4.0.4 ). Afterwards, all edges and operands are resolved, including appropriately assigning the incoming values to all *phi-nodes* (Section 4.0.5). SalSSA is designed to preserve all properties of SSA form via the standard SSA construction algorithm (Sections 4.0.6). Finally, SalSSA integrates a novel optimisation with the SSA construction algorithm, called *phi-node coalescing*, producing even smaller merged functions (Section 4.0.7).

**Working examples** Figure 4.3 shows how the functions from our motivating example align without register demotion. Here, *phi-nodes* are not aligned, similarly to how FMSA handles *landing-pad* instructions. We will use these as working examples to describe step by step how our new code generator works in the next subsections.

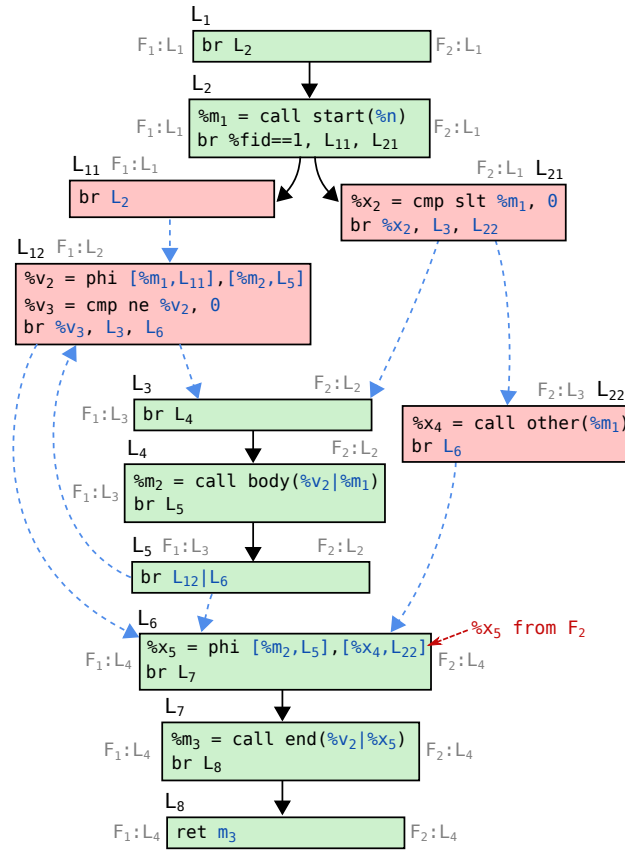


Figure 4.4: Merged CFG produced by SalSSA. Code corresponding to a single input basic block may be transformed into a chain of blocks, separating matching and non-matching code. The generator inserts conditional and unconditional branches to maintain the same order of instructions from the input basic block. Operands and edges highlighted in blue will be resolved by the operand assignment described in Section 4.0.5.

#### 4.0.4 Control-Flow Graph Generation

Our code generator starts by producing all the basic blocks of the merged function. Each original block is broken into smaller ones so that matching code is separated from non-matching code and matching instructions and labels are placed into their own basic blocks. Having one block per matching instruction or label makes it easier to handle control flow and preserve the ordering of instructions from the original functions by chaining these basic blocks as needed.

Blocks with instructions that come originally from the same basic block (of either input function) are chained in their original order with branches. We use either unconditional branches or conditional branches on the function identifier depending on

whether control flow out of this code is different for the two input functions. Because we have one basic block per pair of matching instructions/labels, this tends to generate some artificial branches, most of them are unconditional, but can be simplified in later stages.

Figure 4.4 shows the generated CFG. At this point, the only instructions that actually have their operands assigned are the branches inserted to chain instructions originating from the same input basic block. These branches have no corresponding instruction in the input functions. All other operands and edges, depicted in blue in Figure 4.4, will be resolved later, during operand assignment.

#### 4.0.4.1 Phi-Node Generation

Our code generator treats *phi-nodes* differently from other instructions. For all alignment and code generation purposes, SalSSA treats *phi-nodes* as attached to their basic block's label; that is, they are aligned with their labels and are copied to the merged function with their labels. So, when creating a basic block for a label, we also generate the *phi-nodes* associated with it. For a pair of matching labels, we copy all *phi-nodes* associated with both labels. We have decided for this approach where phi-nodes are tied to labels because phi-nodes describe primarily how data flows into its corresponding basic block. Figure 4.4 shows an example where *phi-nodes* are present in basic blocks with both matching or non-matching labels. The phi-node  $x_5$  is simply copied into the merged basic block labeled  $L_6$ .

Unlike other instructions, we do not merge *phi-nodes* through sequence alignment. Instead, identical *phi-nodes* are merged during the simplification process using existing optimisations from LLVM.

#### 4.0.4.2 Value Tracking

While generating the basic blocks and instructions for the merged function, SalSSA keeps track of two mappings that will be needed during operand assignment. The first one, called *value mapping*, is responsible for mapping labels and instructions from the input functions into their corresponding ones in the merged function. This is essential for correctly mapping the operand values. The second one, called *block mapping*, is a mapping of the basic blocks in the opposite direction, as shown by the light gray labels in Figure 4.4. It maps basic blocks in the merged function to a basic block in each input functions, whenever there is a corresponding one. This *block mapping* will

```

    %m2 = call body(%v2|%m1)
    |||
    %s = select %fid==1, %v2, %m1
    %m2 = call body(%s)

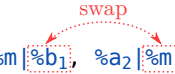
```

Figure 4.5: Operand selection for the `call` instruction in  $L_4$  from Figure 4.4. Mismatching operands chosen with a `select` instruction on the function identifier.

```

    %y = add %m|b1, %a2|%m
    |||
    %s = select %fid==1, %a2, %b1
    %y = add %m, %s

```



The diagram shows two instructions. The first instruction is `%y = add %m|b1, %a2|%m`. The second instruction is `%s = select %fid==1, %a2, %b1`. A red dashed arrow labeled "swap" points from `b1` in the first instruction to `%b1` in the second instruction, and another red dashed arrow points from `%a2` in the second instruction to `a2` in the first instruction, indicating a swap of operands to make them compatible for merging.

Figure 4.6: Optimizing operand assignment for commutative instructions. Example of a merged `add` instruction that can have its operands reordered to allow merging the two uses of `%m`, avoiding a `select` instruction.

be needed to map control flow when assigning the incoming values of *phi-nodes* (see Section 4.0.5.3).

## 4.0.5 Operand Assignment

Once all instructions and basic blocks have been created, we perform operand assignment in two phases. First, we assign all label operands, essentially resolving the remaining edges in the control flow graph (dashed blue edges in Figure 4.4). With the control flow graph complete, we can then create a dominator tree to help us assign the remaining operands while also properly handling instruction domination.

Whenever the corresponding operands of merged instructions are different, we need a way to select the correct operand based on the function identifier. Section 4.0.5.1 describes how we perform label selection. In all other cases, we simply use a `select` instruction, as shown in Figure 4.5.

When assigning operands to commutative instructions, we also perform operand reordering to maximise the number of matching operands and reduce the need for `select` instructions. Figure 4.6 shows an example of a commutative instruction where an operand selection can be avoided by reordering operands.

### 4.0.5.1 Label Selection

In LLVM, labels are used exclusively to represent control flow. More specifically, label operands are used by terminator instructions, where they specify the destination basic

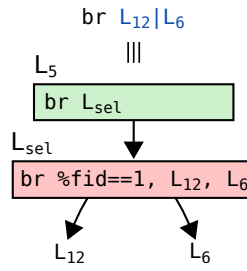


Figure 4.7: Label selection for mismatched terminator instruction operands  $L_{f1}$  and  $L_{f2}$  corresponding to labels of two different basic blocks. We handle control flow in a new basic block,  $L_{sel}$  with a conditional branch on the function identifier targeting the two labels. We use the label of the new block as the merged terminator operand.

block of a control flow transfer, or to represent incoming control flow in a *phi-node* instruction.

Whenever assigning the operands of a merged terminator instruction, if there is a label mismatch between the two input functions, we need a way to select between the two labels depending on the executed function. We do so by creating a new basic block with a conditional branch on the function identifier to each one of the mapped labels. Then we use the new block’s label as the operand of the merged terminator instruction. Figure 4.7 illustrates a CFG that handles label selection for a merged terminator instruction.

#### 4.0.5.2 Landing Blocks

Most modern compilers, including GCC and LLVM, implement the zero-cost Itanium ABI for exception handling [?], which is known as the *landing-pad* model. This model has two main components: (1) invoke instructions that have two successors, one that continues when the call succeeds as per normal, and another, usually called the *landing pad*, in case the call raises an exception, either by a throw or the unwinding of a throw; (2) landing-pad instructions that encode which action is taken when an exception is raised. A landing pad must be the immediate successor of an invoke instruction in its unwinding path. The code generator must ensure that this model is preserved.

Our new code generator delays the creation of landing-pad instruction until the phase of operand assignment. Once we have concluded the remapping of all label operands of an invoke instruction, regardless of whether they are merged or non-merged code, we create an intermediate basic block with the appropriate landing-pad instruction. Then we assign the label of this landing block as the operand of the invoke

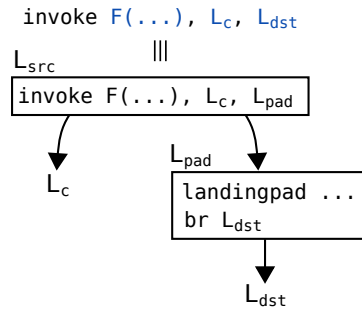


Figure 4.8: Landing blocks are added after operand assignment and are assigned to invoke instructions as operands.

instruction, as shown in Figure 4.8.

#### 4.0.5.3 Phi-Node's Incoming Values

There are two distinct cases for *phi-nodes*: being associated with a matching or with a non-matching label. In both cases, *phi-nodes* are only copied from their input functions and they are not merged. So each *phi-node* in the merged function should capture the incoming flows present in the corresponding *phi-node* of their input function. For matching labels, each *phi-node* in the merged function will have additional incoming flows specific to the *other* input function but these flows should have undefined values.

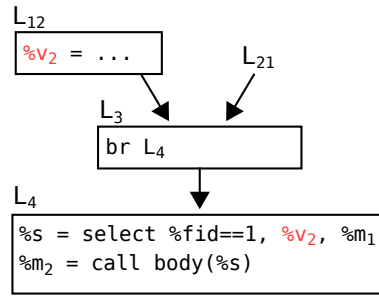
To assign the incoming values of a *phi-node*, SalSSA iterates over all predecessors of its parent basic block and uses the *block mapping* to discover each predecessor's corresponding basic block in the input function. If such a basic block is found, then SalSSA obtains the incoming value associated with that predecessor from the *value mapping*. Otherwise, an undefined value, which by construction should never be actually used, is associated with that predecessor.

#### 4.0.6 Preserving the Dominance Property

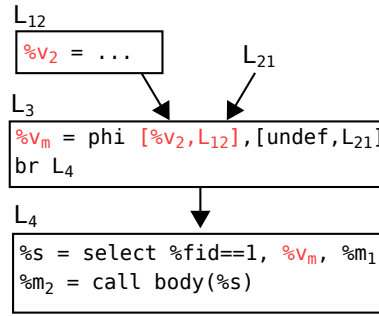
The code transformation process described so far could violate the *dominance property* of the SSA form. This property states that each use of a value must be dominated by its definition. For example, an instruction (or basic block) dominates another if and only if every path from the entry of the function to the latter goes through the former. Figure 4.9a gives one example extracted from Figure 4.4 where the dominance property is violated during code transformation.

SalSSA is designed to preserve the dominance property to conform with the SSA form. It achieves this using a two-step approach. It first adds a *pseudo-definition* at the





(a) Example where the dominance property is violated.



(b) The dominance property is restored by placing phi-nodes where needed.

Figure 4.9: Example of how SalSSA uses the standard SSA construction algorithm to guarantee the dominance property of the SSA form.

entry block of the function where names are defined and initialised with an *undefined* value. This guarantees that every register name will be defined on basic blocks from both functions. Then, SalSSA applies the standard SSA construction algorithm [? ?], which guarantees both the dominance and the single-reaching definition properties of the SSA form. We note that our implementation uses the standard SSA construction algorithm provided by LLVM for register promotion. This algorithm guarantees that names have a single definition by placing extra phi-nodes where needed so that instructions can be renamed appropriately. Figure 4.9b shows how the property violation in Figure 4.9a can be corrected using this strategy.

#### 4.0.7 Phi-Node Coalescing

The approach described in Section 4.0.6 guarantees the correctness of the SSA form but generates extra phi-nodes and registers which increase register pressure and might lead to more *spill code*. In this section, we describe a novel optimisation technique, *phi-node coalescing*, that SalSSA uses to lower register pressure.

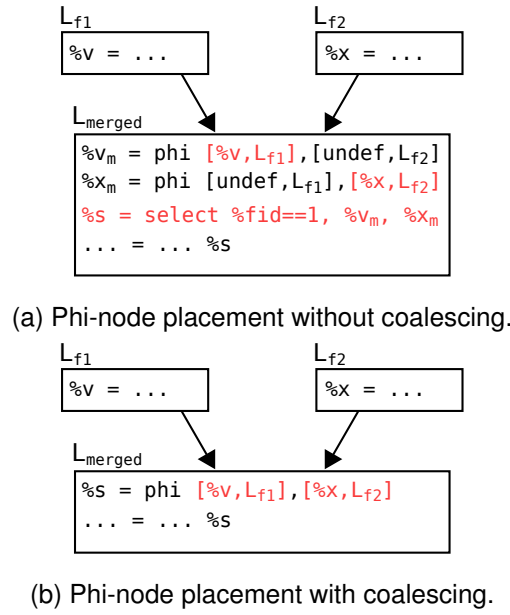


Figure 4.10: Phi-node coalescing reduces the number of phi-nodes and selections.

Figure 4.10 illustrates such an optimisation opportunity. SalSSA is merging an instruction with different arguments, so it needs to select the right one based on the function identifier. The two arguments though,  $v$  and  $x$ , have *disjoint definitions*, i.e. they have non-merged definitions from different input functions. Using the standard SSA construction algorithm would result in the sub-optimal code shown in Figure 4.10a. This code inserts two trivial phi-nodes to select, again,  $v$  or  $x$  based on the executed function. SalSSA optimises this code by coalescing both phi-nodes into a single one and removing the selection statement. As shown in Figure 4.10b, the optimised version has a smaller number of instructions and phi-nodes.

This transformation is valid because a value definition that is exclusive to a function will never be used when executing the other function. Figure 4.11 shows another example illustrating that even disjoint definitions that have no user instructions in common can be coalesced, reducing the number of phi-nodes.

Since SalSSA is aware of which basic blocks are exclusive to each function, it can choose a pair of disjoint definitions for coalescing. Given a pair of disjoint definitions, SalSSA assigns the same name for both of them before applying the SSA reconstruction. SalSSA coalesces the set of definitions that violate the dominance property. Two definitions can be paired for coalescing if they are disjoint and have the same type. The optimisation pairs disjoint definitions that maximise their live range overlap since the goal is to avoid having register names live longer than they should, reducing register

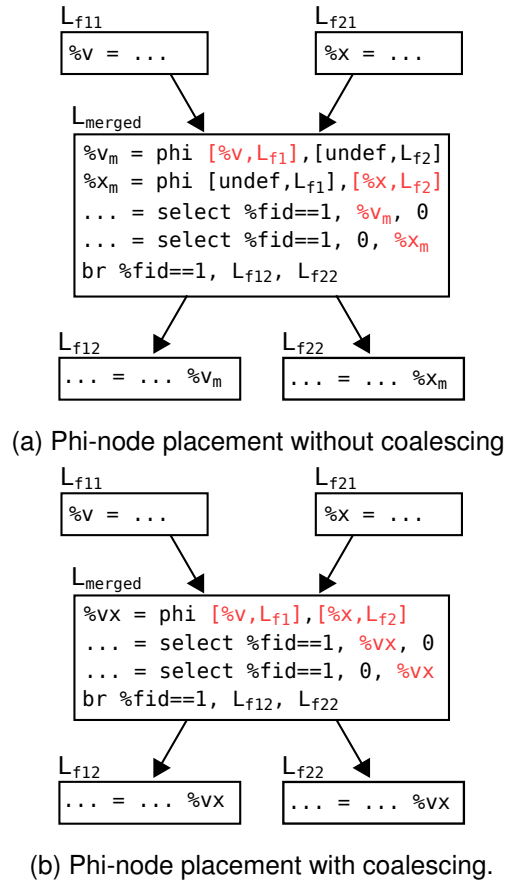


Figure 4.11: Reducing the number of phi-nodes by coalescing disjoint definitions with no user instructions in common.

pressure.

Formally, the heuristic implemented in our phi-node coalescing can be described as follows: Given a set  $S_1 \times S_2$  of disjoint definitions that violate the dominance property, the optimisation chooses pairs  $(d_1, d_2) \in S_1 \times S_2$  that maximise the intersection  $UB(d_1) \cap UB(d_2)$ , where  $UB(d)$  is the set  $\{Block(u) : u \in Users(d)\}$ .

Phi-node coalescing allows SalSSA to produce smaller merged functions and reduce code size. Consequently, it also enables more functions to be profitably merged.



# Chapter 5

## Optimisation Strategy

### 5.1 Profitability Cost Model

After generating the code of the merged function, we need to estimate the code-size benefit of replacing the original pair of functions by the new merged function. In order to estimate the code-size benefit, we first compute the code-size cost for each instruction in all three functions, which is done with the help of the compiler's target-specific cost model.

The compiler's cost model provides a target-dependent cost estimation that approximates the cost of an intermediate representation (IR) instruction when lowered to machine instructions. Compilers usually provide a performance cost model, which estimates the latency of each instruction, and a code-size cost model, which estimates the binary size of each instruction. These cost models play an essential role in the decision making of most compiler optimisations [10, 14]. The absolute cost estimated for one instruction is less important than its relation to the cost of other instructions. Ultimately, these costs are used for comparing two pieces of code, informing the compiler's decision of whether or not a given transformation is profitable.

In order to estimate code-size cost of one function, we sum up the code-size cost of all instruction by querying the cost model. In addition to measuring the difference in size of the merged function, we also need to take into account all extra costs involved: (1) for the cases where we need to keep the original functions with a call to the merged function; and (2) for the cases where we update the call graph, there might be an extra cost with a call to the merged function due to the increased number of arguments.

Let  $c(f)$  be the code-size cost of a given function  $f$ , and  $\delta(f_i, f_j)$  represent the extra costs involved when replacing or updating function  $f_i$  with the function  $f_j$ . Therefore,

given a pair of functions  $\{f_1, f_2\}$  and the merged function  $f_{1,2}$ , we want to maximise the profit defined as:

$$\Delta(\{f_1, f_2\}, f_{1,2}) = (c(f_1) + c(f_2)) - (c(f_{1,2}) + \varepsilon)$$

where  $\varepsilon = \delta(f_1, f_{1,2}) + \delta(f_2, f_{1,2})$ . We consider that the merge operation is profitable if  $\Delta(\{f_1, f_2\}, f_{1,2}) > 0$ .

However, these cost models are expected to contain inaccuracies. Because we are operating on the IR level, one IR instruction does not necessarily translate to one machine instruction. Several number of optimisations and code transformations will run afterwards, modifying the code. Moreover, we cannot know exactly how each IR instruction will be lowered without actually running the compiler's backend. The same IR instruction can be lowered to different machine instructions, depending on its surrounding context, the instruction selection algorithm, and many other factors. Therefore, there is also an inherent limitation of estimating the cost of each instruction separately of its context. However, the use of cost models is still a good trade-off between compilation time and accuracy.

## 5.2 Exhaustive Search

## 5.3 Focusing on Profitable Functions

Although the proposed technique is able to merge any two functions, it is not always profitable to merge them. In fact, as it is only profitable to merge functions that are sufficiently similar, for most pairs of functions, merging them increases code size. In this section, we introduce our framework for efficiently exploring the optimisation space, focusing on pairs of functions that are profitable to merge.

For every function, ideally, we would like to try to merge it with all other functions and choose the pair that maximises the reduction in code size. However, this quadratic exploration over all pairs of functions results in prohibitively expensive compilation overhead. In order to avoid the quadratic exploration of all possible merges, we propose the exploration framework shown in Figure 5.1 for our optimisation.

The proposed framework is based on a light-weight ranking infrastructure that uses a *fingerprint* of the functions to evaluate their similarity. It starts by precomputing and caching fingerprints for all functions. The purpose of fingerprints is to make it easy to discard unpromising pairs of functions so that we perform the more expensive

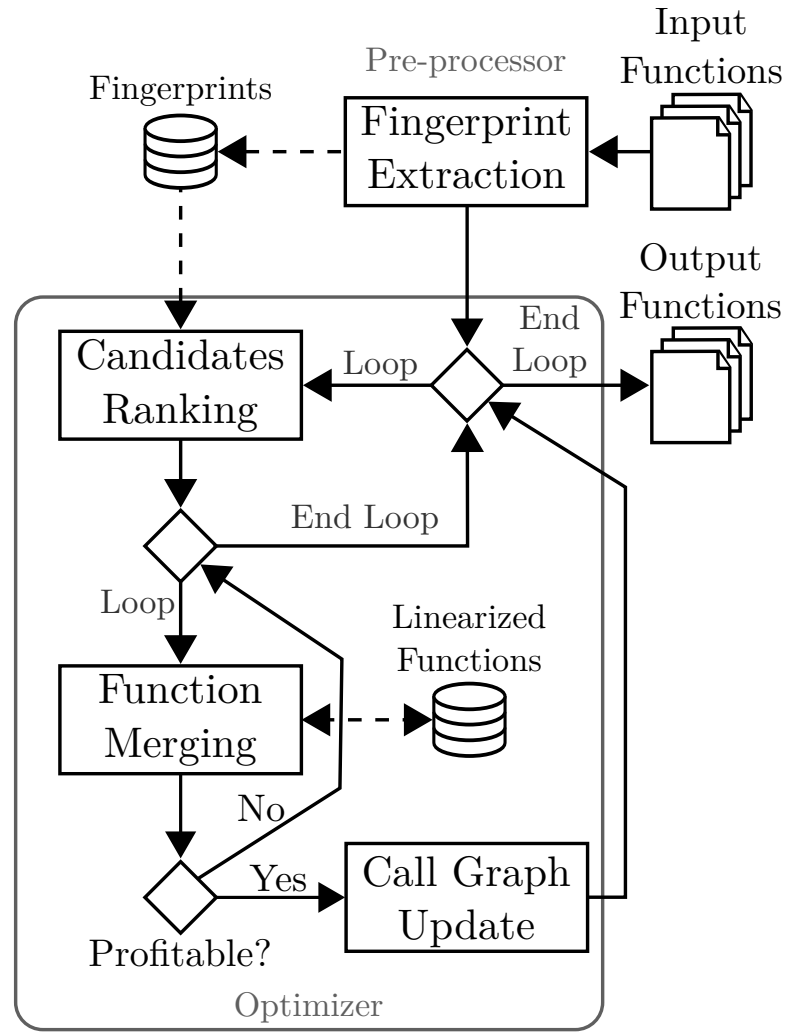


Figure 5.1: Overview of our exploration framework.

evaluation only on the most promising pairs. To this end, the fingerprint consists of: (1) a map of instruction opcodes to their frequency in the function; (2) the set of types manipulated by the function. While functions can have several thousands of instructions, an IR usually has just a few tens of opcodes, e.g., the LLVM IR has only about 64 different opcodes. This means that the fingerprint needs to store just a small integer array of the opcode frequencies and a set of types, which allows for an efficient similarity comparison.

By comparing the opcode frequencies of two functions, we are able to estimate the best case merge, which would happen if all instructions with the same opcode could match. This is a very optimistic estimation. It would be possible only if instruction types and order did not matter. We refine it further by estimating another best case merge, this time based on type frequencies, which would happen if all instructions

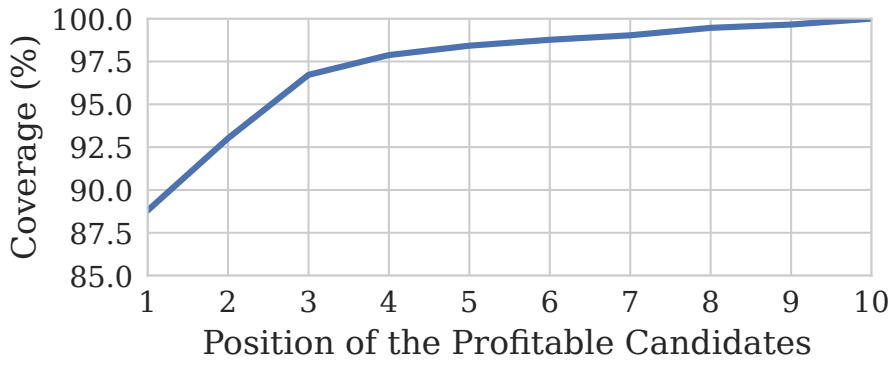


Figure 5.2: Average CDF for the position of the profitable candidate and the percentage of merged operations covered. 89% of the merge operations happen with the topmost candidate.

with the same data type could match.

Therefore, the upper-bound reduction, computed as a ratio, can be generally defined as

$$UB(f_1, f_2, K) = \frac{\sum_{k \in K} \min\{freq(k, f_1), freq(k, f_2)\}}{\sum_{k \in K} freq(k, f_1) + freq(k, f_2)}$$

where  $UB(f_1, f_2, Opcodes)$  computes the opcode-based upper bound and  $UB(f_1, f_2, Types)$  computes the type-based upper bound. The final estimate selects the minimum upper bound between the two, i.e.,

$$s(f_1, f_2) = \min\{UB(f_1, f_2, Opcodes), UB(f_1, f_2, Types)\}$$

This estimate results in a value in the range  $[0, 0.5]$ , which encodes a description that favors maximizing both the opcode and type similarities, while also minimizing their respective differences. Identical functions will always result in the maximum value of 0.5.

For each function  $f_1$ , we use a priority queue to rank the topmost similar candidates based on their similarity, defined by  $s(f_1, f_2)$ , for all other functions  $f_2$ . We use an exploration threshold to limit how many top candidates we will evaluate for any given function. We then perform this candidate exploration in a greedy fashion, terminating after finding the first candidate that results in a profitable merge and committing that merge operation.

Ideally, profitable candidates will be as close to the top of the rank as possible. Figure 5.2 shows the cumulative distribution of the position of the profitable candidates in a top 10 rank. It shows that about 89% of the merge operations occurred with the topmost candidate, while the top 5 cover over 98% of the profitable candidates. These



results suggest that fingerprint similarity is able to accurately capture the real function similarity, while reducing the exploration cost by orders of magnitudes, depending on the actual number and size of the functions.

When a profitable candidate is found, we first replace the body of the two original functions to a single call to the merged function. Afterwards, if the original functions can be completely removed, we update the call graph, replacing the calls to the original functions by calls to the merged function. Finally, the new function is added to the optimisation working list. Because of this feedback loop, merge operations can also be performed on functions that resulted from previous merge operations.

### 5.3.1 Link-Time Optimisation

There are different ways of applying this optimisation, with different trade-offs. We can apply our optimisation on a per compilation-unit basis, which usually results in lower compilation-time overheads because only a small part of the whole program is being considered at each moment. However, this also limits the optimisation opportunities, since only pairs of functions within the same translation unit would be merged.

On the other hand, our optimisation can also be applied in the whole program, for example, during link-time optimisation (LTO). Optimizing the whole program is beneficial for the simple fact that the optimisation will have more functions at its disposal. It allows us to merge functions across modules.

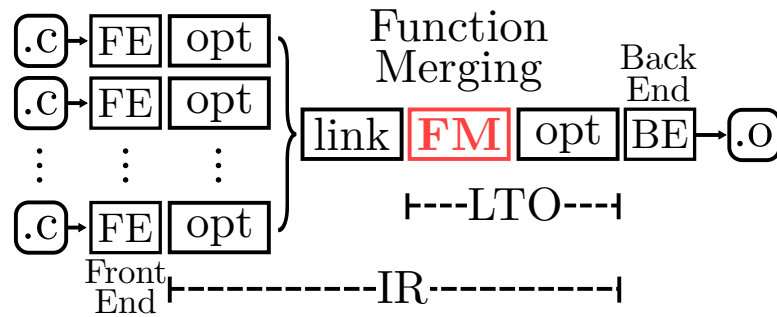


Figure 5.3: In our experiments we use a compilation pipeline with a monolithic link-time optimisation (LTO).

In addition to the benefit of being able to merge more functions, when optimizing the whole program, we can also be more aggressive when removing the original functions, since we know that there will be no external reference to them. However, if the optimisation is applied per translation unit, then extra conditions must be guaranteed, e.g., the function must be explicitly defined as internal or private to the translation unit.



# Chapter 6

## Reducing Runtime Overhead

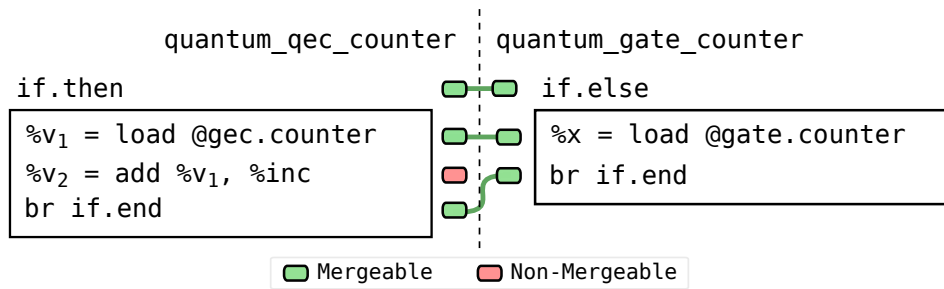
The identical function merging has no runtime penalty since it only alias two identical functions, removing one of the copies. However, merging partially similar functions may introduce runtime overheads in a given execution path.

Figure 6.1 shows an example of how function merging may introduce runtime overheads when merging two partially equivalent basic blocks. The two basic blocks involved in this merge operation were extracted from the `462.libquantum` benchmark. This example illustrates two different types of overheads that may be introduced during function merging: (i) the insertion of value selections; (ii) the insertion of branches to diverge to mismatching code or converge to matching code.

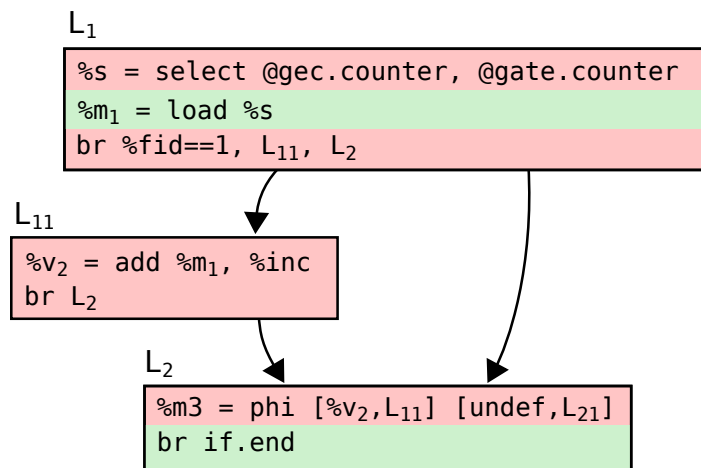
### 6.1 Conservative Function Merging

In this section, we propose a conservative function merging approach that aims at minimal runtime overhead in the absence of profiling information. The function merging approach described in Chapter 4 is capable of partially merging basic blocks, inserting branches between instructions to split matching from non-matching code. In order to minimise runtime overheads on possibly hot basic blocks, our conservative approach consider only merging whole basic blocks.

### 6.2 Profile-Guided Function Merging



(a) Pair of aligned basic blocks extracted from two larger functions in the 462.libquantum benchmark.



(b) Merged code generated for the pair of aligned basic blocks.

Figure 6.1: Example of how function merging may introduce runtime overheads.

## **Chapter 7**

## **Conclusion**



# Bibliography

- [1] Microsoft Visual Studio. Identical COMDAT folding. <https://msdn.microsoft.com/en-us/library/bxwfs976.aspx>, 2020.
- [2] The LLVM Compiler Infrastructure. MergeFunctions pass, how it works. <http://llvm.org/docs/MergeFunctions.html>, 2020.
- [3] Humberto Carrillo and David Lipman. The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.*, 48(5):1073–1082, October 1988.
- [4] Tobias J.K. Edler von Koch, Björn Franke, Pranav Bhandarkar, and Anshuman Dasgupta. Exploiting function similarity for code size reduction. In *Proceedings of the 2014 SIGPLAN/SIGBED Conference on Languages, Compilers and Tools for Embedded Systems*, LCTES '14, pages 85–94, New York, NY, USA, 2014. ACM.
- [5] Glenn Hickey and Mathieu Blanchette. A probabilistic model for sequence alignment with context-sensitive indels. In *Proceedings of the 15th Annual International Conference on Research in Computational Molecular Biology*, RECOMB'11, pages 85–103, Berlin, Heidelberg, 2011. Springer-Verlag.
- [6] Desmond G Higgins and Paul M Sharp. Fast and sensitive multiple sequence alignments on a microcomputer. *Bioinformatics*, 5(2):151–153, 1989.
- [7] Doug Kwan, Jing Yu, and B. Janakiraman. Google's C/C++ toolchain for smart handheld devices. In *Proceedings of Technical Program of 2012 VLSI Technology, System and Application*, pages 1–4, April 2012.
- [8] Martin Liška. Optimizing large applications. *arXiv preprint arXiv:1403.6997*, 2014.

- [9] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453, 1970.
- [10] A. Pohl, B. Cosenza, and B. Juurlink. Cost modelling for vectorization on ARM. In *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 644–645, Sept 2018.
- [11] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195 – 197, 1981.
- [12] SPEC. Standard Performance Evaluation Corp Benchmarks. <http://www.spec.org>, 2014.
- [13] Sriraman Tallam, Cary Coutant, Ian Lance Taylor, Xinliang David Li, and Chris Demetriou. Safe ICF: Pointer safe and unwinding aware identical code folding in gold. In *GCC Developers Summit*, 2010.
- [14] V. Porpodas, R. C. O. Rocha, and L. F. W. Góes. VW-SLP: Auto-vectorization with adaptive vector width. In *Proceedings of the 27th International Conference on Parallel Architectures and Compilation Techniques*, PACT ’18, pages 12:1–12:15, New York, NY, USA, 2018. ACM.
- [15] Lusheng Wang and Tao Jiang. On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337–348, 1994.