Towards Alleviating The Software Parallelization Task

Aleksandr Maramzin

Master of Philosophy

Institute of Computing Systems Architecture

School of Informatics

University of Edinburgh

2020

Abstract

Despite decades of research into parallelizing compiler technology, *software parallelization* still remains a largely manual task, which is complex, time consuming and error-prone. An embarrassingly parallel problem can be hidden behind a serial algorithm, thoughtless software design or unsuccessfully chosen lower level constructs, such as data structures. To elegantly and effectively map a parallel problem onto the exact hardware a programmer must possess an expert-level knowledge in various fields from software design and algorithmic patterns down to automatic vectorization and cache coherence. In this thesis we do not strive to find a ”silver bullet” and solve the problem of automatic parallelization. Neither do we expect an average programmer to be an expert. Instead, we acknowledge the role a programmer plays in the parallelization process and equip the former with an *assistant solution*. Our solution alleviates the task and makes parallelism more accessible to an average programmer.

The assistant solution consists of a tool and a library aiming at different stages of software parallelization. The tool aims at finer granularity levels. Program loops are often the richest source of parallelism and account for the biggest portion of the running time. The tool identifies those loops, which are both worthwhile and feasible to parallelize. For each loop the tool combines its potential contribution to speedup and an estimated probability for its successful parallelization. This probability is predicted using a machine learning model, which has been trained and tested on 1415 labelled loops, achieving a prediction accuracy greater than 90%. We present a methodology which makes a better use of expert time by guiding them directly towards those loops, where the largest performance gains can be expected while keeping analysis and transformation effort at a minimum. We have evaluated our parallelization assistant against sequential C applications from the SNU NAS benchmark suite. We show that our novel methodology achieves parallel performance levels comparable to those from expert programmers while requiring less expert time. On average, our assistant reduces the number of lines of code that have to be inspected manually before reaching expert-level parallel speedup by 20%.

The library implements the novel idea of *computational frameworks*, which are higher level entities that embody both data structures and algorithms together. The use of computational frameworks as parallel software design primitives alleviates the software parallelization task for a wide class of applications. We prototyped the library on the Olden benchmark suite. The parallel library version consistently outperforms the sequential version hitting 5-6x speedups on the major benchmarks.

Acknowledgements

This work was supported by grant EP/L01503X/1 for the University of Edinburgh School of Informatics Centre for Doctoral Training in Pervasive Parallelism. I would like to express my special gratitude to the UKRI and the CDT in Pervasive Parallelism for the great opportunity they gave me.

The work would not have been done without the constant direction and guidance of my primary supervisor Björn Franke. I would also like to separately thank Murray Cole for his wise advice and support. I express my gratitude to my co-supervisors Michael O'Boyle and Kenneth Heafield for their help in reviewing my project and directing my work.

I would also like to thank all my friends and colleagues for the numerous discussions, technical help and their time. Special thanks to Artemy, Chris, Nikolay, Roberto, Rodrigo and many, many others. And, of course I would like to thank my mother for her understanding, encouragement and support.

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.

(*Aleksandr Maramzin*)

Contents

Chapter 1

Introduction

1.1 Overview

Parallelism has become pervasive in the world of computing, with parallel hardware omnipresent across the whole spectrum of various computing systems from low-end embedded devices to high-end supercomputers. Yet, most of the existing software is written in a sequential fashion: be it an old legacy software initially designed to run on the serial hardware available the time or modern applications being developed by application domain experts rather than performance engineers. In order to exploit all available hardware facilities software has to be parallelized.

The software parallelization problem

Manual parallelization challenges The task of software parallelization has characteristically been a very manual process, which is multifaceted, extremely complex, time consuming and error-prone. An embarrassingly parallel problem might end up being hidden behind a thoughtless software design, a serial algorithm or being implemented with unsuccessfully chosen lower level constructs, such as pointers, heap-allocated and pointer-linked data structures, indirect array referencing, etc. To elegantly and effectively map a parallel problem onto the exact hardware a programmer must work on the various abstract levels and possess an expert-level knowledge in various fields from software design and algorithmic patterns in software engineering down to compiler automatic vectorization and hardware cache coherence protocols. It is not always realistic to expect such a wide expertise from an average programmer. Done in the wrong way software parallelization can even slow the program down in comparison to its original sequential version.

Automatic parallelization limitations Given the difficulty of manual software parallelization, there have been numerous efforts aimed at automating the task [5]. For several decades, parallelizing compilers have been the subject of intensive academic research [32] and industrial investment [3]. Yet, for most real-world applications they still fail to deliver parallel performance and to fully exploit the potential of modern parallel hardware one still needs to apply a significant manual effort. Furthermore, automatic parallelization techniques are limited to narrow domains of straightforward scientific Fortran codes and relatively simple computational idioms. When dealing with arbitrary real world codes automatically parallelizing compilers face a number of problems and challenges. The *Data-Centric Parallelization (DCP) problem* is an important demonstrative example. Listings 1.1 and 1.2 illustrate the problem and show how easily automatic parallelization can be hampered. In an array based implementation (Listing 1.1) compiler knows addresses of all sequence elements statically and can generate the code processing different array elements in parallel, while in a linked list based implementation (Listing 1.1) we see a pointer chasing code, where addresses of sequence elements will be resolved only dynamically and compiler cannot generate parallel code in advance. In the real world code the problem is obviously a way more challenging: *data structures are closely entangled with algorithms*. Parallelization of these codes requires a human mind.

```
for (int i=0; i<1024; i++) {
  a[i]=a[i]+1;
}
```

Listing 1.1: Parallelizable loop operating on a **flat array**.

```
for (p=list; p!=NULL; p=p->next) {
  p->value+=1;
}
```

Listing 1.2: Non-parallelizable loop operating on a **linked-list**.

Machine learning based parallelization applicability There were attempts to tackle the problem from another side. There is a vast body of research into utilizing more exotic machine learning based methods to the field of software parallelization. A good overview is provided by [45]. These methods have proved to be extremely useful and high performing on some compilation technology problems like selecting the best compiler flags or finding the most optimal compiler optimization parameters (like loop unroll or function inline factors). However, due to the inherent statistical errors and unavailability of large training data sets for compilation problems these methods have not yet found a widespread application in the area of software parallelization.

The assistant solution

In this thesis we are not trying to find a ”silver bullet” and solve the problem of automatic parallelization. Neither do we try to tune machine learning algorithms to a perfect 100% prediction accuracy. Given the difficulty of the obstacles faced by the field today, we do not expect that programmers will be liberated from performing manual parallelization in the near future [30]. Instead, we acknowledge the role of a human programmer in the software parallelization process, but we do not expect the programmer to be an expert either. What we try to do is to ***reduce the manual effort*** by providing a programmer with a parallelization *assistant solution*. Our solution alleviates the task and makes parallelism more accessible to an average programmer. The assistant solution we propose is as multifaceted as the problem itself. To fully exploit all the potential of software parallelization a programmer has to work on several conceptual levels. Thus, the assistant solution consists of a machine learning based loop parallelization tool [47] aiming at the finest levels of granularity, namely the program loops and a library of computational frameworks [48] aiming at a coarse grained parallelization on a higher level of software architecture design, algorithm and data structure choice.

1.2 Loop Parallelization Assistant

Despite decades of intensive research in automatic software parallelization [32], fully exploiting the potential of modern parallel hardware still requires a significant manual effort. Chapter 3 introduces a novel parallelization assistant that aids a programmer in the software parallelization process in the frequent case where automatic approaches fail. The assistant works at the finer levels of granularity, namely the program loops. Loops are compelling candidates for parallelization, as they are naturally decomposable and tend to capture most of the execution time in a program. The assistant reduces the manual effort in this process by presenting a programmer with a ranking of program loops that are most likely to 1) require little or no effort for successful parallelization and 2) improve the program’s performance when parallelized. Thus, it improves over the traditional, profile-guided process by also taking into account the *probability* of potential parallelization for each of the profiled loops.

At the core of our parallelization assistant resides a novel machine-learning (ML) model of loop parallelizability. Focusing on loops allows the model to leverage a large

amount of specific analyses available in modern compilers, such as generalized iterator recognition [44] and loop dependence analysis [40]. The model encodes the results of these analyses together with basic properties of the loops as machine learning *features*. The loop parallelizability model is trained, validated, and tested on 1415 loops from the SNU NAS Parallel Benchmarks (SNU NPB) [27]. The loops are labelled using a combination of expert OpenMP [11] annotations and optimization reports from the Intel Compiler (ICC), a production-quality parallelizing compiler. The model is evaluated on multiple machine learning algorithms. The evaluation shows that – despite the limited size of the data set – our model achieves a prediction accuracy higher than 90%.

The parallelization assistant combines inference on the parallelizability model with traditional profiling to rank higher those loops with a high probability of being parallelizable and impacting the program performance. An evaluation on eight programs from the SNU NPB suite shows that the program performance tends to improve faster as loops are parallelized in the ranking order suggested by our parallelization assistant compared to a traditional order based on profiling only. On average, following the order suggested by the assistant reduces by approximately 20% the number of lines of code a programmer has to examine manually to parallelize SNU NPB to its expert-level speedup. Given the high level of effort involved in manual analysis, such a reduction can translate into substantial development cost savings.

1.2.1 Contributions of our Loop Parallelization Assistant

In summary, our machine learning based loop parallelization assistant makes the following contributions:

▷ We introduce a machine learning model, which can be used to predict the probability with which sequential C loops can be parallelized (Sections 3.2 and 3.3);

▷ We integrate profiling of execution time with our novel ML model into a parallelization assistant, which guides the user through a ranked list of loops for parallelization (Section 3.4); and

▷ We demonstrate that our tool and methodology increase programmer productivity by identifying parallel loop candidates better than existing state-of-the-art approaches (Section 3.5).

1.3 Computational Frameworks library

The problem of software parallelization is multifaceted. In order to arrive at the final solution a programmer has to work on multiple conceptual levels. Our loop parallelization assistant reduces the programmer's efforts at a fine grained parallelization. To assist in tackling the problem on the higher level we propose the concept of *computational frameworks* and implement it as a prototype library. The concept blends algorithms and data structures together to form an elegant higher level entity. The word *computational* emphasizes the algorithmic component, while the word *framework* hints towards the underlying data structures. We demonstrate the utility and potential of the concept by deploying the library on the subset of Olden benchmark suite. We express benchmark computations in terms of our computational frameworks and rewrite the original legacy C versions of these benchmarks in a modern, better structured and crucially parallel way. The parallel library version consistently outperforms the sequential version hitting 5-6x speedups on the major benchmarks.

The concept of computational frameworks has been inspired by the current problems in the software parallelization field, as well as by the complexities of the real world legacy codes. We have already introduced the data-centric parallelization (DCP) problem. If a programmer had a tool, that could automatically recognize the type and properties of data structures and substitute them with simpler, parallelizable and more suitable alternatives, it would make the parallelization process a lot easier. Unfortunately, as our state of the art review shows (see Section 2.3.2), the problem is not solved yet. There are various techniques and methodologies with their specific limitations and problems. Static techniques like shape analysis [9],[12],[13] or pattern matching [39],[42],[43] are constrained in their program view. Dynamic techniques [22],[37],[41] come the closest to the solution, but are still far from tackling an arbitrary real world codes.

Generally, understanding the data structure type requires a thorough understanding of the algorithm that uses it and vise versa. We call that *the problem of data structure and algorithm inseparability*. For many real world programs the task of separating data structures from algorithms seems infeasible, but for some it does not seem meaningful either. For example, in many Olden benchmarks (see Section 2.7.3) the data structures and algorithms are blended together, but the union they form can be framed into an elegant higher level entity, that can later be parallelized in a nice and structured way. As we have already said we call that entity a computational framework.

There are some specific problems, that could be tackled more effectively with our computational frameworks. Imagine some scientific simulation task, which might be abundant with various higher level algorithmic constructs like *maps*, *reductions*, *folds*, *stencils*, etc. These concepts are characteristic of functional programming. The latter often forbids keeping any mutable states. At the same time, simulations require keeping of significant state being updated and accumulated with every step. The presence of state is characteristic of an imperative programming. Here one can see a contradiction and hence the gap to fill. Computational frameworks fill the gap between imperative and functional programming paradigms. While we keep the backbone algorithmic component immutable, computational frameworks provide a user space for customization through a human friendly, modern and convenient interface combining the elements of both functional, as well as object-oriented programming and embody the best ideas of various software design patterns.

To conclude, computational frameworks are higher level entities that blend algorithms and data structures to form an elegant union and improve program **structuredness**, **separation of concerns**, but crucially they can be used as parallel software design primitives to parallelize a wide class of applications in a portable and effective way hidden from a programmer behind a **modern and convenient** user interface.

1.3.1 Contributions of our Computational Frameworks

▷ We propose a novel idea of *computational frameworks*, which are higher level entities that embody both data structures and algorithms; and

▷ report on a research prototype C++ template library [48] implementing the idea in a modern, convenient, parallel and an easy to use way;

▷ We expressed the computations of some Olden benchmarks in terms of our computational frameworks and have rewritten their original sequential legacy C versions with the help of our library in a modern, better structured, portable and crucially parallel way (see Section 4.3);

▷ We demonstrate the potential of the idea and the performance of the prototype library on the suite of Olden benchmarks (see Section 4.4), achieving consistent parallel speedups of 5-6x on the major benchmarks;

▷ Finally, we propose an idea of alternative software parallelization approach based on our computational frameworks as a future work.

Chapter 2

Background and Motivation

Parallel computing and *software parallelization* are vast, overlapping and complementary computer science areas with a history dating back to 1950s. With the advances in semiconductor industry the topics have left the niche of high-end scientific supercomputers and spread to a much wider area spanning across all consumer electronic devices and have become of major importance.

Nowadays parallelism is pervasive. Parallel hardware is omnipresent across the whole wide spectrum of computing systems. In order to exploit all the available hardware capabilities software has to be parallel as well. And thus, every computer scientist and software developer would benefit from having an insight into the area. Nonetheless, the topics are extremely difficult, require a serious time investment and a great deal of knowledge in various subdomains. It is not always realistic to expect an average programmer to possess such a deep expertise. For that reason we propose a solution aimed at alleviating the challenging task of manual software parallelization. Our solution consists of the two components. We describe them in Chapters 3 and 4.

In this chapter we accumulate and structure the material, which stresses the importance of software parallelization, highlights its challenges, describes the parallel software engineering process and finally lays the ground for our proposed solutions from Chapters 3 and 4. **The major ideas leading towards the solutions from Chapters 3 and 4 are highlighted with bold face text.** We express our special gratitude to Lawrence Livermore National Laboratory (LLNL) [49] for their great parallel computing tutorials. We heavily relied on those to prepare the material of this chapter.

The background chapter is structured as follows. Section 2.1 stresses the importance of parallel computing and software parallelization in the modern world. There are numerous software parallelization methods and techniques, but all of them run into

their specific challenges and problems. Section 2.2 highlights the major problems of various software parallelisation methods. First, it presents the challenges of manual and then automatic software parallelization techniques. There have been various experiments and works employing machine learning (ML) based methods [45] for the task of software parallelization. These challenges form the ground for our program loop parallelization assistant solution [47] to grow on. We describe our solution in Chapter 3. Then, in Section 2.3 we discuss the problem of data structure choice and how it affects the software parallelization. In Section 2.3.2 we present a thorough literature review on the topic of Data Centric Parallelization (DCP) dealing with the task of automatic data structure recognition. Our review concludes with an observation that an automatic data structure recognition techniques are still an unsolved grand challenge. Data structures are often inseparable from the algorithms they support. Our computational frameworks build on that fact by defining the blend of data structures and algorithms. We propose our solution to the problem in Chapter 4. Furthermore, modern software design and engineering tasks are extremely rich and complex topics. That is true of parallel software engineering as well of course. In Sections 2.4 and 2.5 we talk about imperative, functional and object-oriented programming paradigms, as well as various OOP software design patterns. Our computational frameworks take the best from these principles and alleviate the task of parallel software architecture design for an average programmer.

2.1 The importance of parallel computing

Parallelism is pervasive and the future of computing is parallel. There are many factors which stress the importance of parallelism in the modern computing world.

Abundance of natural parallelism The field of High Performance Computing (HPC) has traditionally been concerned with scientific modelling and simulation of various natural phenomena (climate change, fluid flows, etc.). Such systems consist of numerous often independent parts. When we compile a highly parallel algorithm to a serial sequence of CPU instructions or process a huge data set with independent parts sequentially we are artificially constraining a vastly parallel computation to a serial one. Parallelism is not limited by the natural world, instead many algorithms have inherent parallelism in them.

Semiconductor technology advances and power limits With advances in transistor

density it became feasible to design more complicated CPUs. Initially the trend went into complex microarchitecture, for example, deeper pipelines, but running into power limits the industry design shifted towards multi core CPUs and multiprocessor systems. Such systems require of software to mirror the trend and become parallel as well.

Domain inherent parallelism and specialized computations The areas like computer graphics for instance have a lot of problems that can be processed in a Single Instruction Multiple Data (SIMD) fashion. That naturally led to specialized coprocessors like GPUs. Hardware systems become more complex and heterogeneous now.

2.2 Challenges in software parallelization

The problem of software parallelization is extremely complex and multi-faceted. There are various approaches to the problem, but all of them have their pros and cons. Although, the process of software parallelization has characteristically been a very *manual* task, which is time-consuming and error-prone, there are also *automatic* and *machine learning based* techniques. In this section we highlight inherent challenges of these approaches. The solution we propose aims to alleviate these.

2.2.1 Manual parallelization challenges

Parallel software development has characteristically been a very manual process. As any software development process it consists of a number of stages and parts. The major problems are described below.

Problem understanding and partitioning As the best software engineering practices dictate, before diving into software development one needs to thoroughly understand the problem and decide on the requirements and restrictions the final piece of software must meet. The whole algorithm and software architecture might change with the decision of developing a parallel software version instead of a serial one. If one starts from an already implemented serial software version, the parallelization might be even more difficult to do. Source code comprehension is a hard task. The algorithm chosen for a serial version might be completely unsuitable for a parallel implementation. The problem must be partitioned into relatively independent chunks of work to be processed in parallel. The partitioning can be done in numerous ways and a programmer

needs to choose the way to do that (data set decomposition, functional decomposition or a hybrid of the two).

Communications and synchronization Very often the parts of the problem are not completely independent and require an exchange of information. Designing the way that exchange is going to work is a complex task. Almost always communication results in an overhead. Sending the data over congested network or waiting on a synchronization barrier slows the program down. The slowdown might even diminish all performance benefits obtained from parallelization.

Implementation and data dependencies When the problem partitioning is done, all communication and synchronization points are determined and the high level parallel algorithm is designed, a programmer might start the actual implementation. Here a programmer will run into other types of problems. Consider two functionally equivalent code samples below.

```
for (int i=1; i<n; i++) {
  a[i]=a[i-1]+1;
}
```

Listing 2.1: Non-parallelizable loop with planted loop-carried data dependence.

```
for (int i=0; i<n; i++) {
  a[i]=a[0]+i;
}
```

Listing 2.2: Parallelizable loop free of any data dependencies.

The actual shape of the code can break parallelization by introducing fake (not required by the algorithm) dependencies.

Performance analysis and tuning One needs to know where the program's hotspots are. Hotspots are the places where the most of the real work is being done. The majority of programs spend most of the CPU time in a few places. The task of a programmer is to find those places and concentrate all parallelization and optimization efforts right there. Finding hotspots might be difficult before the programmer has the whole program implemented. Modern hardware architectures have a multi level memory hierarchy, memory data prefetchers, TLBs, out-of-order execution, etc. It might be surprising how the actual program execution performance differs from the one inferred from the algorithm. Profilers and other analysis tools can be of help here.

Finally, all the above challenges are interrelated and very often depend on each other. The parallel software development process can go iteratively with numerous dead ends and redesign efforts. With a long research history into the topic, all these problems are still actual now.

2.2.2 Limitations of automatic techniques

Given the enormous challenges of manual software parallelization there have been numerous attempts to automate the task. There are various tools available to assist a programmer in the task of software parallelization. Parallelizing compilers [1][3] are the most widely used. Below we present their classification:

Fully Automatic The compiler analyzes the source code and identifies opportunities for parallelisation. The analysis includes identifying inhibitors to parallelism and possibly a cost weighting on whether or not the parallelisation would actually improve performance [14]. Loops are the most frequent target for automatic parallelization [5].
Programmer Directed Using compiler directives or possibly compiler flags, a programmer explicitly tells the compiler how to parallelize the code. These directives and flags may be also used in conjunction with some degree of automatic parallelization. The most common compiler generated parallelization is done using on-node shared memory and threads (such as OpenMP [11]).

If one is starting with an existing serial code and has time or budget constraints, then automatic parallelization may be the answer. However, there are several important caveats that apply to automatic parallelization.

Performance Performance may actually degrade.

Limitations Limited to a subset (mostly loops) of code.

Effectiveness May actually not parallelize code i.e. it may be ”over conservative” or the code is too complex.

To clearly illustrate the problems an automatic software parallelization faces we conducted several experiments with the suite of NASA Parallel Benchmarks (NPB) [28]. These benchmarks target performance evaluation of highly parallel supercomputers. Consequently, the suite has a great amount of inherent parallelism and is supposed to be relatively easy parallelizable. NPB benchmarks do not provide the exact implementation, they rather specify what should be computed and how. We used Seoul National University's (SNU) implementation [33] of NPB benchmarks. SNU NPB implementation comes in two versions: sequential legacy C implementation and the one parallelized with OpenMP pragmas. The main data structure used everywhere in the suite is a simple flat array. There are numerous loop nests operating over arrays and computing simple reductions.

reason	num	reason	num	reason	num
unrecognised reduction	18	**array privatization**	7	**AA conservativeness**	60
unknown iteration number	7	**static dependencies**	46	**too complex**	22
uninlined calls	4	**other**	4	**total**	168

Table 2.1: Classification of loops missed by Intel Compiler for various reasons.

For these experiments we used a desktop Ubuntu 18.04 machine with installed Intel C/C++ Compiler (ICC) 18.0 and measured the running time of benchmarks with the help of UNIX time() utility. To minimize the errors, we ran experiments several times and took the mean average. The machine has 4 Intel Core i5-6500 CPUs with 3.20 GHz frequency and vectorization support up to AVX2. The RAM is 16Gb.

The first experiment we conducted was aimed at assessing the effectiveness of the state-of-the-art automatic parallelization. The experiment is platform agnostic as long as the target supports vector instructions and parallel primitives and can be reproduced on any such platform. We took the Intel C/C++ Compiler (ICC), configured it for the most aggressive parallelization (*-par-threshold0*), i.e parallelize all parallelizable code independent of its potential cost weighted profitability. Also, we configured ICC to do all enabling loop transformations (*-O3* flag) before the actual parallelization (*-parallel* flag) and vertorization (*-vector* flag). In other words, the experiment measures the maximum parallelization coverage the best state-of-the-art compiler can have on embarrassingly parallel problems, which still represent the real world code. Our results show a significant potential for improvement. Among all 1415 SNU NPB loops, 980 are truly parallelizable, but the ICC compiler manages to find only 812 parallelizable cases and misses 168 loops. Table 2.1 shows the classification we conducted by manually examining the ICC parallelization reports as well as looking at the source code of the benchmarks. The biggest problems are statically unknown pointers, which might potentially overlap at the running time, as well as other statically unresolvable dependencies. There are some unrecognized reductions as well as loops that can be parallelized with prior array privatization and function inlining.

While parallelization coverage is important it is not the primary goal. A parallelized loop might not make a significant contribution to the total running time of the application. We should strive to parallelize those loops, which are on the critical paths and hot

spots. And finally, only the running time is the ultimate parallelization performance measure. Given that, we conducted a further experiment. We used the same machine and compiled SNU NPB benchmarks with different sets of ICC automatic parallelization options. Figure 2.1 illustrates the running times of resulting codes. Bars marked as serial (s) show running times of original legacy C sequential versions. Bars marked as omp (o) show running times of an expertly manually parallelized versions. Other bars show running times of versions produced with various combinations of ICC compiler options (vectorize, parallelize or do both), as well as complete cases of parallel OpenMP versions, which have been additionally parallelized and vectorized by the ICC compiler. One can see that the best performance is still attributed to benchmark versions, which have been expertly parallelizad by their developers. Vectorization and parallelization help parallel versions a bit, but not that significantly and the profits can be neglected altogether. When we automatically vectorize serial versions we get a little improvement, but when we try to automatically parallelize them we get striking slowdowns on some benchmarks. Overall, automatic vectorization gives us a tiny 1.1x running time improvement in the geometric mean compared to 1.73x of manual parallelization. And the automatic parallelization results into 0.79x slowdown in the geometric mean across all the benchmarks.

Our machine learning based loop parallelization assistant [47] we propose in Chapter 3 extends parallelism recognition capabilities of the Intel C/C++ Compiler (ICC) by learning the loop parallelizability property and making predictions regarding it with an acceptable false positives rate. These predictions cover most of the cases missed by ICC. Moreover, our assistant helps to reach the best possible manual expert performance faster by guiding the programmer towards to the most fruitful code segments to parallelize.

2.2.3 Limits of machine learning based methods

Correctness is the most important property of the code. Although important, the running time or any other type of code performance characteristic is not always vitally critical. As all machine learning based techniques have always been characterised by their inherent and ineradicable errors [35], the field of compilers has never been the primary target for these methods. Nonetheless, these techniques have found their application to some problems withing the field.

ML in Compiler Optimization. Usually, machine learning based methods target prob-

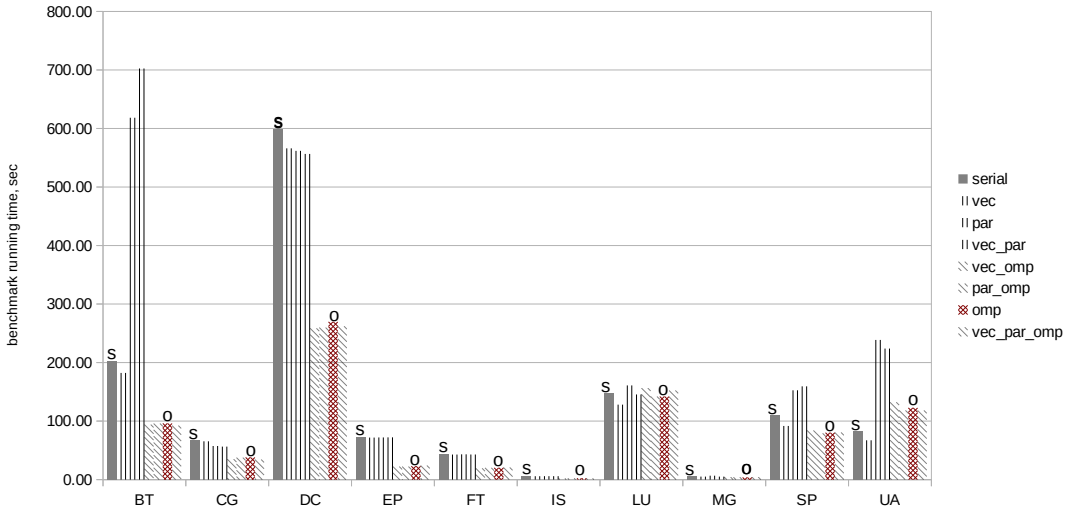

Figure 2.1: The running time of various NPB benchmarks versions.

lems, where a misprediction will only lead to a hampered performance and not a functional failure. For example, machine learning based methods have been used for finding the most optimal compiler optimization parameters like predicting the optimal loop unroll factor [23, 21] or determining whether or not a function should be inlined [17, 18]. These works are supervised classification problems and they target a fixed set of compiler options, by representing the optimization problem as a multi-class classification problem where each compiler option is a class. Recent works try to find a scheduling and optimization of parallel programs for heterogeneous multi-cores. For example, Hayashi *et al.* [36] extracts various program features during compilation for use in a supervised learning prediction model aiming at optimality of CPU vs. GPU selection. Evolutionary algorithms like generic search are often used to explore a large design space. Prior works [15, 19, 38] have used evolutionary algorithms to solve the phase ordering problem (i.e. in which order a set of compiler transformations should be applied).

Machine Learning and Parallelization. Application of machine learning based methods to the problem of software parallelization has not yet found a widespread practical utility. Mispredictions regarding the code parallelizability can lead to a broken dependency property and thus incorrect program execution. Nonetheless, there have already been works on prediciting loop parallelizability, like the approach of Fried *et al.* [34]. Fried *et al.* train a supervised learning algorithm on code hand-annotated with OpenMP parallelization directives in order to create a loop parallelizability predictor. These directives approximate the parallelization that might be produced by a human

expert. Fried *et al.* focus on the comparative performance of different ML algorithms and studies the predictive performance that can be achieved on the problem and does not produce any practical application.

In Chapter 3 we describe a practical ML based loop parallelization assistant [47]. Contrary to Fried *et al.* [34] it uses static source code features. Moreover, we use a richer training set, which is not limited to OpenMP pragmas only, but additionally takes the information from the Intel C/C++ Compiler. While Fried *et al.* focus on the comparative performance of different ML algorithms and study the possibility to learn loop parallelizability property, we do the same, but additionally we contribute a practical assistant capable of ranking loop candidates in their order of merit.

2.3 Data-Centric Parallelization (DCP) problem

2.3.1 The problem

The problem of data-centric parallelization (DCP) is the central motivation for the concept of computational frameworks we propose in Chapter 4. As it has already been stated the problem of software parallelization is multifaceted. There is a vast range of lower level technical issues, which can turn a perfectly parallelizable at a higher level computation into a non-parallelizable implementation. For example, in Section 2.2.2 we showed that the main reasons of Intel Compiler failures on SNU NPB benchmarks are alias analysis conservativeness, uninlined function calls and statically unresolvable dependencies. These reasons do not close the set of all possible parallel algorithm implementation failures. Listings 2.3 and 2.4 clearly illustrate a yet another potential implementation failure.

```
for (int i=0; i<1024; i++) {
  a[i]=a[i]+1;
}
```

Listing 2.3: Parallelisable loop operating on a **linear array**.

```
for (p=list; p!=NULL; p=p->next) {
  p->value+=1;
}
```

Listing 2.4: Non-parallelisable loop operating on a **linked-list**.

The above code snippets present two alternative implementations of the same simple and embarrassingly parallel computation. We increment all sequence elements by one. Listing 2.3 implements the sequence with a regular array linearly laid out

in the memory. Listing 2.4 chooses a linked list as an implementing data structure. While in an array based implementation the compiler knows all element addresses statically and can generate parallel code in advance in the linked list based implementation element addresses can be resolved only dynamically, which leads to a source code non-parallelizability.

The data-centric parallelization problem is how to automatically recognise what kind of a data structure is used in the code: is it a tree, a linked-list or a directed acyclic graph? The DCP problem is not solved yet. Automatic methods are limited to relatively simple code bases such as libraries of well known data structures. The most successful methods rely on dynamic analysis of memory graphs. Static techniques such as shape analysis are undecidable and highly conservative and might not finish in a reasonable time for the real software projects. The next section gives a comprehensive literature review on the topic.

2.3.2 Literature review

The idea of automatic discovery of higher level entities in programs is not a new one. This discovery problem is closely interlinked and entangled with alias analysis techniques [10] like points-to analysis [7]. Points-to analysis is a variation on data flow analysis techniques. The final output is the sets of pairs of the form (p, x) (pointer variable p points to a stack allocated variable x). These techniques are aimed at getting aliasing information regarding stack-allocated pointers.

The problem of understanding heap-directed pointers and heap-allocated linked data structures these pointers might point to is addressed with a family of static analysis techniques collectively known as shape analysis. Shape analysis techniques can be used to verify properties of dynamically allocated data structures in compile time. These are among the oldest and most well known techniques. Three-valued logic [12][13] can be used as an example. The technique proposes a construction of a mathematical model consisting of logical predicate expressions. The latter correspond to certain pointer operating imperative language program statements. Abstract interpretation of these statements leads to a construction of sets of shape graphs at various program points. Shape graphs approximate the possible states of heap-allocated linked data structures and answer the questions such as node reachability, data structure disjointness, cyclicity, etc. The major limitation of these simplified mathematical models is the lack of precision high level of abstraction leads to. The problem of precise shape

analysis is provably undecidable.

The work of [9] proposes a simplified and hence more practical implementation of shape analysis. Authors propose to use direction *D* and interference *I* matrices instead of complex mathematical models in order to derive shape information on heap allocated data structures. The entry of direction matrix *D[p,q]* says if there exists a path from a node referred to by *p* to a node referred to by q. In other words, if we can enter a path withing the data structure through *p* and exit through *q*. The entry of interference matrix *I[p,q]* says if the paths started from *p* and *q* are going to intersect at some point. Authors implement their technique withing McCAT compiler, which uses SIMPLE intermediate representation with a total of 8 statements (*malloc()*, pointer assignments *p=q*, structure updates *p->next=q*), which are capable of changing *D* and *I* matrices. Statements generate and kill entries in matrices. Moreover, they are capable of changing *Shape* attribute of pointers. The technique has been assessed on various benchmarks (bintree, xref, chomp, assembler, loader, sparse, etc.) from the era before the standard benchmark suites became available. The technique mostly reported shapes as *Trees* (be it a binary tree or a linked-list) or sometimes as *DAGs* or *Cycles* but with higher error rates in these last cases. The latter shows that the technique is imprecise and conservative.

One of the more recent techniques designed and developed by Ginsbach et al. is based on the pattern matching on LLVM IR level. The main idea is to specify computational idioms to be recognized in a domain specific constraint based programming language CAnDL [42]. Constraints are specified over LLVM IR entities such as instructions, basic blocks, functions, etc. The CAnDL language allows for a rapid prototyping of new compiler optimisations based on pattern recognition and its substitution with an optimised versions of matched idioms. The language and its relatively fast backtracking constraint solver are capable of recognizing not only simple arithmetic idioms (thus performing different peephole optimizations), but more complex computations like general reductions and histograms [39], vector products in graphics shaders [43], sparse and dense linear algebra computations and stencils [43]. Having recognized these computational idioms the work [43] replaces them with a code for various heterogeneous APIs (MKL, libSPMV, Halide, clBLAS, CLBlast, Lift) and compares the resulting performance demonstrating an improvement over sequential versions and matching performance to a hand-written parallel versions. The technique has been deployed on the sequential C versions of SNU NPB, the C versions of Parboil and the OpenMP C/C++ versions of Rodinia demonstrating improved detection capabilities

over the state-of-the-art techniques.

The other principally different technique has been recently proposed by Changhee Jung and Nathan Clark [22]. The authors developed a Data-structure Detection Tool (DDT) based on the LLVM framework. The tool instruments loads, stores and calls withing program binaries and gathers dynamic traces for sample inputs. The traces are used to recreate a memory allocation graph for program data structures. Call graphs are used to identify interface functions interacting with the built memory graph. DDT traces memory graph properties (number of nodes, edges, etc.) before and after interface function calls into another Daikon tool to compute dynamic invariants (the number of nodes in a memory graph decreses by 1 after every *(*delete()) interface method call, etc.). At the end manually constructed decision tree is used to probabilistically match observed behavioral patterns against known data structure invariant properties. The technique has been deployed to recognise data structure implementations withing standard libraries like STL, Apache (STDCXX), Borland (STLport), GLib, Trimaran achieving almost perfect recognition accuracy. Moreover, the technique has been able to recognise linked lists in Em3d and Bh Olden benchmarks, along with red-black trees implementing vectors in the Xalancbmk benchmark.

There has recently been other published works on the application of dynamic techniques to the problem of dynamic data structure recognition [41][37]. The technique used in the DDT tool [22] makes an assumption, that all modifications and interactions with memory graphs representing data structures happen through a set of interface functions. That is not true, when we deal with aggressively optimising compilers, which may eliminate some code or inline some functions. The MemPick tool [37] searches data structures directly on a built dynamic memory graph by analyzing its shape. The graph is built with the help of Intel Pin binary instrumentation tool during quiescent periods, when pointer operations are absent. DSIbin tool [41] operates with the source code rather than program binaries. Instead of memory points-to graphs it uses strands as primitives, which abstract such entities as singly-linked lists.

The work of Dekker [6] addresses software design recovery problem in a completely different way. Contrary to the approaches described above, which operate on the IR and dynamic instruction stream levels, work of Dekker operates at the level of abstract syntax tree. Dekker's tool tries to compact the tree down to a recognizable syntactic patterns by transforming it in accordance to a special grammar.

2.4 Imperative and functional programming

Programming languages can be classified by different *programming paradigms* they support. Among the most general classifications are *imperative* and *declarative* programming languages.

Imperative programs are written in a form of instruction sequences, which read and write the *state* of a program. The concept of state is the main characteristic of imperative programming paradigm. Instruction sequences can be structured in various ways. In *procedural programming* paradigm instructions are grouped inside procedures and functions. In *object-oriented programming (OOP)* paradigm instructions are grouped with the data they operate on inside objects of various types or classes. Programs are built either out of various procedures calling each other and exchanging the data or on the interaction of objects of various types. Imperative programs specify the exact sequence of steps to take in order to compute the final result.

Declarative programs do not specify the exact sequence of steps and state updates a program needs to do in order to get the desired result. Declarative programs declare the properties of the desired result. The properties can be specified as a set of constraints like in *constraint programming* or a set of linear inequalities like in *linear programming*. *Functional* programming is another subtype of declarative programming. In functional programming the final result is specified as a sequence of stateless function evaluations, which form a tree of expressions. Among the most common constituents are functions like *map*, *reduce*, *fold*, etc. Functions can be passed as arguments and returned from other functions ultimately composing bigger programs.

Functional programming is sometimes treated as synonymous with purely functional programming, a subset of functional programming which treats all functions as deterministic mathematical functions, or pure functions. When a pure function is called with some given arguments, it will always return the same result, and cannot be affected by any mutable state or other side effects. This is in contrast with impure procedures, common in imperative programming, which can have side effects (such as modifying the program's state or taking input from a user). Proponents of purely functional programming claim that by restricting side effects, programs can have fewer bugs, be easier to debug and test, and be more suited to formal verification.

There are no universally optimal programming paradigms and languages. Some languages are more convenient and suitable for one sort of problem, some languages are better at tackling other problems. For example, functional languages are more con-

venient in addressing certain domains such as R for statistics and financial analysis. Imperative languages are certainly better for simulations and other state based scientific computations. For that reason, major languages are often multi-paradigm to cover a potentially larger set of problems. Largely imperative C++ language included support for functional programming with its newer standards starting from C++11.

There is still an unfilled gap between functional and imperative programming paradigms: some problems contain computations, which are better expressed with standard functional concepts (like *maps*, *folds*, *reductions*, etc.), but at the same time require some state keeping. These problems lie at the boundary of functional and imperative programming. In Chapter 4 we propose an idea of *computational frameworks*, which fills that gap. Moreover, we develop a prototype library [48] with a modern and convenient programmer interface, which takes the best features of both object-oriented and functional paradigms and reflects the core ideas behind the concept of computational frameworks.

2.5 OOP and software design patterns

2.5.1 Object-Oriented Programming (OOP)

Object-oriented programming (OOP) is arguably the most widely used programming paradigm nowadays. It is supported by almost all major programming languages. At the very essence, in OOP computer programs are designed by making them out of objects that interact with one another. Object interactions are very close to human level of reasoning and logic and thus the paradigm fits quite naturally to a human developer.

Objects are instances of different types or classes in OOP terminology. Classes are object specifications. They specify the data objects contain (like an integer *age* field for an object of class *Person*) and the methods used to operate on the data. Classes define the public part of objects as well as their internal implementation. Object-oriented (OO) languages provide a rich set of facilities and features to build programs.

Encapsulation is used for protection against object misuse and unintended outside interference: data and methods concerned with internal workings are declared *private*, while those designed to form an outward appearance are declared *public*. This facilitates code refactoring, for example allowing the author of the class to change how objects of that class represent their data internally without changing any external code. It also eases debugging by better localizing functionality and thus possible bugs.

Dynamic dispatch is the responsibility of the object, not any external code, to select the procedure to execute in response to a method call, typically by looking up the method at run time in a table associated with the object. This feature allows a programmer to write general code, which works with abstract interface methods and leave the exact method resolution to be made during the running time of a program.

Dynamic dispatch is closely related to the technique of *inheritance*. Inheritance allows classes to be arranged in a hierarchy that represents ”is-a-type-of” relationships. Inheritance can be of two types: interface and implementation inheritance. The first one allows a parent class to require its descendants to stick to the same interface. A common interface allows the objects of different classes from the same hierarchy to be operated on by the same type agnostic code. The latter is called a *polymorphism*. The user code can be more concise and abstract. The call of the same method on the parent class or one of its descendants can result into a varying behaviour.

Features that are described above do not close the set of all available OOP techniques and mechanisms. These are just the main features we rely on in our project of computational frameworks and are the most important to know.

2.5.2 Software design patterns

The presence of all above features makes OOP languages extremely rich with various facilities. That creates a vast design space for software architects and turns the OOP software design into an art. *Software design patterns* [8] are reusable solutions to common design problems in OOP. Design patterns have been well tested and are proven to be the most reliable and elegant solutions for design problems they target.

Design patterns live at an intermediate level between the exact algorithm and a programming paradigm and are language agnostic. Design patterns specify and document solutions to common design problems. These solutions consist of a set of classes, objects and the ways they interact with one another. It is agreed to classify patterns into 4 distinct categories: creational, structural, behavioural and concurrency patterns. Here we are going to mention only those patterns, which are the most relevant to our project of computational frameworks.

The *command* pattern is a behavioural pattern, where a command object is used to encapsulate all the information needed to perform an action or trigger some event at a later time. The other participants are the receiver, invoker and client. The central ideas of this design pattern closely mirror the semantics of first-class functions and higher-

order functions in functional programming languages. Specifically, the invoker object is a higher-order function of which the command object is a first-class argument.

One can see the command pattern used in combination with the *chain of responsibility pattern*, which is a design pattern consisting of a source of command objects and a series of processing objects. Each processing object contains logic that defines the types of command objects that it can handle; the rest are passed to the next processing object in the chain. In a variation of the standard chain of responsibility model, some handlers may act as dispatchers, capable of sending commands out in a variety of directions, forming a *tree of responsibility*. The chain of responsibility pattern promotes the idea of *loose coupling*, where the system consists of multiple components, which know little or nothing about the definitions of other components.

Another very relevant pattern is a *template method* pattern. The template method is a method in a superclass, usually an abstract superclass, and defines the skeleton of an operation in terms of a number of high-level steps. These steps are themselves implemented by additional helper methods in the same class as the template method. The helper methods may be either abstract methods, for which case subclasses are required to provide concrete implementations, or hook methods, which have empty bodies in the superclass. Subclasses can (but are not required to) customize the operation by overriding the hook methods. The intent of the template method is to define the overall structure of the operation, while allowing subclasses to refine, or redefine, certain steps.

The *visitor* pattern is a way of separating an algorithm from an object structure on which it operates. A practical result of this separation is the ability to add new operations to existing object structures without modifying the structures. In essence, the visitor allows adding new virtual functions to a family of classes, without modifying the classes. Instead, a visitor class is created that implements all of the appropriate specializations of the virtual function. The visitor takes the instance reference as input, and implements the goal through double dispatch.

Software design patterns can be more down to the language like the *curiously recurring template pattern (CRTP)*, which is an idiom in C++ where a class X derives from a class template instantiation using X itself as template argument.

Software design patterns specify and document the best practices to solve various commonly reoccurring problems. At the end it is the experience, mastery and ingenuity of a programmer which determine the final software design.

2.5.3 Computational frameworks and design patterns

When we introduce the concept of computational frameworks, it is very important to mention software design patterns. While not exactly the same, the two concepts have an overlap. Computational frameworks do not specify the exact algorithms and data structures that must be used to implement the task, but alike declarative programming rather define the requirements on a higher level computation. For instance, a fractal can be implemented as a tree or as a heap and a fold can be implemented with an array or with a linked list, but the higher order computational requirements must be met. One can see a similarity to the template method design pattern, which also specifies the main computation and leaves some lower level steps to be defined by a user. One can see a similarity between the chain of responsibility pattern and the fold computational framework. The latter is also a visitor pattern in its own way. While the backbone computation of our frameworks is implemented with a template method pattern, the command pattern is used to specify the user defined custom part. Computational frameworks also promote the idea of loose coupling by keeping the computational procedure strictly decoupled from the side effects accumulation.

Computational frameworks embody the best ideas of all relevant software design patterns (template method, command, curiously recurring template pattern, etc.) and can be also considered as such. They relieve a programmer from a complex task of software architecture design by providing a ready solutions with all computational, communicational and data access patterns known and implemented in advance.

2.6 Parallel algorithmic skeletons

Algorithmic skeletons or parallelism patterns [31] are a high-level parallel programming model, that take advantage of common programming patterns to hide the complexity of parallel and distributed applications. For example, the basic algorithmic skeletons are *map*, *pipeline*, *reduce*, *scan*, *stencil*, *divide and conquer*, *task farm*, etc. More complex parallelism patterns can be composed out the basic ones. When the computational and data access patterns are known in advance, algorithmic skeletons hide all complexities (synchronization, communication, etc.) of parallel programming and let a user to concentrate on the algorithmic part. Algorithmic skeletons are higher than other parallel programming models like MPI, OpenMP, POSIX threads and they

have a vast array of various implementations and libraries. The latter vary by the programming language they target, distribution library used to implement parallel/distributed computations internally (like MPI, OpenMP, etc.), the sets of supported skeletons and the allowed level and way of skeleton nesting, i.e composing a more complex patterns out the basic ones.

While algorithmic skeletons or parallelism patterns stress an algorithmic component of the computation, computational frameworks contain an underlying data structure that might accumulate computation's side effects.

2.7 Benchmark studies

In our projects we used three benchmark suites. We trained our machine learning (ML) based loop parallelization assistant [47] (see Chapter 3) with Seoul National University's implementation [33] of the NASA Parallel Benchmarks (NPB) [33]. The suite comes in two versions: sequential and parallelized with OpenMP. We used the latter to extract ML training labels. Moreover, the suite contains a load of various sorts of parallel loops, which makes it a good training set. For the data-centric parallelization (DCP) project we started with the SPEC CPU2006 benchmarks. The complexity level of that suite and the perspective we derived from the suite study led us to a simpler suite of Olden benchmarks. We used the latter for the project of computational frameworks (see Chapter 4). Below we provide descriptions of benchmarks, so a reader can develop a better feel for the problems we tackle and the code we work with.

2.7.1 NASA Parallel Benchmarks (NPB)

NAS Parallel Benchmarks (NPB) [28] target performance evaluation of highly parallel supercomputers. NPB are "paper-and-pencil" benchmarks, i.e they do not provide the exact implementation, but rather specify various computational problems at a higher level. There are various implementations. We used the one from Seoul National University (SNU) - SNU NPB [33].

There are 10 various benchmarks in the suite. Benchmarks perform various scientific computations. They solve various systems of linear equations, compute gradients, work with matrices (compute matrix transpose, inverse, etc.), solve differential equations, solve heat and diffusion equations on the mesh, compute 3d grids, etc. The main data structure used in all SNU NPB benchmarks is a flat multidimensional array.

These arrays are processed in loops of various complexity levels including embarrassingly parallel loops like (*a[i] = b[i] + c[i]*), parallel copy and initialization loops (*a[i] = b[i]* and *c[i] = 0.0*), loops computing simple reductions, unrolled loops, multilevel nested loops and loops with uninlined function calls, as well as more complex loops with *if* and *switch* statements and loops with statically unknown iteration numbers and indirect array references (*a[i] = b[c[i]]*).

To the biggest part, these benchmarks are inherently parallel, but surprisingly their automatic parallelization with the best state-of-the-art Intel C/C++ Compiler (ICC) [3] results into a slowdown (see Section 2.2.2). Moreover ICC fails to recognize a lot of parallel loops. Our parallelization assisatant increases parallelization coverage and leads a programmer to achieve a manual expert level parallel benchmark performance faster.

2.7.2 SPEC CPU2006

The project of Data-Centric Parallelization (DCP) started with the feasibility studies of the SPEC CPU2006 benchmark suite. We studied the feasibility of the automatic data structure recognition techniques on these benchmarks. Although the benchmarks proved to be extremely complex for such techniques, these studies directed our further efforts and ultimately led to the concept of computational frameworks being an inseparable blend of data structures and algorithms. The key lessons we learnt from the SPEC CPU2006 benchmarks are an enormous complexity of the real legacy code and a very close relationship between data structures and algorithms. Let alone automatic techniques, it might take some weeks for an expert engineer to understand what a single benchmark is actually doing. Below we describe some of the benchmarks we looked at.

429.mcf The benchmark is derived from MCF, a program used for single-depot vehicle scheduling in public mass transportation. The benchmark operates with a complex network of nodes and arcs linearly allocated on the heap memory. Despite the simplicity of allocation, every node and arc has numerous pointers forming several object linking chains. Pointers are set in different places withing the source code base (during allocation as well as during consecutive network structure updates), making the deduction of the actual data structure shape a task of grand complexity. The network forms a spanning tree with several properties true of its nodes: every node has only one child pointer and if a node has several children, then the latter are connected

through sibling pointers starting from the first child.

The tree data structure presents a high interest from the point of its recognition. But even a manual source code analysis and transformation requires a serious effort. Static automatic techniques seem infeasible, while dynamic ones seem to be a grand challenge.

456.hmmer Searches a DNA sequence database given a Profile Hidden Markov Model (HMM). The benchmark uses the Viterbi algorithm. The implementation works with four dynamic programming matrices allocated linearly as arrays. Algorithm walks either horizontally or diagonally along these matrices and computes reductions of maps. The computation is parallelizable and there has been successful work [25][24] doing it manually for specialized hardware.

Despite the complexity of its core function *P7Viterbi()*, the benchmark presents a very high interest for the application of our computational frameworks. Reductions of maps perfectly fit the purpose. We view it as a future work.

400.perlbench This benchmark is a cut-down Perl interpreter, which implements the regular expression matching state machine. The benchmark processes the bitcode of a compiled regular expression instruction by instruction. Although, instructions have the same size and are laid out linearly in memory, there might be branches and the whole processing happens in a linked-list offset directed fashion. That requires sequential execution and is far beyond the capabilities of any existent techniques.

The benchmark is neither parallel nor a simple one. It makes no point to apply any recognition techniques here.

470.lbm The benchmark implements a Lattice Boltzmann Method (LBM) and is a relatively simple one (around 1400 LOC). The main underlying data structures the benchmark works with are the two 3D grids mapped onto a linear array space, which simulate incompressible fluids in 3D. The benchmark runs over arrays a specified number of time steps. Every array element represents a point from a 3D grid and consists of a number of velocity vector projections at this point. The values of these projections are being combined and mapped in a stencil fashion. The computation is highly parallel.

The benchmark operates with 3D grids laid out on regular arrays. The latter do not present a great deal of interest from the point of data structure recognition. Although, it would be interesting to try to recognize an algorithmic stencil.

2.7.3 Olden

Although the studies of SPEC CPU2006 benchmarks have proved their enormous complexity for the task of automatic data structure recognition, we acquired a good perspective and narrowed our research path to a much simpler suite of Olden benchmarks. Computations and algorithmic patterns present in Olden benchmarks have ultimately led us to the concept of *computational frameworks* (see Chapter 4).

Olden benchmark suite consists of 10 benchmarks. For our project we looked at 6 of those (*bisort*, *health*, *perimeter*, *treeadd*, *mst* and *tsp* benchmarks). The nature of different Olden benchmarks varies. Benchmarks health, treeadd and perimeter perform essentially the same computational pattern, but for different problems. We call that pattern a fractal and it is basically a parallel tree growth and processing. Benchmarks tsp and mst solve 2 well-known graph problems namely travelling salesman problem (TSP) and minimum spanning tree (MST) construction. The other 4 benchmarks perform scientific numerical computations and are bigger, more complex and less interesting from the perspective of data structure recognition.

bisort ***Definition*** *A sorted sequence is a monotonically non-decreasing (or non-increasing) sequence. A bitonic sequence is a sequence with* $x_0 \leq ... \leq x_k \geq x_{k+1} \geq ... \geq x_n - 1$ *for some k, or a circular shift of such a sequence.*

The sequence is implemented as a binary tree (recursive calls to the left and right subtrees). The algorithm is based on a sorting comparator network consisting of several layers. The network can be and is implemented in a divide and conquer way similar to that of a well-known merge sort. Sort() function is called on the left and right array halves recursively. The merging step of the merge sort algorithm is substituted with compare-and-swap step. The latter is possible due to input sequences required to be bitonic.

The benchmark is heavily based on pointers, tree swaps and rotations. It presents an interest from the point of divide and conquer algorithm recognition. Static techniques are unlikely to handle the legacy source code of that complexity and style. Dynamic techniques might be able to see the binary tree.

health The health benchmark fits into the fractal computational framework the best and hence delivers the most promising performance results. Health benchmark does a simulation of the Columbian healthcare system and is based on a complete 4-ary tree of villages. One can view the tree as the hierarchical structure reflecting various levels of municipal divisions: the root is the capital, leaves represent villages

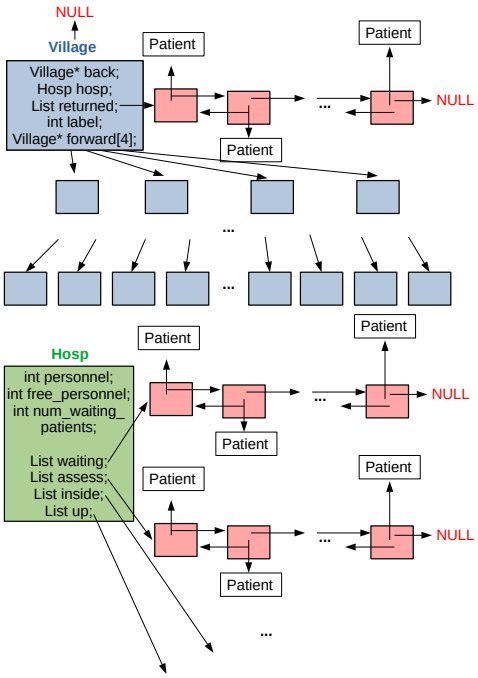

Figure 2.2: The *health* benchmark.

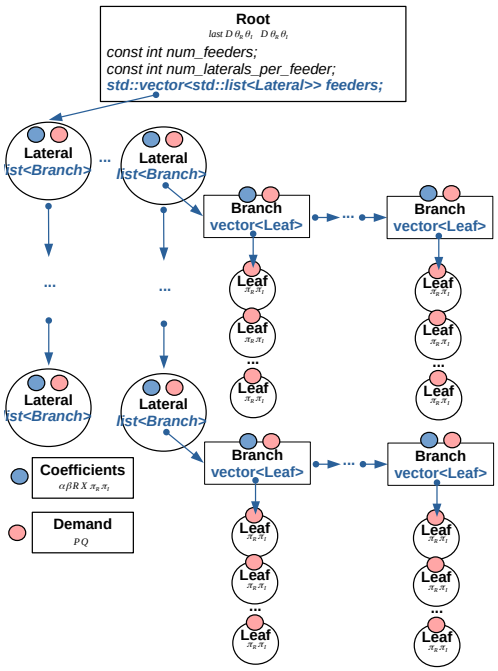

Figure 2.3: The *power* benchmark.

and the nodes on intermediate tree levels represent small towns and bigger cities. The closer to the root, the bigger the settlement and higher the level of hospital expertise. Every village has its own hospital. Figure 2.2 illustrates the tree.

Simulation starts at the tree root and goes down to the leaves. There are people who fall ill in every village and when they do they go into the local hospital for an assessment and in case they are ill for treatment. If local staff cannot give an accurate diagnosis a patient goes up the tree to a bigger settlement with a higher staff expertise level. As the simulation goes the lists of patients waiting, under assessment or inside the hospital for treatment grow and the benchmark state becomes bigger. So does the workload. The computation is parallel: all tree node child sub-trees can be simulated independently. Child nodes pass lists of patients to their parents and the latter take them to their local hospitals.

The benchmark operates with quad tree structures and does it in a highly parallel fashion. In our work we call that pattern a fractal. The latter is a computational framework, i.e. the blend of an algorithm and a data structure. The whole structure can be aimed at for an automatic recognition (not just a separate tree or an algorithmic skeleton, but both) even with static techniques.

perimeter The perimeter benchmark is another example of the fractal framework. The benchmark computes the perimeter of a ring (r=1024, R=2048). Figure 2.4 illus-

trates the process. The ring is placed onto a square, which is then being continuously and recursively divided into 4 equal sub-squares (southwest, northwest, southeast and northeast). Only guided by the stop condition the process continues further. We divide the square further if it falls on the intersection with the ring boundary. If the square falls completely inside the ring or completely outside we stop the division. Division also stops, when the square area becomes less than a preset granularity (or equivalently the depth of the tree). The process can be represented as a growing unbalanced quad tree. Squares are the nodes of the tree. Squares inside the ring (all 4 square corners (x,y) are $r^2 < x*x+y*y < R^2$) are painted black, while those outside are painted white. In other words, the growth stops at black and white tree nodes and continues for those representing squares on the intersection with the ring. Finally, the resulting grid is being traversed to catch all flips of color. When the flip is detected we increment the perimeter counter by the square area adjacent to the flip boundary. The latter results into the final perimeter approximation.

The benchmark operates with quad tree structures and does it in a highly parallel fashion. This fits into the fractal pattern as well. The only difference from the health benchmark is the tree growth stop condition resulting into unbalanced tree. Automatic fractal recognition seems harder, but still possible.

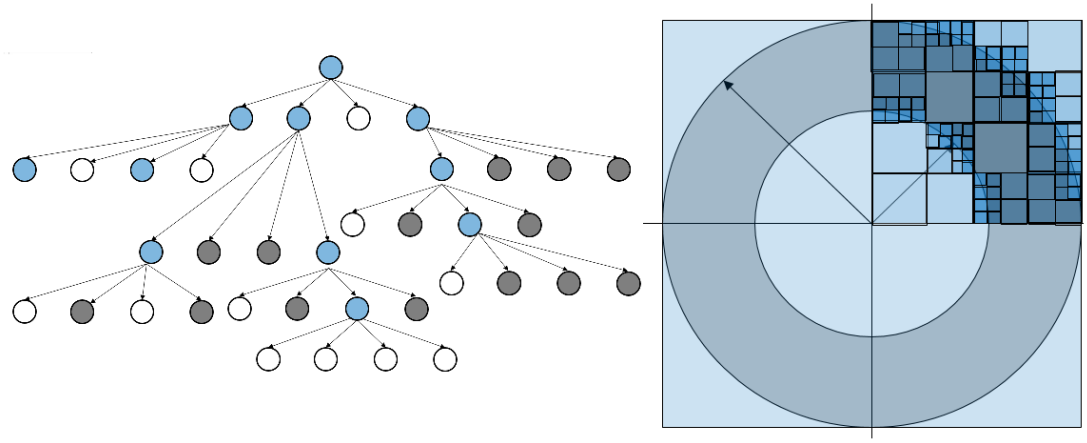

Figure 2.4: The **perimeter** benchmark. Computational pattern can be represented by an unbalanced quad tree.

power Another key benchmark for us is the power pricing computation benchmark. The benchmark is based on a composite hierarchical structure of C arrays and pointer-based linked lists of various objects: Root, Laterals, Branches and Leaves. Figure 2.3 illustrates the data structure. The algorithm is basically composed of folds and

reductions. C arrays are reduced and linked lists are folded. The algorithm works recursively and starts with the Root::compute() call. The method accumulates the power demand Demand(P,Q) from all lists of Laterals and does a reduction to a final Demand(P,Q) value. Accumulation starts with the end of each lateral list. Each lateral accumulates its power demand from all its branches and the latter in turn accumulate their power demand from all their leaves. The leaves call optimize_node() method, which does a chunk of scientific computations (gradients, vectors, etc.), which compute P and Q. Finally, Root::compute() is called iteratively from a while loop of power_pricing_problem(). Iteration continues up until P and Q error becomes less than the required epsilon. Before every simulation step the benchmark does an injection of coefficients into the structure. Every simulation step leaves computed side effects on the coefficients stored withing various objects.

The power benchmark operates with sequences and linked lists. Sequences are reduced and linked lists are folded. These patterns correspond to and inspire our Fold and Reduce computational frameworks.

treeadd The treeadd benchmark is a very small and simple one. It does a recursive reduction on a pointer-linked binary tree. To create a workload the reduction is done repetitively inside the loop.

The treeadd benchmark is a straightforward example of the fractal computational framework. Although, the workload is to small for its effective application.

2.7.3.0.1 mst Minimum Spanning Tree (MST) benchmark does an MST weight computation of a complete graph. Computation is approximate and the algorithm looks like it might finish with incorrect result. Nevertheless, the benchmark can be used for the purpose of computational workload. Graph is represented as a linked list of vertices. Every node in the list has a hash table of incident edges. Graph is complete: each vertex is connected to all other vertices in the graph (except itself). Algorithm repeatedly traverses the list of vertices and gradually accumulates the MST weight. On every traversal algorithm picks the node in the list to use as an input for the next traversal. In that sense, there is a cross iteration/traversal dependency. The code below summarises the benchmark.

```
vertex = list;
list = list->next;
while (num_vertices) {
    ret = traverse_vertex_list(vertex, list);
    mst_weight += ret.distance; // accumulate the final result
    vertex = ret.vertex; // next vertex to measure the distance
   against
    num_vertices--;
}
return mst_weight;
```

Listing 2.5: The main algorithm of mst benchmark.

2.7.3.0.2 tsp Travelling Salesman Problem (TSP). The benchmark generates a set of dots scattered on a 2D plane. Dots represent cities and are specified by their (x,y) coordinates. The TSP problem is to visit all the cities and return to the city of origin having passed the minimal distance. In other words, the algorithm returns a cycled sequence of cities, where proximity of elements in the sequence in terms of order means their spatial proximity on the 2D plane. The algorithm's work resembles that of an insertion sort. The sequence is divided into 2 parts. Ordered part and unordered part. At first, ordered part consists of just 1 element. On every iteration the algorithm takes the next element out of unordered part, finds it a pairing element inside the ordered part (with the minimal distance between them) and inserts the element into the ordered subsequence next to its pair. At the end we get the sequence with the property that closest dots stand the closest in the sequence. The benchmark is based on a binary tree being transformed into doubly linked list. Every node of the tree represents a city located on 2D plane with randomly generated (x,y) coordinates. The build_tree() method is written in a way to generate a uniform distribution of dots on the plane.

Chapter 3

Software Parallelization Assistant

3.1 Introduction

Parallel hardware is ubiquitous through the entire spectrum of computing systems, from low-end embedded devices to high-end supercomputers. Yet, most of the existing software is written in a sequential fashion. Despite decades of intensive research in automatic software parallelization [32], fully exploiting the potential of modern parallel hardware still requires a significant manual effort. Given the difficulty of the obstacles faced by automatic parallelization today, we do not expect that programmers will be liberated from performing manual parallelization in the near future [30].

This chapter introduces a novel parallelization assistant that aids a programmer in the process of parallelizing a program in the frequent case where automatic approaches fail. The assistant reduces the manual effort in this process by presenting a programmer with a ranking of program loops that are most likely to 1) require little or no effort for successful parallelization and 2) improve the program's performance when parallelized. Thus, it improves over the traditional, profile-guided process by also taking into account the *probability* of potential parallelization for each of the profiled loops.

At the core of our parallelization assistant resides a novel machine-learning (ML) model of loop parallelizability. Loops are compelling candidates for parallelization, as they are naturally decomposable and tend to capture most of the execution time in a program. Furthermore, focusing on loops allows the model to leverage a large amount of specific analyses available in modern compilers, such as generalized iterator recognition [44] and loop dependence analysis [40]. The model encodes the results of these analyses together with basic properties of the loops as machine learning *features*.

The loop parallelizability model is trained, validated, and tested on 1415 loops from

the SNU NAS Parallel Benchmarks (SNU NPB) [27]. The loops are labelled using a combination of expert OpenMP [11] annotations and optimization reports from Intel Compiler (ICC), a production-quality parallelizing compiler. The model is evaluated on multiple machine learning algorithms, including tree-based methods, support vector machines, and neural networks. The evaluation shows that – despite the limited size of the data set – using support vector machines allows the model to achieve a prediction accuracy higher than 90%. The model improves over the ICC Compiler across the sequential C version of the SNU NPB suite by detecting 13% more parallel loops. Albeit this improvement comes at the cost of introducing *false positives*, where non-parallelizable loops are misclassified as parallelizable. However, the false positive rate in our evaluation is as low as 6.5%. We feel this is acceptable, as our parallelization assistant does not automatically restructure code, but leaves the parallelization decision in the hands of the programmers.

The parallelization assistant combines inference on the parallelizability model with traditional profiling to rank higher those loops with a high probability of being parallelizable and impacting the program performance. An evaluation on eight programs from the SNU NPB suite shows that the program performance tends to improve faster as loops are parallelized in the ranking order suggested by our parallelization assistant compared to a traditional order based on profiling only. On average, following the order suggested by the assistant reduces by approximately 20% the number of lines of code a programmer has to examine manually to parallelize SNU NPB to its expert-level speedup. Given the high level of effort involved in manual analysis, such a reduction can translate into substantial development cost savings.

3.1.1 Motivating example

Consider the sequential C implementation of the *Conjugate Gradient (CG)* benchmark from the SNU NPB suite. Table 3.1 shows the top three CG loops as ranked by the Intel Profiler (based on their execution time) and by our parallelization assistant (additionally taking into account their parallelizability). Both rankings include the same loops, but, crucially, the loops are **ranked in a different order**. Following the profiler ranking (second column in Table 3.1), a parallelization expert would concentrate on analyzing the loop `cg.c:326` first. Analyzing this loop turns out to be costly (it consists of 100+ lines of code) and unfruitful (it is actually not parallelizable due to inter-iteration dependencies and side effects, see Listing 3.1). This analysis would be

Ranking	Profiler	Assistant	
	loop	loop	parallelizability
1	cg.c:326	**cg.c:509**	85%
2	cg.c:484	cg.c:326	29%
3	**cg.c:509**	cg.c:484	8%

Table 3.1: Comparison of the profiler and assistant rankings for the CG benchmark loops (limited to the top three loops).

followed by an equally unfruitful analysis of loop cg.c:484.

In contrast, our parallelization assistant (two last columns in Table 3.1) ranks the loop **cg.c:509** before loops cg.c:326 and cg.c:484, as it finds that the former has a significantly higher probability of being parallelizable (see last column in Table 3.1). Following the assistant's ranking, a parallelization expert would thus concentrate on analyzing the loop **cg.c:509** first. Analyzing this loop is inexpensive (it consists of only six lines of code, see Listing 3.2) and fruitful: its parallelization speeds up CG by a factor of 2.8, which is 70% of the speedup obtained by parallelizing the entire benchmark. Hence, following the ranking proposed by our assistant, a parallelization expert can achieve most of the available speedup in CG in a fraction of the time required by a traditional profile-guided parallelization process.

3.1.2 Contributions of our Loop Parallelization Assistant

In summary, this chapter makes the following contributions:

- ▷ We introduce a machine learning model, which can be used to predict the probability with which sequential C loops can be parallelized (Sections 3.2 and 3.3);
- ▷ we integrate profiling of execution time with our novel ML model into a parallelization assistant, which guides the user through a ranked list of loops for parallelization (Section 3.4); and
- ▷ we demonstrate that our tool and methodology increase programmer productivity by identifying parallel loop candidates better than existing state-of-the-art approaches (Section 3.5).

```
for (it = 1; it <= NITER; it++) {
  ...
  if (timeron) timer_start(T_conj_grad);
  conj_grad(colidx,rowstr,x,z,a,p,q,r,&rnorm);
  if (timeron) timer_stop(T_conj_grad);
  ...
  printf("    %5d       %20.14E%20.13f\n", it, rnorm, zeta);
  ...
}
```

Listing 3.1: `cg.c:326`. Longest running loop in CG. The loop cannot be parallelized due to inter-iteration dependences and side effects caused by system calls.

```
for (j = 0; j < lastrow-firstrow+1; j++) {
  suml = 0.0;
  for (k = rowstr[j]; k < rowstr[j+1]; k++)
    suml = suml + a[k]*p[colidx[k]];
  q[j] = suml;
}
```

Listing 3.2: **`cg.c:509`**. Longest running loop in CG among those *that can be parallelized*.

3.2 Predicting parallel loops

We approach the prediction of parallel loops as a *supervised probabilistic machine learning classification problem*. Based on sequential reference applications and their manually parallelized counterparts, as well as Intel C/C++ Compiler's (ICC) parallelization reports, we create a data set of parallelizable and non-parallelizable loops. We extract loop features and use the data set to train a machine learning model, which links feature vectors describing the loops with their observed parallelizability. We then use the trained model as a probabilistic predictor: for each new loop we determine its feature vector and then predict the probability of the loop being parallelizable [20]. For naturally probabilistic models like trees we directly use the computed classification probabilities (fraction of parallelizable training samples in the leaf node). For the support vector machines classifier, we use Platt scaling to derive the probabilities.

In this section we introduce the parallelizability model, whereas Section 3.3 presents a standard ML performance assessment including accuracy, precision and recall scores. Descriptions and definitions of the machine learning techniques we use can be found

in [35]. We used the *scikit-learn* library [26] for all ML related tasks.

3.2.1 Loop analysis & Feature engineering

For the purpose of machine learning, program loops are represented by numerical *feature vectors*. We derive these features using standard compiler analyses operating on the Program Dependence Graph (PDG) [4] of a loop. The PDG is a representation that captures both data and control information and is constructed using dependence analysis [14]. Figure 3.1 shows an example of a PDG for a simple loop, where SCC stands for *Strongly Connected Component*. We construct the PDG on a program's

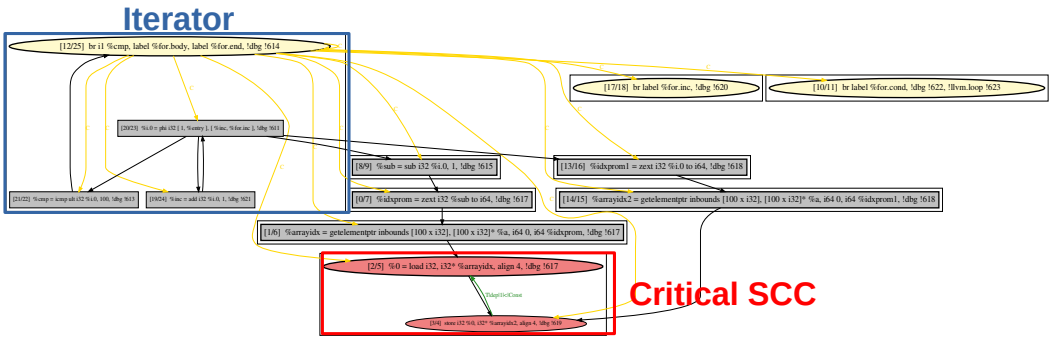

Figure 3.1: Example of PDG of a simple loop with a cross-iteration dependency.

LLVM IR using standard dependence analysis [14]. In addition, we use *generalized loop iterator recognition* [44] to separate *loop iterators* from *loop payloads*. This enables us to define and extract features relating to each of those loop components. In total, we extract a set of 74 static loop features which are based on structural properties of the PDG and the types of instructions constituting it. Table 3.2 summarizes these features.

Many features have simple and intuitive motivations behind them. Loop proportion features are backed up by the fact that larger loops tend to be harder to parallelize. Complex iterators include non-trivial cross-iteration transitions (e.g. linked-list update), unknown iteration numbers, etc. Payload SCCs introduce cross-iteration dependencies. Cohesion features characterize how tightly components of loops are coupled together in terms of the number of edges between them. Loop dependence features count the number of edges in different loop parts as well as their types. Loop instruction features characterize the loop's instruction mix, assigning more importance to memory reads/writes, calls and branches. Non-inlined function calls usually prevent

loop parallelization. Intensive memory work (memory read/write fraction features) complicates parallelization as well.

Feature groups	Features	Description
loop proportion	absolute size	number of LLVM IR instructions
	payload fraction	payload instructions / total loop instructions
	proper SCCs number	number of payload SCCs with more than 1 instruction
loop dependencies	number of PDG edges for different dependence classes: read/write order (*true*, *anti*, *output*), dependency type (*register*, *memory*, *control*), other (*cross-iteration*, etc.)	
loop cohesion	iterator/payload	$\frac{\text{edges between iterator/payload}}{\text{total loop edges}}$
	critical/regular payload	$\frac{\text{edges between critical/regular payload}}{\text{total loop payload edges}}$
loop instruction nature	numbers and fractions of different parallelization critical instructions (memory loads and stores, branches, calls, etc.)	

Table 3.2: Static features used for the characterization of loops.

3.2.2 Feature extraction

To extract all devised loop features from SNU NPB benchmarks and get the train/test data sets we developed a tool based on the LLVM compiler infrastructure [1][16]. The tool is a set of LLVM function passes working on the SSA-based LLVM IR and can be found on the GitHub [46]. The tool works by building data, memory and control dependence graphs (DDG, MDG, CDG) and combining them into the final program dependence graph (PDG) [4] for all loops found in program functions. Once all graphs are built we run the search of SCCs [44] on them and recognise loop iterators. The final step is to traverse all these graphs computing devised metrics (ML features) and dump all that information into the file to be later fed into scikit-learn based ML scripts.

3.2.3 Feature selection

The feature engineering task produced a quantitative description of program loops being characterised by feature vectors of length 74. To avoid overfitting, we discard irrelevant or redundant features using a pipeline of automatic feature selection methods

from the scikit-learn library. First, we eliminate features with a low variance score, then we fit a decision tree-based model and select features with importance scores above a given threshold. After that we repeatedly run *Recursive Feature Elimination, Cross-Validated (RFECV)* to improve accuracy, precision and recall scores. This yields the final set of features. Table 3.3 presents the 10 highest-ranked features in the set. SNU NPB benchmarks contain a lot of uninlined function calls and it is unsurprising that the amount of call instructions in the payload of a loop ranks the highest. Despite the absence of straightforward intuition behind cohesion metrics, they tend to correlate with loop parallelizability well. Loops heavy on memory writes also significantly affect the parallelizability property.

Feature	**Importance**
payload call fraction	23.5
iter/payload non-cf cohesion	18.5
payload mem write fraction	6.1
loop absolute size	5.7
critical payload pointer access count	5.3
payload memory dependence count	4.0
critical payload non-cf cohesion	2.9
payload pointer access fraction	2.7
critical payload total cohesion	2.6

Table 3.3: Relative importance of static loop features, ranked by fitting a tree-based ML model.

3.2.4 Model & Hyper-parameter selection

We evaluate several machine learning classification algorithms in our parallelization assistant, including tree-based methods like decision trees (DT), random forests (RFC) and boosted decision trees (AdaBoost); support vector machine classifiers (SVC) and multi-layer perceptron neural networks (MLP). Section 3.3.1 shows that these models perform similarly with SVC and MLP performing slightly better. For each ML model we use exhaustive hyper-parameter grid search and pick the grid node with the best cross-validation score on the validation set. The details of all ML pipeline stages are available in our repository [46].

3.2.5 Training data & ML model training

In order to train and test our ML model in a supervised way, first we need to provide it with the ”right answers” regarding loop parallelizability. In other words, we need to prepare a labelled training data set. For training our ML model we use a total of 1415 loops from the SNU NPB benchmark suite. Out of those loops, 210 have been annotated by (external) human experts with parallel OpenMP *pragmas*. Like [34] we use these annotations as labelled data to indicate parallelizable loops. However, the data is not complete. Human programmers strive to capture only coarse-grain parallelism and do not annotate every parallelizable loop. Hence, we augment the training data with the help of the ICC Compiler, which finds additional parallelizable loops. We combine the results into our final training set comprising a total 1415 loops, of which 995 are labelled as parallelizable. Then we use K-fold and Leave-One-Out Cross-Validation (LOOCV) methodologies to train and test our ML models.

To extract loop parallelizability labels from the Intel compiler's optimization reports we developed a parser [2], but the task has presented us with a number of technical challenges. Before ICC can actually parallelize or vectorize a loop, it applies a number of enabling loop transformations such as loop interchange, distribution, tiling, etc. These transformations help ICC to eventually uncover more parallelism. The detailed description of all these transformations can be found in [5]. Applied to a loop nest, these optimizations might significantly restructure and distribute the parts of a loop across the whole ICC optimization report. Moreover, ICC might parallelize only certain parts of transformed loop. In the end, we considered a loop to be parallelizable by the ICC compiler if the latter hasn't found any dependencies and either vectorized or parallelized it. In the case of distributed loops, all parts must be parallelizable for an original loop to be considered as such. For a final correctness we conducted a manual verification on top of automatically extracted results.

Table 3.4 presents a parsing report, which summarises the number of times ICC applied a certain optimization. The major cells are *parallel* and *icc*, which report the total number of truly parallelizable loops and the number of loops parallelized by ICC. As can be seen, ICC compiler does not exploit all the parallelism available in SNU NPB benchmarks. Section 2.2.2 presented a study of the reasons why ICC fails to parallelize certain loops. To obtain the data we used ICC 18.0. As we are interested in obtaining the most complete set of parallelization labels, and not in the optimal program running time, we instructed ICC to do the most aggressive parallelization, i.e. parallelize a loop,

Labels		Intel Compiler (ICC)			
		Optimization		Parallelization	
loop	ranking	loop	ranking	parallel	
total loops	**1415**	distrs	34	parallel	653
parallel	**995**	fusions	214	vector	737
icc	**812**	collapses	58	parallel deps	535
openmp	**210**	tilings	27	vector deps	266

Table 3.4: Report on loop classification labels derived from expertly added OpenMP annotations of SNU NPB benchmarks and ICC optimization reports. Out of all 995 parallelizable loops the ICC discovered and parallelized only 812.

whenever it can be parallelized, ignoring ICC cost model and independent of the expected profitability of loop parallelization (*-parallel -vector -par-threshold0* options). To make ICC uncover all possible parallelism we ran it with a wide set of enabling loop transformations (*-O3* option) applied prior to the parallelization. All the options had a matching hardware support: we installed ICC on the Ubuntu 18.04 machine with 4 Intel Core i5-6500 CPUs having a vectorization support all up to AVX2.

3.3 ML predictive performance

In our work we employ 2 cross-validation (CV) techniques. We evaluate the overall predictive performance our trained ML model is capable of achieving on SNU NPB benchmarks using K-fold CV. To deploy our assistant against single benchmarks of the suite and assess its effectiveness (Section 3.5) we have to use a modified Leave-One-Out CV.

3.3.1 Overall model performance

Table 3.5 shows the overall predictive performance of different ML models measured with K-fold CV on the whole SNU NPB data set. Training and testing have been done for different values of K (5, 10, 15, 20, 25, 30) and the accuracy remains stable across the entire range. The same is true of recall and precision scores. We used the baselines (constant ”parallelizable” prediction and uniform) available in scikit-learn to compare our models against. The SVC model has the highest average accuracy and successfully manages to recall 95.24% of all parallel loops. The ICC Compiler suc-

ML model	accuracy	recall	precision
constant	70.32	100	70.32
uniform	46.27	41.50	69.79
SVC	90.04	95.24	91.06
AdaBoost	86.96	92.92	89.06
DT	84.36	89.57	87.90
RFC	86.65	93.22	88.47
MLP	89.40	93.77	91.39

Table 3.5: Average predictive performance of different ML models measured with K-fold CV on the whole set of 1415 SNU NPB loops.

ceeds in parallelizing 812 out of 995 parallelizable loops available in SNU NPB. Thus, on average SVC extends the ICC Compiler's parallelization capabilities to 948 loops. Figure 3.2 shows that out of the 10% of mispredictions that SVC makes, 65% are false positives. Hence, the average unsafe error rate is 6.5%. In this project though we devise a scheme, that protects us and makes these errors not critical.

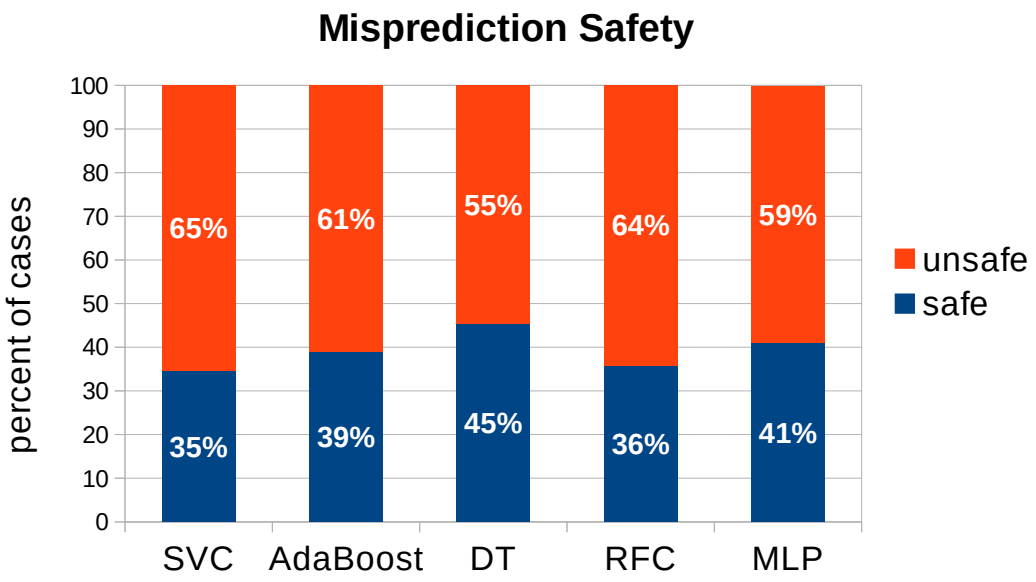

Figure 3.2: Breakdown of misclassification errors.

3.3.2 Model performance within assistant

Our proposed assistant (Section 3.4) is trained and tested using LOOCV rather than K-fold CV. Instead of treating the entire set of loops from all SNU NPB benchmarks

as a single data set, in this context we train the model on nine benchmarks and test it on the remaining one. Doing so completely excludes the loops of the target benchmark out of a training set, but allows us to get predictions for all benchmark loops, parallelize them if advised so, and test the effectiveness of our assistant. The drawback of this scheme is that it might potentially reduce the accuracy if the nature of loops in the target benchmark dramatically differs from that of loops seen in the training set. Figure 3.3 compares LOOCV accuracy against that of K-Fold CV for all SNU NPB benchmarks, where K-Fold CV is conducted on a data set consisting of loops from a single benchmark only. The comparison proves that lower LOOCV accuracies are attributed to a reduced training data set and not to our ML model.

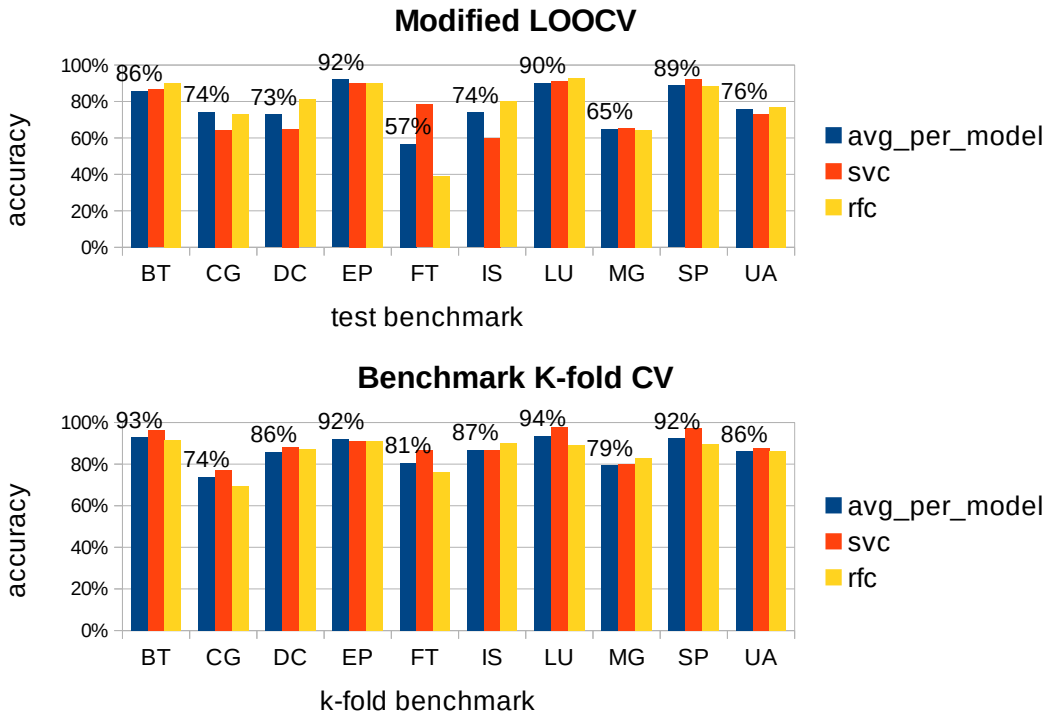

Figure 3.3: Prediction accuracy measured with modified LOOCV and compared against that of K-fold CV for single benchmarks.

3.4 Parallelization assistant

The ML-based predictor developed and assessed in the 2 previous sections can have a real practical application. Due to the statistical nature inherent to all machine learning techniques it is impossible to eliminate all prediction errors completely. While false negatives might just miss available parallelization opportunities and lose some perfor-

mance, false positives can break the program and are the most critical in the context of our ML problem. Given that, we develop a parallelization assistant, which does not seek to replace programmers, but aims to assist and increase their productivity. The predictor is the core component of our novel parallelization assistant. The assistant incorporates prediction results and combines them with profiling information. It then produces a ranking of all loops in an application to guide a programmer towards the most beneficial loop candidates for their *manual* parallelization effort.

Loop Ranking. The loop ranking computed by our parallelization assistant combines a loop's contribution to the overall program execution time with its predicted probability of being parallelizable. In particular, we obtain the ranking by applying a shifted sigmoid function to the predicted probability multiplied by the application runtime fraction a loop takes to run as shown in Figure 3.4. The intuition for using this

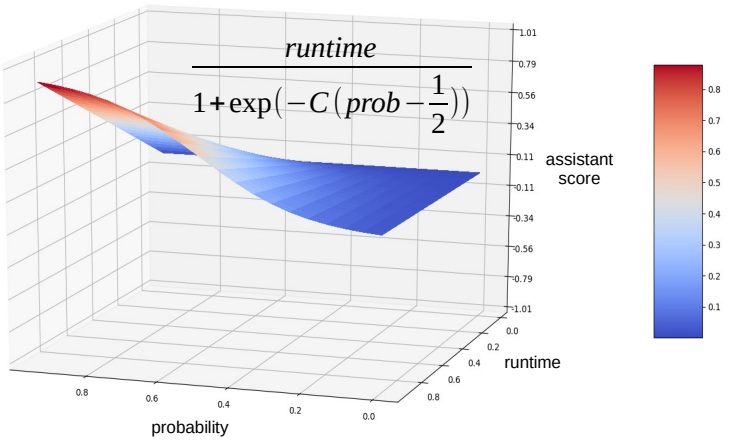

Figure 3.4: For each loop the ranking function combines its contribution to the application's execution time and its predicted probability of being parallelizable.

function to combine the two metrics is that it prioritizes parallelizable long-running loops and scales down the weight of non-parallelizable loops irrespective of their contribution to execution time. The effect of this can be seen in Figure 3.5 for the loops of the FT benchmark. If programmers attempt to parallelize loops in the order prescribed by their execution time, they will inevitably waste their efforts trying to parallelize loops which may be long-running, but offer little or no opportunity for extracting parallelism. Instead, by taking into account predicted parallelizability our ranking directly guides the programmer towards loops that significantly contribute to overall execution time **and** offer a realistic prospect of parallelization.

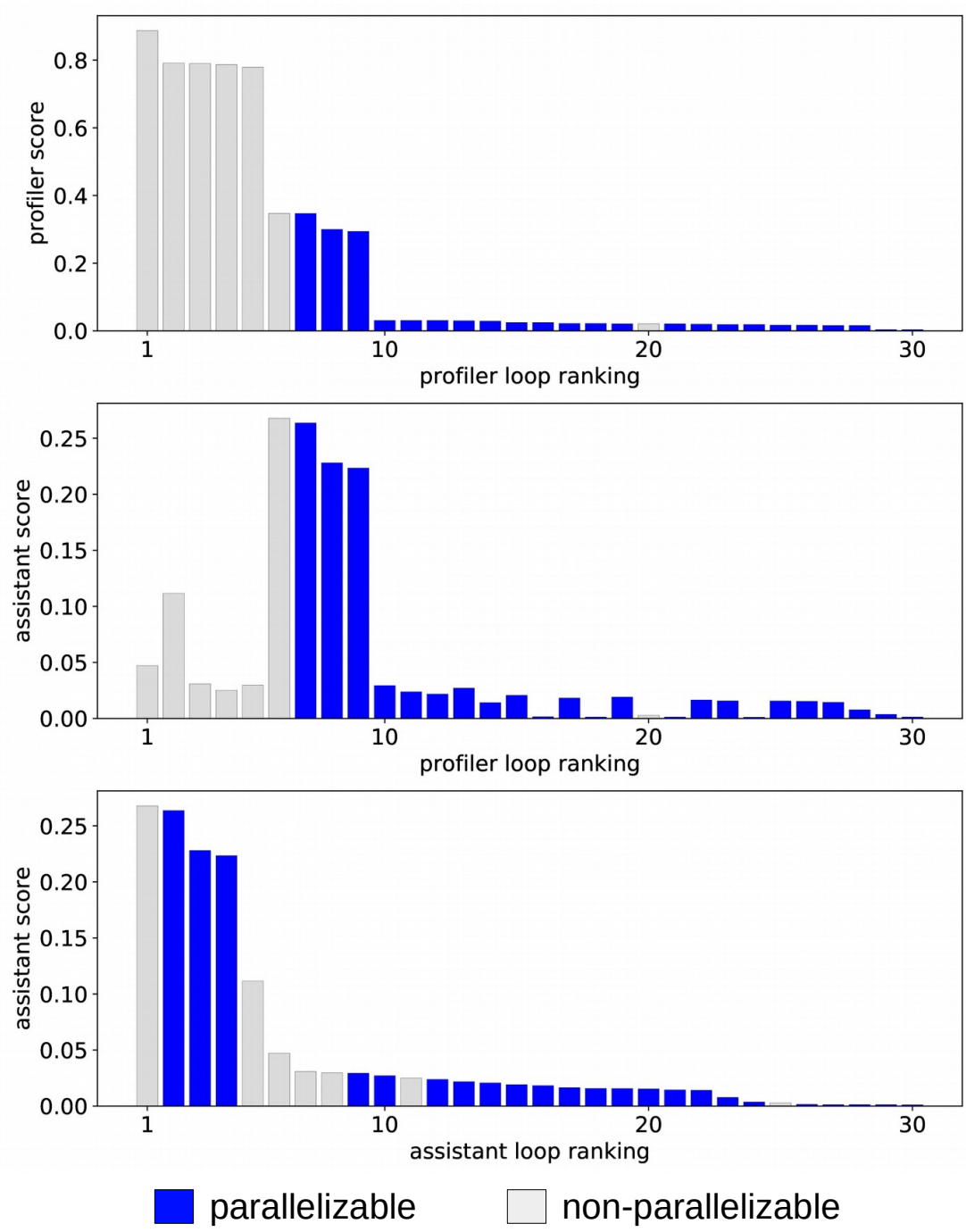

Figure 3.5: Change of loop rankings with the application of the assistant ranking function for the 46 loops of the FT benchmark. Ranking based on loop execution time alone (top figure) results in some high-ranked, but non-parallelizable loops. Combining profiled execution time and parallelizability in a single score (middle figure) results in a ranking that prioritizes parallelizable loops (bottom figure).

3.5 Assistant evaluation

3.5.1 Comparison to static analysis

We have compared the generated loop parallelizability classifications of our assistant against that of the ICC Compiler, which due to its use of static analysis is conservative and occasionally misses some parallelization opportunities. The ML approach to parallelization with a human programmer responsible for final code transformation allows our parallelization assistant to be more aggressive than the ICC Compiler. In other words, our model can discover more parallelizable loops.

We have set up an experiment where we apply our ML predictor side-by-side with the ICC Compiler, which has been configured to do the most aggressive parallelization (see Section 3.2.5) on the same machine, that provides all hardware support needed. Both, the assistant and the ICC aim at independently classifying loops as parallelizable or not. There is a total of six possible classification combinations that our scheme might produce. Figure 3.6 shows their relative distribution as a pie chart. To calculate the relative frequency of different loop classification combinations illustrated in Figure 3.6 we repeatedly ran K-fold CV on the whole set of SNU NPB loops and sorted the outcomes into separate classification buckets. In the "agreement" cases, which number around 80% of cases the ICC Compiler and ML predictor identically classify truly (non-)parallelizable loops as (non-)parallelizable. This is an expected result, since we used ICC (along with OpenMP annotated loops) to train our ML model. The "missed opportunity" cases, where both ICC Compiler and ML predictor miss parallelizable loops also represent the agreement and are not interesting. The most interesting cases though are those, where ML predictor and the ICC Compiler disagree. While ICC Compiler is conservative and will never classify a non-parallelizible loop as parallelizible, the statistical ML predictor can make a "false positive" error. The rate of false positives in this experimental setting is 8%. That works in the opposite direction as well. The ML predictor can discover truly parallelizible loops, which escape compiler analysis. The rate of such cases is 10%. These cases are classified as "discovery" and have been manually checked in the source code of SNU NPB. The results are summarized in Table 3.6, which reports on the reasons behind ICC conservativeness. False negative mispredictions make ML predictor miss some real parallelization opportunities, but in the fraction of these cases the ICC Compiler can catch them and "shield" the ML predictor.

Reason	Num	Reason	Num	Reason	Num
missed reduction	18	array privatization	7	conservative analysis	60
unknown iteration number	7	static dependencies	46	too complex	22
non-inlined calls	4	other	4	total	168

Table 3.6: Classification of parallelizable loops rejected for parallelization by the ICC Compiler.

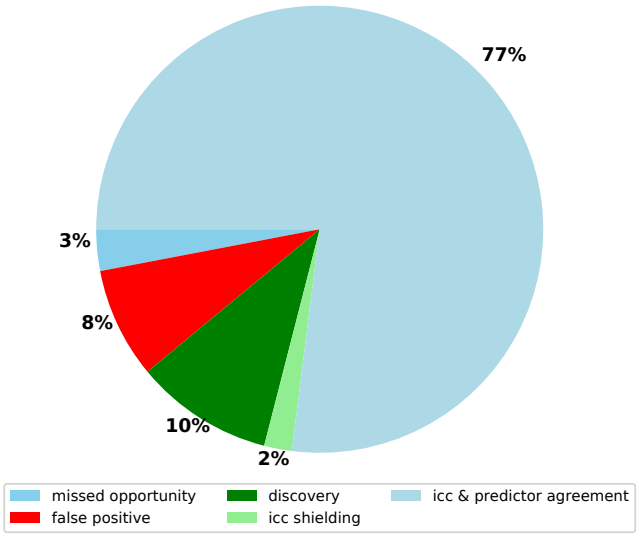

Figure 3.6: Distribution of loop classifications by the ICC Compiler and our predictor.

3.5.2 SNU NAS parallelization

In this section we evaluate the effectiveness of our parallelization assistant. In particular, we are interested in the potential programmer productivity gains delivered by our tool and savings on human expert time. Our study assumes that the human expert starts with a sequential version of the SNU NPB benchmarks. The goal is to parallelize these applications to a performance level matching that of their existing parallel versions. By using our assistant we expect the human expert to consider fewer loops than by using a profiling-based approach, i.e. considering loops in decreasing execution time.

For this experiment we used a desktop Ubuntu 18.04 machine and compiled the benchmarks with the Intel C/C++ Compiler (ICC) 18.0. We compiled the serial versions with *-O3* and *-ipo* flags and explicitly prohibited any vectorization (*-no-vec*) or parallelization (*-no-par*). The OpenMP parallel versions have also been compiled with *-O3* and *-ipo* flags and prohibited automatic parallelization. When we followed the

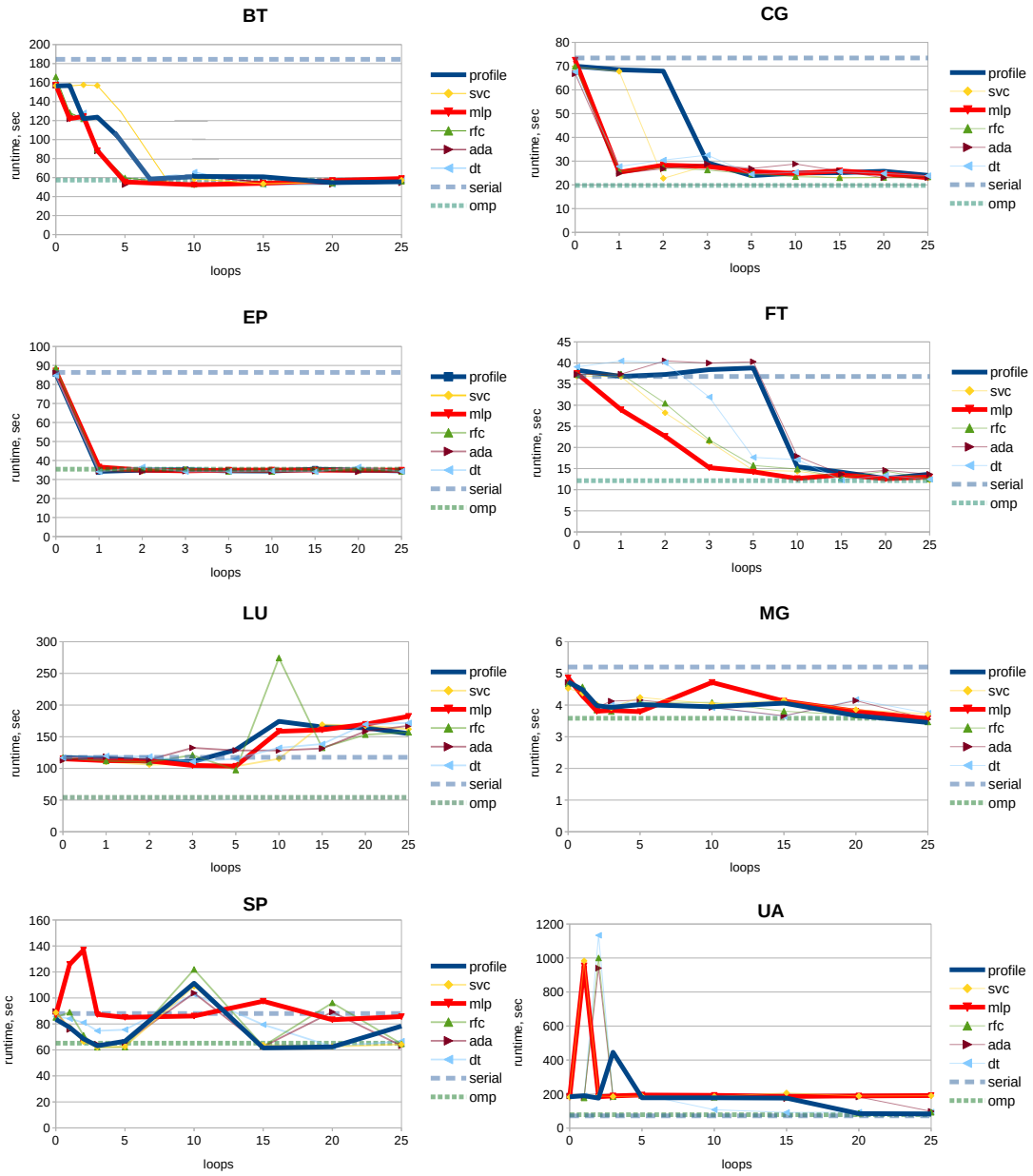

Figure 3.7: From left to right more loops are parallelized for each benchmark. As we parallelize more loops, program execution times improve over the initial sequential performance and reach the performance level of the reference OpenMP implementations. Our ML based parallelization assistant requires the user to parallelize fewer loops than a purely profile-guided approach to reach the maximum parallel speedup.

rankings and parallelized benchmarks loop by loop we used OpenMP pragmas and compiled the code with *-O3* and *-ipo* flags as well. Benchmark running times have been measured with the help of UNIX time() utility. To minimize the errors, we ran

experiments several times and took the mean average. The machine has 4 Intel Core i5-6500 CPUs with 3.20 GHz frequency and vectorization support up to AVX2. The RAM is 16Gb.

Figure 3.7 summarizes our results as performance convergence curves. For each benchmark the curves plot its execution time (y-axis) as a function of the number of analyzed and possibly parallelized loops (x-axis). The runtime is bounded within the runtime of the serial execution (top dashed line) and the time of the reference parallel OpenMP version (bottom dashed line). Our goal is to reach the performance of the reference parallel versions of each program by parallelizing them one loop at a time following the rankings offered by our assistant and the profiler. The neural network based MLP model of our assistant provides the fastest overall performance convergence. While there is some variation depending on the ML model used for the parallelization prediction, in general ML-assisted parallelization outperforms or equals the profile-guided schemes across all benchmarks.

Following the rankings of our assistant in parallelizing the BT, CG and FT benchmarks, we reach their maximum potential performance faster. For BT, maximum parallel performance can be reached after the user has parallelized the first three loops (3061 LOC in Table 3.7) suggested by our assistant, while profile-guided parallelization requires 6 loops (6122 LOC) to be parallelized first before reaching the same performance level. For CG, if we follow the suggestion of our assistant to first parallelize a small loop (6 LOC), we are able to achieve 70% of the maximum potential speedup (see Section 3.1.1). On the other hand, using the profiler ranking requires examining three loops, totalling to 330 LOC, to yield the same performance gains. Moreover, for the SP and UA benchmarks, some of the assistant rankings require the programmer to examine more loops than the profiler. However, the loops proposed by the assistant in the UA benchmark are actually simpler, since they consist of fewer LOC – 508 LOC for MLP and 579 for DT versus the 882 LOC offered by the profiler. By following our assistant's suggestions, a programmer would be required to examine 20% less LOC on average across all models.

In some cases, partial benchmark parallelization might result in a slowdown. In the case of LU, after having parallelized the first 25 loops we do not converge to the best achievable parallel version performance. There are a total of 40 OpenMP pragmas in the benchmark and we need to parallelize all the respective loops to reach the best performance level. In the case of UA, all rankings suggest analyzing a long-running innermost loop first. Its parallelization actually increases running time due to a syn-

Bench.	Bench. Runtime, *sec*			Speedup, *times*		*Loops Number*$_{LOC}$					
	Serial	**OpenMP**	**Critical**	**OpenMP**	**Critical**	**Profile**	**SVC**	**MLP**	**RFC**	**AdaBoost**	**DT**
BT	158.76	57.36	56.57	2.77	2.81	6_{6122}	8_{6392}	4_{4088}	5_{5105}	3_{3061}	5_{5105}
CG	69.38	19.77	25.06	3.51	2.77	3_{330}	2_{118}	1_{6}	1_{6}	1_{6}	1_{6}
DC	698.82	254.29	698.82	2.75	1.00	∞	∞	∞	∞	∞	∞
EP	86.35	35.40	35.07	2.44	2.46	1_{45}	1_{45}	1_{45}	1_{45}	1_{45}	1_{45}
FT	36.81	12.13	14.69	3.03	2.51	9_{338}	4_{187}	3_{140}	4_{187}	9_{338}	5_{193}
IS	4.75	1.35	4.63	3.53	1.03	∞	∞	∞	∞	∞	∞
LU	115.46	55.00	140.53	2.10	0.82	∞	∞	∞	∞	∞	∞
MG	5.20	3.58	3.94	1.45	1.32	3_{43}	3_{43}	3_{43}	3_{43}	3_{43}	3_{43}
SP	86.65	65.19	62.90	1.33	1.38	3_{801}	3_{801}	∞	3_{801}	3_{801}	20_{1257}
UA	71.82	78.56	189.66	0.91	0.38	19_{882}	30_{918}	30_{508}	19_{861}	22_{883}	10_{579}

Table 3.7: SNU NPB benchmark parallelization reports. The left part of the table shows execution times of serial, OpenMP and partially parallelized (critical) versions. The partially parallelized versions have only several critical (top ranked) loops parallelized. The right hand part of the table shows the number of top-ranked loops one needs to parallelize in order to reach the critical performance. The Profile column gives the reference number a profiler requires. The total lines of code (LOC) in the loops are written down as underscript. In most cases, ML based models converge to the critical performance faster than a profiler based approach. There are also cases where a profiler outperforms our assistants.

chronization barrier being introduced at a wrong program point. It takes 30 loops for the MLP model to achieve the parallel version performance.

Finally, we observe that neither our parallelization assistant nor the profiler reach the performance of the reference OpenMP versions on the DC and IS benchmarks. Manual inspection reveals that these benchmarks have been parallized using OpenMP parallel sections, but do not contain any OpenMP parallel loops. Both our parallelization assistant and the profiler incorrectly suggest to parallelize some of the benchmark loops, though. Table 3.7 summarizes the results of applying our parallelization assistant to the SNU NPB suite.

Chapter 4

Computational Frameworks

4.1 Overview

The problem of *software parallelization* is multifaceted. In order to arrive at the final solution a programmer has to work on multiple conceptual levels starting from problem decomposition and algorithm choice down to software architecture design, data structure choice and finer grained low level loop optimizations. If our loop parallelization assistant (see Chapter 3) aims at reducing programmer's efforts at the finest granularity level, *computational frameworks* we propose in this chapter alleviate the tasks of software architecture design, as well as algorithm and data structure choice for parallelization of certain types of applications. The central motivating ground for the concept of computational frameworks and the prototype library implementing it is formed by the *Data-Centric Parallelization (DCP)* problem, limitations of the background work tackling the problem and the properties of the real world legacy codes.

Motivating ground for Computational Frameworks

The grand problem of Data-Centric Parallelization (DCP) Among all lower level implementation questions, the problem of successful data structure choice stands particularly prominent and important. Listings 4.1 and 4.2 illustrate how easily parallelization can be hampered. Both implementations solve an embarrassingly parallel problem of incrementing every sequence element by one. In the linear array based implementation compiler knows all element addresses statically and can generate parallel code. In the linked list based implementation compiler does not know element addresses in advance, since the latter will be resolved only dynamically and can generate only sequential code.

```
for (int i=0; i<1024; it++) {
  a[i]=a[i]+1;
}
```

Listing 4.1: Parallelizable loop operating on a **linear array**.

```
for (p=list; p!=NULL; p=p->next) {
  p->value+=1;
}
```

Listing 4.2: Non-parallelizable loop operating on a **linked-list**.

If a programmer had a tool, that could automatically recognize the type and properties of data structures and automatically substitute them with a simpler, parallelizable and more suitable alternatives, that would make the parallelization process a lot easier. Unfortunately, as our state of the art overview discovered, the solution to the problem is not available yet. There are a lot of various techniques and methodologies with their specific limitations and problems.

Limitations of "The state of the art" work The discovery of higher level entities in programs is definitely not a novel idea. There are various *static* and *dynamic* techniques in the literature. *Shape analysis* is one of the most well-known static techniques, which aims at understanding of various heap-allocated data structures and their properties (node reachability, cyclicity, disjointedness, etc.) at compile time [9][12][13]). *Shape analysis techniques give a very rough and conservative approximation (a tree instead of a linked list), work with high error rates (DAG instead of graph with cycles) and are provably undecidable [10].* One of the more recent static techniques is based on pattern matching on the code intermediate representation (IR) level and aims at recognition of various computational idioms (such as reductions, stencils, sparse and dense linear algebra computations, etc.)[39][42][43]. Computational idioms are specified in a constraint-based domain specific programming language CAnDL. *The technique allows for a rapid prototyping of new compiler optimizations based on pattern recognition and its substitution with an optimized versions of matched idioms, but it is limited for a relatively simple computational idioms and entities.* All static methods are limited in their program view to a single compilation unit. The challenge of data structure recognition requires a much broader view. Dynamic techniques come the closest to the solution of the DCP problem. There are numerous works available [37][41]. These techniques are based on program instrumentation and construction of various dynamic memory allocation graphs. The idea is that these graphs along with their update operations can reflect the actual shape of the data structure. Probably, the most promising and the best performing of all is the work of a Data-structure Detec-

tion Tool (DDT) [22]. The DDT tool can successfully recognise data structures in most of the standard libraries, such as STL, Apache (STDCXX), Borland (STLport), GLib, Trimaran achieving almost perfect recognition accuracy. Moreover, the technique has been able to recognise linked lists in Em3d and Bh Olden benchmarks. But, it is still far from solving the problem for an arbitrary real world code.

Data structure and algorithm inseparability An illustrative example in Listings 4.1 and 4.2 is obviously far below the complexity level of the real world code. The work with data structure can be spread all around the code base: allocation and initialization happen in one translation unit, update operations on the data structure are scattered between various functions in multiple other translation units. The exact way an operation works determines the properties (cyclicity, reachability, etc.) of a data structure and ultimately its type. It is crucial to understand how an algorithm actually calls data structure update operations. *This all points to the inseparability of algorithms and data structures: understanding the type of the data structure might require understanding of the algorithm and vise versa; and the data structure substitution might lead to algorithm transformation.* Indeed, as our feasibility studies with the SPEC CPU2006 benchmark suite have shown (Section 2.7.2) let alone automatic techniques, it might take some weeks for even an experienced human software engineer to understand what kind of data structures a benchmark uses and how to optimize it.

The ongoing trend to higher abstraction levels For decades there has been an ongoing trend in the process of software engineering to move up in the levels of abstraction from bare hardware to higher level concepts closer to human reasoning and understanding. We have moved from assembly languages to languages like Fortran and C, followed by the development of object-oriented languages and a supplement of imperative programming languages with functional programming concepts. All these steps increase programmers productivity, improve program structuredness and modularity and move the process of software design closer to a human level. The trend of moving to higher abstraction levels is not only true for software engineering in general, but for parallel software engineering in particular. For example, standards like POSIX, OpenMP and MPI aim at abstracting a programmer from various hardware and operating system details and work on the level of platform agnostic parallel programming models. Parallel algorithmic skeletons [31] (see Section 2.6) move software parallelization process even higher and allow a programmer to specify a computation on the algorithmic level with various concepts like *map*, *stencil*, *divide and conquer*, etc.

A higher level solution

The task of separating data structures from algorithms for an arbitrary real world code seems infeasible at the moment. Moreover, for some applications and benchmarks it does not seem meaningful either. For example, the suite of Olden benchmarks is much simpler, than SPEC CPU2006, but also exhibits the same inseparability problem. For many of Olden benchmarks the data structures and algorithms are blended together, but the union they form can be framed into an elegant higher level entity, that can later be used to parallelize the benchmarks in a nice and structured way.

In this thesis we propose a novel notion of *computational frameworks*, which exploit this possibility and help a programmer with a coarse-grained parallelization tasks, such as software design, algorithm and data structure choice on applications, where computations can be expressed with our computational frameworks. We describe the concept and the major frameworks in Section 4.3. We provide a prototype C++ template library implementing the notion [48] (see Section 4.3.5). We prototyped the library on the suite of Olden benchmarks, which inspired and shaped the concept. The parallel library version consistently outperforms the sequential version hitting 5-6x speedups on the major benchmarks (see Section 4.4).

4.1.1 Contributions of our Computational Frameworks

▷ We propose a novel notion of *computational frameworks*, which are higher level entities that embody both data structures and algorithms; and

▷ report on a prototype C++ template library [48] implementing the notion in a modern, convenient, parallel and an easy to use way;

▷ We expressed the computations of some Olden benchmarks in terms of our computational frameworks and have rewritten their original sequential legacy C versions with the help of our library in a modern, better structured, portable and crucially parallel way (see Section 4.3);

▷ We demonstrate the potential of the notion and the performance of the prototype library on the suite of Olden benchmarks (see Section 4.4), achieving consistent parallel speedups of 5-6x on the major benchmarks;

▷ Finally, we propose an idea of alternative software parallelization approach based on our computational frameworks as a future work.

4.2 Usage example

To get an initial feeling for what it is like to use our computational frameworks from a user perspective, let's consider a motivating example. A full discussion of the concept follows in Section 4.3. Suppose we want to calculate a well-known functional programming concept of the *left fold*, which also updates its elements with corresponding folded values, thus leaving some *side effects*.

We can code that simple computation using the C++ Standard Template Library (STL). Listing 4.3 shows the code implementing the task with an STL $list <>$ class template. First, we construct a list with 5 elements and initialize them to the value of 1. Then we loop through the list updating its elements and return the final left folded value. The task is small and the code is concise, but it is still possible to see its drawbacks. The code does not clearly separate concerns: list traversal, result computation and list transformation all happen in the same place and are mixed up together. For a better structuredness and comprehensibility it would make sense to put these different pieces of functionality into separate places. Alternatively, we could use STL's $std :: accumulate()$ function template. The latter provides a programmer with a more concise and abstract interface, but does not allow any side effects. We would still need to write a separate chunk of code aimed at implementing an update of the list elements.

Listing 4.4 shows an alternative implementation of the left fold using our **Fold** computational framework. The main computation here is expressed with a single line of code and a reader familiar with the concept of fold will comprehend the purpose of the code instantly. The fold defines the backbone of the computation, which is hidden from a user, while the definition of the custom part is left for a user to complete and is passed into a higher order functional style interface $Fold < Elem >:: compute()$ as a function object $ComputeFunc$, which has been derived from a Fold-specific base class $Fold < Elem >:: ComputeFunction < int >$. The latter is a template with the main computational result specified as a parameter. The base class frames the interface. Operator $Fold < Elem >:: ComputeFunction < int >:: operator()(Elem\&, int)$ takes an element of the Fold as an argument and combines it with the value folded so far in a user defined way. The user is supposed to pass the result further and has a freedom to update the element, thus leaving some side effects. The code in Listing 4.4 has a clear structure, but might feel a bit heavy for such a simple computation, but when a computational task is significant the advantages of the shown code design will become clear.

```
#include <list>
using namespace std;

int main() {
    list<int> lst(5, 1);
    // lst = [ 1 <- 1 <- 1 <- 1 <- 1 ]

    int result = 0;
    for (auto it = lst.rbegin(); it != lst.rend(); it++) {
        *it += result;
        result = *it;
    }
    // lst = [ 5 <- 4 <- 3 <- 2 <- 1 ]
    // result = 5

    return 0;
}
```

Listing 4.3: Left fold computation using standard STL list class template

```
#include "Fold.h"
using namespace abstract;

class Elem : public Fold<Elem>::Element {
    public:
        void grow() override { value = 1; }
        int value;
};

class ComputeFunc : public Fold<Elem>::ComputeFunction<int> {
    int operator()(Elem& elem, int fold) override {
        elem.value += fold;
        return elem.value;
    }
};

int main() {
    int fold_depth = 5;
    Fold<Elem> fold(fold_depth);
    // fold = [ 1 <- 1 <- 1 <- 1 <- 1 ]

    int result;
    ComputeFunc comp_func;
```

```
    result = fold.template compute<int>(comp_func);
    // fold = [ 5 <- 4 <- 3 <- 2 <- 1 ]
    // result = 5

    return 0;
}
```

Listing 4.4: Left fold computation using our Fold computational framework

4.3 Computational Frameworks

In this section we describe the general concept of **computational framework** along with all frameworks we propose and implement in the prototype library. As our frameworks have been inspired by computations found in the Olden benchmark suite, it would be helpful to review the Section 2.7.3 describing the key benchmarks.

4.3.1 The concept

As we have already mentioned the concept of *computational framework* has been inspired by the current problems in the software parallelization field along with computational patterns we have observed in the suite of Olden benchmarks. Very often programs are written with sub optimal from the point of software parallelization data structures. It might be hard to substitute them with parallelizable alternatives as the former might be closely entangled with algorithms the program is based on. Hence:

▷ Computational frameworks grow on the problem of data structure and algorithm inseparability;

We might keep the two together, but still tackle the problem at a higher level. We call the higher level entity we operate with a computational framework. Computational emphasizes the algorithmic component, while framework hints towards the underlying data structure.

▷ Computational frameworks combine algorithms and data structures to form a higher level entity;

Moreover, there are some specific problems, that could be tackled more effectively with specialized constructs. Imagine some task of scientific simulation. The scientific computation might be abundant with various higher level algorithmic constructs like

maps, reductions, folds, stencils, etc. These concepts are characteristic of functional programming. The latter often forbids any mutable states. At the same time simulations often keep a significant state being updated and accumulated with every step. The presence of state is characteristic to an imperative programming. Here one can see a contradiction and hence the gap to fill:

▷ Computational frameworks fill the gap between imperative and functional programming paradigms;

With all the things said above it is human programmers, who are going to use the concept in the end. Hence, it must be human friendly and convenient. At the same time the code must be modern and effective. We designed an object-oriented library with a functional style interface consisting of higher order functions. While the algorithmic component of the given framework is immutable there is a custom user defined functionality to apply to the framework. That functionality comes as a function object through the framework's interface in accordance with a command and template method (see Section 2.5.2) software design patterns.

▷ Computational frameworks provide a modern and convenient interface with elements of both functional, as well as object-oriented programming; and

▷ Computational frameworks embody the best ideas of various software design patterns.

When the software architecture has been designed, the exact efficient algorithm has been chosen along with all the most optimal and suitable data structures supporting its operation, a programmer can move onto the task of software parallelization. Here:

▷ Computational frameworks implement an effective and portable parallelization under the hood.

To conclude:

▷ Computational frameworks improve program **structuredness**, **modularity**, **separation of concerns** and hide possible **program parallelization** behind the **modern and convenient** user interface.

4.3.2 Fractal

The **Fractal** computational framework is the key framework in our work. It has been inspired by the theoretical work on tree reductions [29] and 3 Olden benchmarks

(*health*, *treeadd* and *perimeter*) with a very similar computational pattern. Although, the work on tree reductions has laid the theoretical foundation, but it did not provide a practical implementation. While there are numerous computations, processes and examples, which can be framed into the fractal, for the simplest and the most illustrative example let's consider an n-ary tree data structure being processed as follows.

Fractal's growth. First, we grow the tree from the root down to its leaves. A node takes a seed to grow, grows and spawns the next set of seeds for all its children nodes. The latter continue the process of growth. In the most general case the growth can happen asymmetrically, i.e it can stop on some paths down the tree, when some user specified condition is met for some nodes. In other words, the tree becomes incomplete with some leaves growing deeper than others and some nodes not having some children. We saw that pattern in the *perimeter* benchmark (see Section 2.7.3), when squares fell completely outside or inside the ring, they stopped to split further. Benchmarks *health* and *treeadd* grow complete trees and do not specify any stop conditions. While the algorithmic pattern of growth is fixed, the node's growth itself along with the node's data are custom and user defined parts. The same is true of the type of seeds and seed spawning procedures. Moreover, the growth of the tree along its various branches happens independently and can be parallelized.

Fractal's processing. When the tree is grown, we can start to process it. On that stage we go in the opposite direction from the bottom up by starting to process the leaves of the tree first. For every node we do some custom computation along with updating its custom data and pass the computation's result up to its parent. Before we process any node we need to process all its children first. The processing of different children happens independently and can safely be parallelized. When the tree has been processed we obtain the final result from its root along with all the side effects left in the nodes of the tree. The user defines these side effects per node in the custom processing function along with defining the exact computation from children up to the parent. We can repeat the process numerous times. This is the way some simulations work. The side effects are going to accumulate and make the nodes heavier. Figure 4.1 illustrates the pattern described above, as well as proposes an object-oriented design of the fractal concept. The backbone components (circles and arrows) of the tree are immutable and keep the main computational and growth patterns. These components can be implemented as base classes. Along with the backbone logic implementation these base classes provide a customizable interface to be overriden by derived classes (stars). Derived classes specify custom data contained in the tree nodes along with the

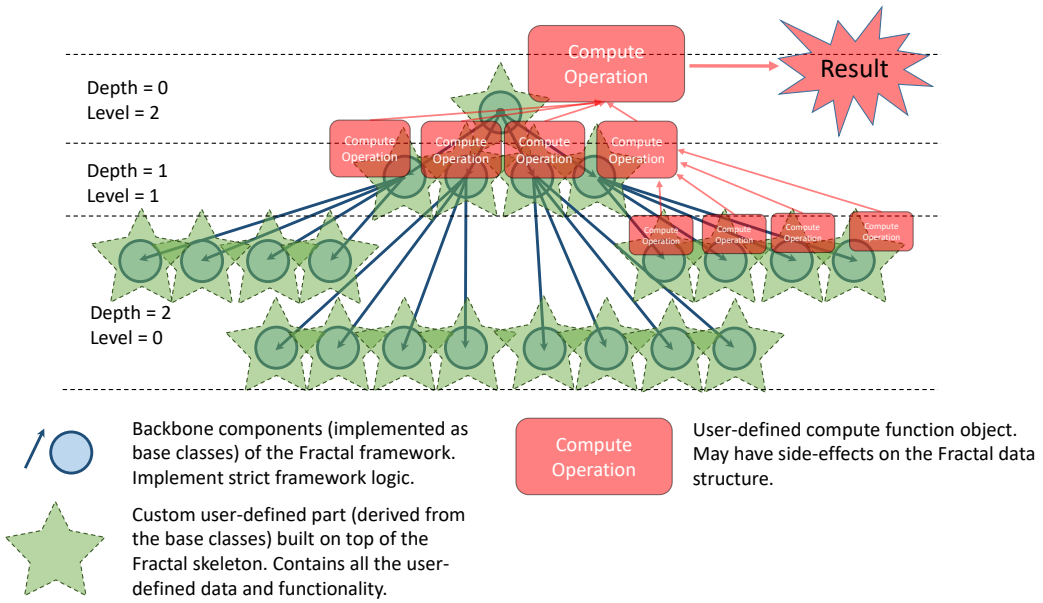

Figure 4.1: Our object-oriented design of the Fractal computational framework.

exact growth and processing procedures, which can touch and alter the data. Given the fixed custom data in the tree nodes, computations operating on the data may still vary. We customize computations as function objects (red rectangles) being passed into higher order functional interfaces of the fractal. In other words, our fractal provides functional style interfaces above its object-oriented structure with the main backbone logic hidden under the hood. The backbone logic can be implemented sequentially or be parallelized in a multiple ways. In all its generality the fractal is a pattern, which can be characterized with self-similarity, repeatedness, structuredness, inherent parallelizability and the exact numeric values such as its depth and arity.

4.3.3 Fold

The **Fold** computational framework has been inspired by the computation done in the *power* benchmark (see Section 2.7.3). The fold is not a new concept and has found a wide application in many functional languages. The C++ language provides *std::accumulate()* function template as a component of its Standard Template Library (STL), which performs a functional fold over a given data structure, but contrary to our computational framework it does not allow any side-effects and modifications to the elements of the data structure must be coded separately. Our C++ *Fold* class template provides an alternative interface to a user with an extensible customization space and clear separation of concerns.

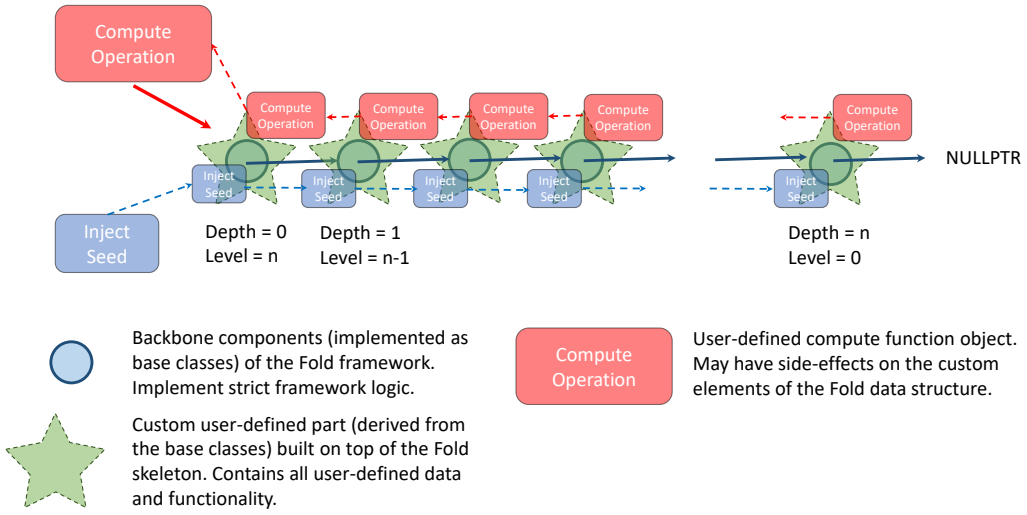

Figure 4.2: An object-oriented design of the Fold computational framework.

Figure 4.2 illustrates the framework. One can think of a Fold as a set of elements arranged into a linked-list. We grow the list to the specified depth given a seed value. Then we may inject some data into the head of the list and propagate the data along with its custom changes to the last element at the tail of the list. All propagation modifications are user-defined. The injection of the data might happen as needed (before every fold iteration for example). Once every element of the list is ready with its injected data planted, the computation starts at the tail element and passes computed values back to previous elements of the chain. The computation may leave some side effects on the elements of the fold, which can accumulate with fold repetitions. The pattern is not parallelizable, but defines a strict order and helps to structure the code to separate various concerns. The object-oriented design of the fold framework is coherent to that of the fractal. Base classes form the skeleton logic and define the interface. All customization happens through overriding inherited base class interface methods withing definitions of derived classes. Custom compute operation is a function object going into functional interfaces of higher order *Fold* class methods.

4.3.4 Reduce

The **Reduce** computational framework is a well-known one. And again, the difference between our computational framework and *std::reduce()* from C++ Standard

Template Library (STL) is the possibility of having side effects and an alternative user customization interface. All general remarks made regarding **Fold** and **Fractal** frameworks also apply to the **Reduce** framework, although the specific details might differ. For example, the reduce framework takes a function object with two overloaded and virtual *operator()* methods. One specifies how to reduce the value from a single element (possibly changing the element in the process) and the other defines the way of combining all the reduced values into the final return value. Our framework implements sequential as well as parallel **Reduce** versions.

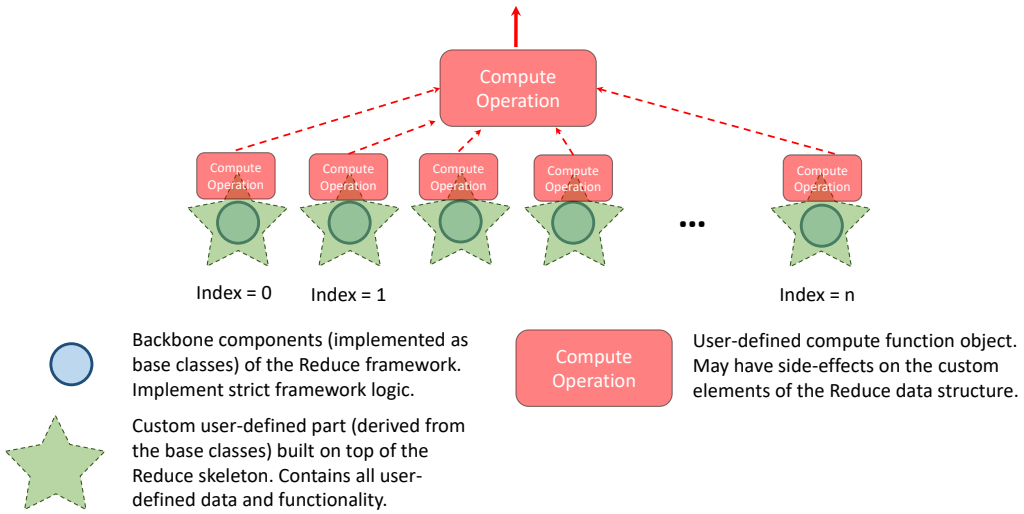

Figure 4.3: An object-oriented design of the Reduce computational framework.

4.3.5 Frameworks library design and implementation

Despite a number of framework specific differences, all our frameworks stick to the same user interface and design. The design and implementation of computational frameworks library have been done along the curved and iterative path running into numerous dead ends and doing redesign efforts over and over before we managed to get the final library version [48]. The major questions raised during the design process of the library were the trade-off between employing static vs. dynamic polymorphism, interoperability of different frameworks (how do we handle the reduction of folds of folds of reductions like in the case of the *power* benchmark, for instance?), as well as the overall library coherence questions (how to design a common interface for patterns

as different as fold and fractal?). But the final design goals have always been clear and follow below.

Modern C++ The implementation of the library is based on the Standard Template Library (STL) data structures, uses move semantics and unique pointers to achieve efficiency and smart memory management. For parallelization library uses OpenMP standard, thus making source code portable. The library is composed of a set of header files with class templates, which are supposed to be included into the user application.

Convenience We designed the library in an object-oriented fashion with a functional look of its user interfaces and the portability in mind.

Coherence Despite variations in the prescribed behaviour, all computational frameworks should stick to the same user interface as well as internal design. The coherence would improve interoperability of different frameworks (say composing a reduction of folds), improve library usability, as well as ease its maintenance and extension. Among less intuitive things, a more general design that handles folds, reductions, fractals, etc. along with all the applications using them altogether will be of the higher quality overall.

Sound design We strove to use the best software design practices. The user side of the library has been inspired by the LLVM Pass Framework [1] and alike it the library uses the *Curiously Recurring Template Pattern (CRTP)* to decrease the number of template parameters and avoid some of the dynamic polymorphism overheads. The concept of computational frameworks fits exactly into the well-known *algorithm template* pattern and the functional user side is implemented with function objects and higher order template methods following the *command* pattern.

We used 4 Olden benchmarks as an inspiration and guide. Benchmarks *health*, *perimeter* and *treeadd* defined our **Fractal** computational framework. The *power* benchmark shaped **Fold** and **Reduce** frameworks. Moreover, striving for coherence and common interface, the designs of different frameworks had a profound effect on each other.

Listing 4.5 shows the essential parts of the final framework design. Initially, custom framework element growth methods such as $Element :: grow()$ and $Element :: growth\_stop\_condition()$ were specified as separate $Framework <>$ template parameters, but were moved to become virtual functions of the base $Element$ class as a trade-off between static vs. dynamic polymorphism. The template method $Framework <>::$

$compute < ComputeType > ()$ is a higher order functional interface, which takes a function object of framework specific type $Framework <>:: ComputeFunction <>$ and applies it to the framework along with computing the main result of the *ComputeType* type. Listing 4.5 shows the one for **Reduction** class. A user is supposed to override two overloaded operators(), which specify a reduction from a single element complemented with the final combining operation. $Framework <>:: grow()$ method defines the backbone data structure and representation of the framework and makes calls to custom user defined functions controlling the growth process.

```
template <typename ElemType, typename SeedType>
class Framework {
    public:
        class Element {
            // user-exposed customization iface
            virtual void grow(SeedType) = 0;
            virtual bool growth_stop_condition() { return false; }
        };
        template <typename ComputeType>
        class ComputeFunction {
            public:
            // framework specific application function API
            virtual ComputeType operator()(ElemType& elem) = 0;
            virtual ComputeType operator()(const std::vector<
    ComputeType>&) = 0;
        };
        void grow(size_t size, SeedType seed) {
            // organise framework elements
            // into a data structure
            ... = new ElemType();
        }
        template<typename ComputeType>
        ComputeType compute(ComputeFunction<ComputeType>& apply_func
    );

    private:
        // framework data structure organisation
        // (list, tree, array, etc.)
};
```

Listing 4.5: Computational framework class template skeleton

4.3.6 Implications of side effects

Compared to pure functional patterns our computational frameworks give a programmer more freedom by allowing to have side effects, but this comes at the price of programmers having more responsibility. It is an interesting trade-off by itself, but a programmer still needs to understand computational patterns our frameworks support and a high level parallelization they do. Quite naturally, our computational frameworks are not universal and are limited to only those programs they are applicable to.

4.4 Performance study of the library

To assess the potential and utility of the concept of computational frameworks, we implemented a prototype library and conducted its thorough performance examination on the subset of Olden benchmarks. We used 4 benchmarks we are interested in. Benchmarks *health*, *treeadd* and *perimeter* fit into our **Fractal** framework and stress it from different angles. We used the *power* benchmark to test the practical operation of our **Fold** and **Reduce** frameworks. We expressed computations in these benchmarks through applicable computational frameworks and have rewritten the original sequential legacy C implementations of the benchmarks with our prototype library in a rejuvenated, structured, well-designed and crucially parallel way. On the opposite side, these benchmarks inspired the concept of computational frameworks and shaped the design of the library. *All our benchmarks have been compiled with GCC 10.1 and -O3 optimization sets. Benchmark running times have been measured with the help of UNIX* ***time()*** *on a powerful compute cluster running Ubuntu 20.04 (focal) and having 64 AMD EPYC 7302 3303 MHz 16-core processors with a total of 0.5 Tb of RAM.*

Performance plots below (Figures 4.5, 4.6, 4.7, 4.8) show the strengths of the library, as well as its equally important weaknesses. The latter characterize applicability and limitations of the library, rather than the problems of the idea. Although *perimeter* benchmark still demonstrates the potential it also highlights the prototype research nature of the library and shows where it needs an optimization effort. The latter presents a matter of software engineering and not that of a research.

The implementation can be easily configured to be sequential or parallel, thus making it easy to conduct measurement experiments. Moreover, the Fractal framework has 2 implementations under the hood. In the most general case the Fractal is based on N-ary unbalanced tree. In this case we allocate it dynamically and every node has an

array of pointers to its children. But in the case of a perfect (all leaves are at the same depth and every non-leaf node has all children present) tree, we can allocate all nodes linearly in an array-based heap manner. In the latter case parallelization happens per level, spawning the number of threads equal to the number of nodes at the given depth (provided that enough CPU cores are available). In the general case though we do not know where the growth of the fractal is going to stop and cannot index nodes on the array. Parallelization in the case of unbalanced fractal happens per child: spawning the number of threads equal to the number of children (again, provided that enough CPU cores are available). Figure 4.4 illustrates two implementation strategies and performance plots below compare them.

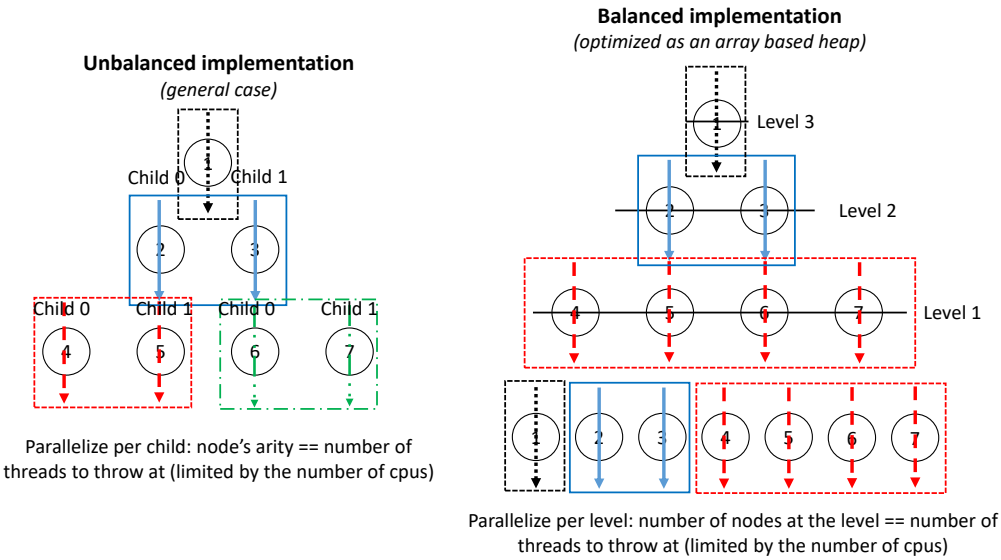

Figure 4.4: *Balanced* vs. *Unbalanced* implementations of the fractal framework. Unbalanced implementation handles arbitrary trees and thus is based on pointers. Balanced implementation optimizes the case of a perfect tree and can be implemented with an array.

Figure 4.5 demonstrates performance of various versions of the *health* benchmark. The benchmark was described in Section 2.7.3. The benchmark grows a tree of hospitals and performs a simulation of the Columbian healthcare system step by step. The number of simulation steps done is plotted on the horizontal axis. The time a simulation took is plotted on a vertical axis. Nodes of the tree, which represent hospitals grow various lists of patients. As simulation goes the lists grow and the nodes become heavier. In other words, the state of the benchmark becomes bigger. The latter has

an accumulating effect and contributes to the workload. We can see this phenomenon reflected in an exponential time growth of the sequential version. Parallel versions diminish this time by tackling the task with several threads. This results into 5-6x speedups of the parallel versions relative to sequential ones. This benchmark is an ideal one to tackle with our fractal framework. And indeed, as Figure 4.5 shows, the thick lines representing the real time (wall clock time) indicate that a parallel versions consistently outperform sequential versions. The sequential versions of our library perform roughly as well as the original legacy C implementation (thin lines vs. a thick *original* line). The latter shows that our library does not introduce any overheads on a benchmark with a significant workload. Dotted lines show the CPU time and illustrate how much of the CPU time has been used by several computational threads of the application. This measure can be used to judge on the aggressiveness of parallelization.

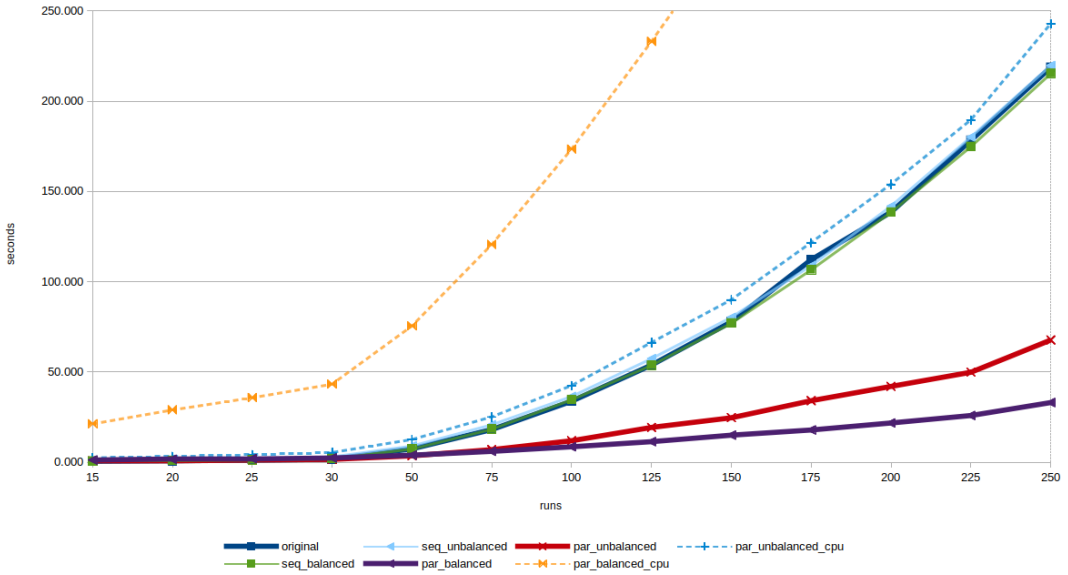

Figure 4.5: Library performance on the *health* benchmark. The legacy C implementation is represented by a thick *original* line. Implementations of the benchmark based on our library are represented by 2 thick (*parallel balanced* and *unbalanced*) and 2 thin (*sequential balanced* and *unbalanced*) lines. Dotted lines represent the CPU time, which is not the real time and has a secondary importance. CPU time reflects the computational workload as we throw several threads at a task and can be used to judge on parallelization aggressiveness.

Figure 4.6 illustrates the results for the *power* benchmark. The benchmark performs a workload of scientific computations. The pattern is basically a reduction of folds of folds of reductions. So, the benchmark tests our **Fold** and **Reduce** frameworks. We

can vary the width of reductions as well as the depth of folds to get various amounts of workload. The benchmark runs repeatedly until the computation result falls withing the set epsilon error. Figure 4.6 illustrates how the running time of the benchmark scales with an increasing top level reduction width. The latter increases the workload and one can see the growing running time of the sequential original legacy C version. The sequential version based on our library runs roughly as well as the original one, while the parallel version consistently outperforms both sequential versions. Dotted line again represents the CPU time and not the real time. CPU time approximately reflects how many threads were used to run the benchmark. As a validation, one can see that CPU time divided by parallel wall clock time roughly corresponds to the reduction width. Behaviour of the *power* benchmark does not change significantly, when we vary reduction widths or fold depths. We do not include the other experiments we ran, as they do not change the picture.

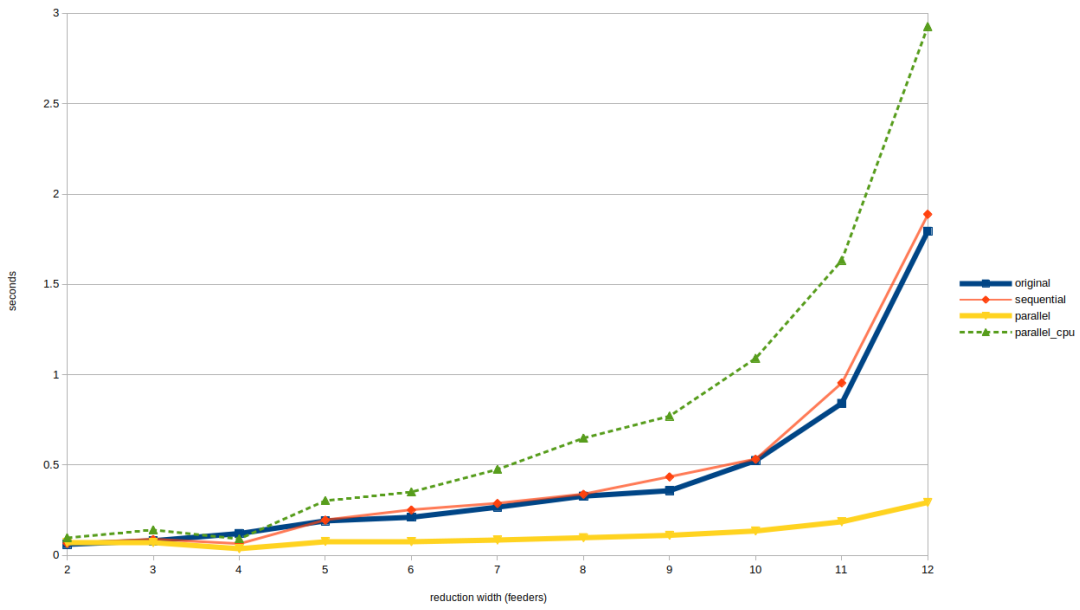

Figure 4.6: Library performance on the *power* benchmark. We vary the top level (feeders) reduction width to scale the workload.

The *health* and the *power* benchmarks show the strengths of our library. The *treeadd* and *perimeter* benchmarks highlight its weaknesses. The weaknesses stress the research prototype nature of the library and show places for further optimization effort. Figure 4.7 shows the behaviour of the *treeadd* benchmark. This benchmark grows a tree with values at its nodes. Then it runs iteratively computing the reduction over the tree and updating node values. The computation is very lighweight. Basically it is just

one addition per node processing. The state is minimal and does not accumulate as we run the benchmark over and over. One can see a roughly linear running time of the original legacy C version. Overheads of our sequential library versions are striking for such a small benchmark. Although the asymptotic complexity of all shown versions is the same, various bookkeeping overheads of the library diminish performance relative to the original version with just a single addition per node as the workload. These overheads can be optimized away with some engineering effort, but present in the research prototype library. The more threads we throw the bigger our overheads become. One can see it looking at the running time of balanced and unbalanced versions. It is worthless to tackle such a small benchmark with our library.

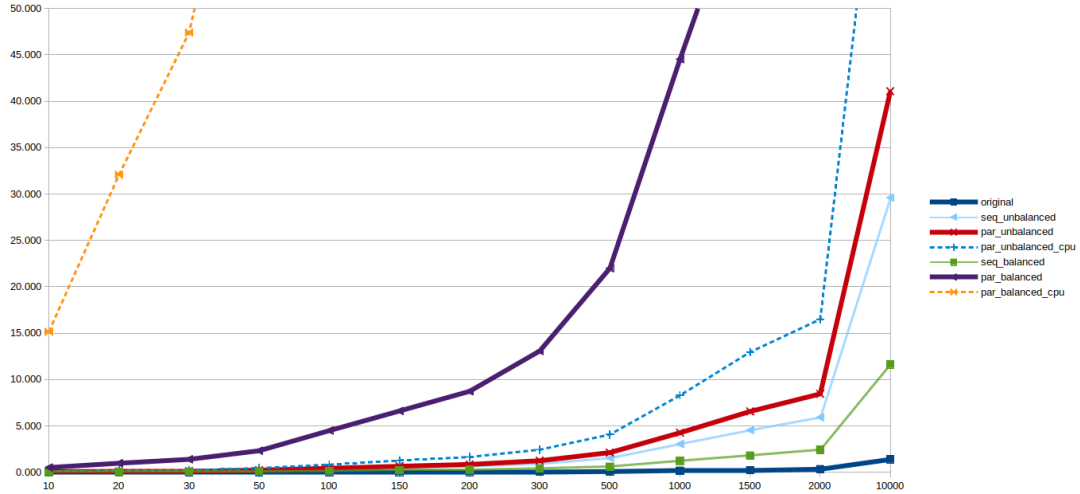

Figure 4.7: Library performance on the *treeadd* benchmark.

The *perimeter* benchmark shows a different picture, but still highlights the current weaknesses of the library implementation. Figure 4.8 shows the bar chart. The benchmark does not run an iterative simulation as the 3 other benchmarks do - it is a single perimeter computation. Horizontal axis plots the depth of the tree underlying perimeter computation, a workload size in other words. As usual, the vertical axis plots the time it took to run the computation. As the benchmark is based on an unbalanced tree version we cannot run it with a balanced implementation. Here we have only the original sequential legacy C implementation, the sequential implementation based on our library and its parallel counterpart. One can see that the benchmark also highlights the current implementation problems: original does better than sequential. At the same time the benchmark has a potential: there is a good parallel speedup when we compare the sequential versus parallel library based versions.

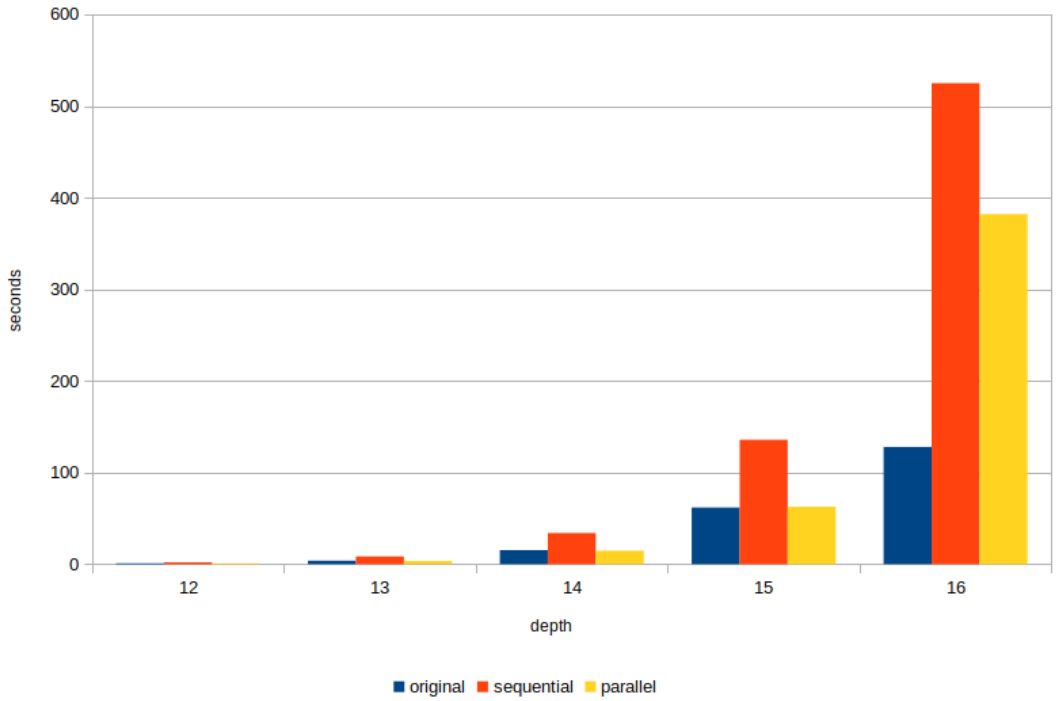

Figure 4.8: Library performance on the *perimeter* benchmark. Horizontal axis plots the depth of the perimeter tree (see Section 2.7.3). Vertical axis shows the time it takes to compute a single perimeter value with various tree depths (splitting granularity).

4.5 Future work

The concept of computational frameworks and the prototype library form the basis for the future work. We believe that the process of software parallelization with the help of computational frameworks can be further automated for a relatively simple applications (like Olden benchmarks). Figure 4.9 shows the scheme. It takes an original legacy C source code, where we supposedly have a computation that fits into one or several of our frameworks. The task of the recognizer is to identify a computational pattern. Like in the case of the *health* Olden benchmark we have a recursive *grow()* and *compute()* procedures, which build and process the tree correspondingly. That recursive pattern can be matched to a fractal framework. The next step would be to strip the code corresponding and implementing the pattern (*backbone logic*) and leave the rest as the *business logic*. The business logic must be further classified into various computational framework template class methods (like *grow()*, *growth_stop_condition()*, *inject()*, etc.). Then, the class template must be instantiated with the business logic inserted into the right places. Once that is done, the parallelization is done, as the compiler will take care of everything else.

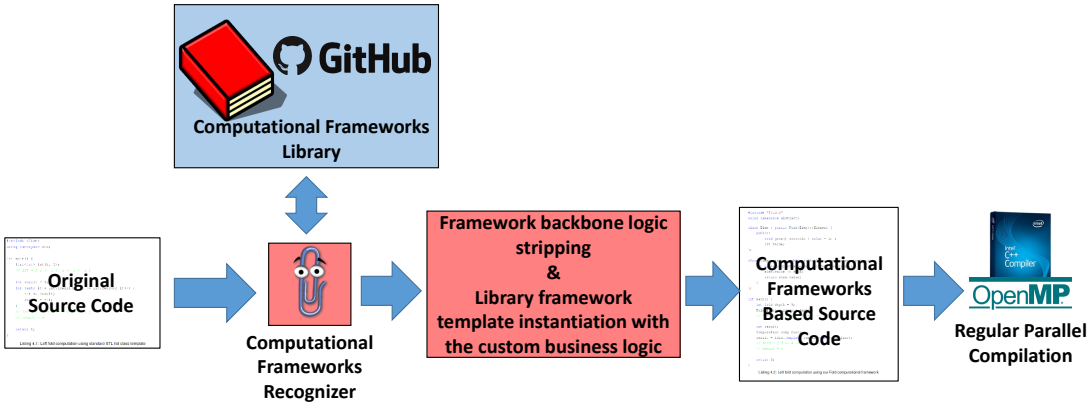

Figure 4.9: An alternative software parallelization scheme based on our library of computational frameworks.

Technically, for benchmarks as simple as Olden the technique can work on the level of compiler's front end: source code or an abstract syntax tree (AST). That sets the technique apart from the regular parallelization approaches based on dependence graphs and working on the level of compiler's intermediate representation (IR).

Chapter 5

Summary & Conclusions

Parallelism has become pervasive in the modern computing world and the task of *software parallelization* is extremely important. Despite decades of academic research and industrial investment into the area, the human expert still has a major role to play. Our state-of-the-art literature review shows that there are still no automatic solutions that could fully replace an expert programmer and at the same time achieve performance results comparable to those of manually parallelized software.

In order to achieve the best result a programmer has to work on multiple conceptual levels starting from problem decomposition and algorithm choice down to low level loop transformations. In this thesis we fully acknowledge the role of the human expert, but provide the latter with an *assistant solution*, which aims at alleviating the software parallelization process and is as multifaceted as the problem itself. The solution consists of a tool with a parallelization methodology to employ it and a library implementing novel parallel software design primitives, namely computational frameworks, which aim at higher levels of program parallelization.

The tool aims to assist human experts by guiding them directly towards the most interesting loops from the perspective of software parallelization, thus alleviating the parallelizable code search process and delivering savings for this costly human resource. We have developed a novel machine learning based approach to predicting whether or not a loop is parallelizable. We combine this prediction with traditional profiling information and develop a ranking function that prioritizes low-risk, high-gain loop candidates, which are finally presented to the user.

We have evaluated our parallelization assistant against the sequential C implementations of the SNU NPB suite. We show that our assistant recognizes parallelizable loops more aggressively than conservative parallelizing compilers, thus improving par-

allelism discovery. We also show that our parallelization assistant can increase programmer productivity. Our experiments confirm, that equipped with our assistant, a programmer is required to examine and parallelize substantially fewer loops to achieve performance levels comparable to those of the reference OpenMP implementations of the benchmarks.

Our work has demonstrated that there is scope for machine learning based tool support in parallelization despite its inherent lack of safety. By assisting human programmers rather than replacing them, machine learning techniques have the potential to deliver productivity gains beyond what is possible by relying on traditional parallelization approaches alone.

The second component of the assistant solution is the concept of *computational frameworks* along with a research prototype library implementing it. In our benchmark feasibility studies we show that the problem of successful data structure choice stands particularly important and can vastly affect the parallelizability of programs. Moreover, we observe *the problem of algorithm and data structure inseparability*. These issues has led us to a novel concept of computational frameworks, which are higher level entities that embody algorithms and data structures together into an elegant and well-structured construct. Computational frameworks can be used as parallel software design and construction primitives alleviating the task and ultimately parallelizing a wide class of applications, which fit into their computational patterns.

We shaped the concept and designed the library using a subset of the Olden benchmarks. We expressed benchmark computations through our *Fractal*, *Fold* and *Reduce* computational frameworks and rewrote the benchmarks in a modern, well-structured and crucially parallel way combining the elements of both object-oriented and functional programming. Moreover, the rejuvenated C++ benchmark versions demonstrate a good parallel performance compared to their serial legacy C counterparts. On the major benchmarks we achieve 5-6x speedups. However, given the research prototype nature of the library, some further engineering effort is still needed.

In this thesis we demonstrate that when decades old and well-known methods of software parallelization such as various automatic techniques run into their limits and fail to tackle the challenges of the real world codes, and more exotic methods of machine learning based techniques run into their principal problems of inherent statistical errors and the lack of safety, it is possible to acknowledge the role of a human expert and resort to various assisting solutions. The latter demonstrate promising results and pave an attractive future research direction.

Bibliography

[1] LLVM Developer Group. *The LLVM Compiler Infrastructure*. 2002-2019. URL: https://llvm.org/.

[2] Aleksandr Maramzin. *Pervasive Parallelism Tool*. 2018-2019. URL: https://github.com/av-maramzin/icc-opt-report-compiler.

[3] Intel Corporation. *Intel C/C++ Compiler (ICC)*. 1985-2018. URL: https://software.intel.com/en-us/c-compilers.

[4] Jeanne Ferrante, Karl J. Ottenstein, and Joe D. Warren. "The Program Dependence Graph and Its Use in Optimization". In: *ACM Trans. Program. Lang. Syst.* 9.3 (July 1987), pp. 319–349. ISSN: 0164-0925. DOI: 10.1145/24039.24041. URL: http://doi.acm.org/10.1145/24039.24041.

[5] David F. Bacon, Susan L. Graham, and Oliver J. Sharp. "Compiler Transformations for High-performance Computing". In: *ACM Comput. Surv.* 26.4 (Dec. 1994), pp. 345–420. ISSN: 0360-0300. DOI: 10.1145/197405.197406. URL: http://doi.acm.org/10.1145/197405.197406.

[6] René Dekker and Frans Ververs. "Abstract Data Structure Recognition". In: *Proceedings of the 9th International Conference on Knowledge-Based Software Engineering*. KBSE'94. Monterey, CA, USA: IEEE Press, 1994, pp. 133–140. ISBN: 0-8186-6380-4. DOI: 10.1109/KBSE.1994.342669. URL: https://doi.org/10.1109/KBSE.1994.342669.

[7] Maryam Emami, Rakesh Ghiya, and Laurie J. Hendren. "Context-sensitive Interprocedural Points-to Analysis in the Presence of Function Pointers". In: *Proceedings of the ACM SIGPLAN 1994 Conference on Programming Language Design and Implementation*. PLDI '94. Orlando, Florida, USA: ACM, 1994, pp. 242–256. ISBN: 0-89791-662-X. DOI: 10.1145/178243.178264. URL: http://doi.acm.org/10.1145/178243.178264.

[8] Erich Gamma et al. *Design Patterns: Elements of Reusable Object-Oriented Software*. USA: Addison-Wesley Longman Publishing Co., Inc., 1995. ISBN: 0201633612.

[9] Rakesh Ghiya and Laurie J. Hendren. "Is It a Tree, a DAG, or a Cyclic Graph? A Shape Analysis for Heap-directed Pointers in C". In: *Proceedings of the 23rd ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. POPL '96. St. Petersburg Beach, Florida, USA: ACM, 1996, pp. 1–15. ISBN: 0-89791-769-3. DOI: 10.1145/237721.237724. URL: http://doi.acm.org/10.1145/237721.237724.

[10] Steven S. Muchnick. *Advanced Compiler Design and Implementation*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. ISBN: 1-55860-320-4.

[11] Leonardo Dagum and Ramesh Menon. "OpenMP: An Industry-Standard API for Shared-Memory Programming". In: *IEEE Comput. Sci. Eng.* 5.1 (Jan. 1998), pp. 46–55. ISSN: 1070-9924. DOI: 10.1109/99.660313. URL: https://doi.org/10.1109/99.660313.

[12] Mooly Sagiv, Thomas Reps, and Reinhard Wilhelm. "Parametric Shape Analysis via 3-valued Logic". In: *Proceedings of the 26th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*. POPL '99. San Antonio, Texas, USA: ACM, 1999, pp. 105–118. ISBN: 1-58113-095-3. DOI: 10.1145/292540.292552. URL: http://doi.acm.org/10.1145/292540.292552.

[13] Reinhard Wilhelm, Shmuel Sagiv, and Thomas W. Reps. "Shape Analysis". In: *Proceedings of the 9th International Conference on Compiler Construction*. CC '00. London, UK, UK: Springer-Verlag, 2000, pp. 1–17. ISBN: 3-540-67263-X. URL: http://dl.acm.org/citation.cfm?id=647476.760384.

[14] Ken Kennedy and John R. Allen. *Optimizing Compilers for Modern Architectures: A Dependence-based Approach*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002. ISBN: 1-55860-286-0.

[15] L. Almagor et al. "Finding Effective Compilation Sequences". In: *Proceedings of the 2004 ACM SIGPLAN/SIGBED Conference on Languages, Compilers, and Tools for Embedded Systems*. LCTES '04. Washington, DC, USA: ACM, 2004, pp. 231–239. ISBN: 1-58113-806-7. DOI: 10.1145/997163.997196. URL: http://doi.acm.org/10.1145/997163.997196.

[16] Chris Lattner and Vikram Adve. “LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation”. In: *Proceedings of the International Symposium on Code Generation and Optimization: Feedback-directed and Runtime Optimization*. CGO ’04. Palo Alto, California: IEEE Computer Society, 2004, pp. 75–. ISBN: 0-7695-2102-9. URL: http://dl.acm.org/citation.cfm?id=977395.977673.

[17] Peng Zhao and José Nelson Amaral. “To Inline or Not to Inline? Enhanced Inlining Decisions”. In: *Languages and Compilers for Parallel Computing*. Ed. by Lawrence Rauchwerger. Berlin, Heidelberg: Springer, 2004, pp. 405–419. ISBN: 978-3-540-24644-2. DOI: 10.1007/978-3-540-24644-2_26.

[18] J. Cavazos and M. F. P. O’Boyle. “Automatic Tuning of Inlining Heuristics”. In: *SC ’05: Proceedings of the 2005 ACM/IEEE Conference on Supercomputing*. Washington, DC, USA: IEEE Computer Society, Nov. 2005, pp. 14–14. DOI: 10.1109/SC.2005.14.

[19] Keith D. Cooper et al. “ACME: Adaptive Compilation Made Efficient”. In: *Proceedings of the 2005 ACM SIGPLAN/SIGBED Conference on Languages, Compilers, and Tools for Embedded Systems*. LCTES ’05. Chicago, Illinois, USA: ACM, 2005, pp. 69–77. ISBN: 1-59593-018-3. DOI: 10.1145/1065910.1065921. URL: http://doi.acm.org/10.1145/1065910.1065921.

[20] Alexandru Niculescu-Mizil and Rich Caruana. “Predicting Good Probabilities with Supervised Learning”. In: *Proceedings of the 22nd International Conference on Machine Learning*. ICML ’05. Bonn, Germany: ACM, 2005, pp. 625–632. ISBN: 1-59593-180-5. DOI: 10.1145/1102351.1102430. URL: http://doi.acm.org/10.1145/1102351.1102430.

[21] M. Stephenson and S. Amarasinghe. “Predicting unroll factors using supervised classification”. In: *International Symposium on Code Generation and Optimization*. Washington, DC, USA: IEEE Computer Society, Mar. 2005, pp. 123–134. DOI: 10.1109/CGO.2005.29.

[22] Changhee Jung and Nathan Clark. “DDT: design and evaluation of a dynamic program analysis for optimizing data structure usage”. In: *MICRO 42: Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*. New York, New York: ACM, 2009, pp. 56–66. ISBN: 978-1-60558-798-1. DOI: http://doi.acm.org/10.1145/1669112.1669122.

[23] H. Leather, E. Bonilla, and M. O'Boyle. "Automatic Feature Generation for Machine Learning Based Optimizing Compilation". In: *2009 International Symposium on Code Generation and Optimization*. Washington, DC, USA: IEEE Computer Society, Mar. 2009, pp. 81–91. DOI: 10.1109/CGO.2009.21.

[24] Naeem Abbas et al. "Accelerating HMMER on FPGA using parallel prefixes and reductions". In: Aug. 2010, pp. 37–44. DOI: 10.1109/FPT.2010.5681755.

[25] Narayan Ganesan et al. "Accelerating HMMER on GPUs by Implementing Hybrid Data and Task Parallelism". In: *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*. BCB '10. Niagara Falls, New York: ACM, 2010, pp. 418–421. ISBN: 978-1-4503-0438-2. DOI: 10.1145/1854776.1854844. URL: http://doi.acm.org/10.1145/1854776.1854844.

[26] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[27] Sangmin Seo, Gangwon Jo, and Jaejin Lee. "Performance Characterization of the NAS Parallel Benchmarks in OpenCL". In: *Proceedings of the 2011 IEEE International Symposium on Workload Characterization*. IISWC '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 137–148. ISBN: 978-1-4577-2063-5. DOI: 10.1109/IISWC.2011.6114174. URL: http://dx.doi.org/10.1109/IISWC.2011.6114174.

[28] NASA Advanced Supercomputing (NAS) Division. *NAS Parallel Benchmarks*. Aug. 2012. URL: https://www.nas.nasa.gov/publications/npb.html.

[29] Kento Emoto and Hiroto Imachi. "Parallel Tree Reduction on MapReduce". In: *Procedia Computer Science* 9 (Dec. 2012). DOI: 10.1016/j.procs.2012.04.201.

[30] Per Larsen et al. "Parallelizing More Loops with Compiler Guided Refactoring". In: *Proceedings of the 2012 41st International Conference on Parallel Processing*. ICPP '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 410–419. ISBN: 978-0-7695-4796-1. DOI: 10.1109/ICPP.2012.48. URL: http://dx.doi.org/10.1109/ICPP.2012.48.

[31] Michael McCool, James Reinders, and Arch Robison. *Structured Parallel Programming: Patterns for Efficient Computation*. 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2012. ISBN: 9780123914439.

[32] S. Midkiff. *Automatic Parallelization: An Overview of Fundamental Compiler Techniques*. San Rafael, CA, USA: Morgan & Claypool, 2012. ISBN: 9781608458424. URL: https://ieeexplore.ieee.org/document/6813266.

[33] Seoul National University. *SNU NAS Parallel Benchmarks*. Aug. 2012. URL: http://aces.snu.ac.kr/software/snu-npb/.

[34] Daniel Fried et al. "Predicting Parallelization of Sequential Programs Using Supervised Learning". In: *Proc. of the 12th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA*. Miami, FL, USA: IEEE Computer Society, Dec. 2013, pp. 72–77. DOI: 10.1109/ICMLA.2013.108. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6786084.

[35] Gareth James et al. *An introduction to statistical learning : with applications in R*. Heidelberg, Germany: Springer, 2013. ISBN: 978-1-4614-7138-7. DOI: 10.1007/978-1-4614-7138-7.

[36] Akihiro Hayashi et al. "Machine-Learning-based Performance Heuristics for Runtime CPU/GPU Selection". In: *Proceedings of the Principles and Practices of Programming on The Java Platform*. PPPJ '15. Melbourne, FL, USA: ACM, 2015, pp. 27–36. ISBN: 978-1-4503-3712-0. DOI: 10.1145/2807426.2807429. URL: http://doi.acm.org/10.1145/2807426.2807429.

[37] Istvan Haller, Asia Slowinska, and Herbert Bos. "Scalable Data Structure Detection and Classification for C/C++ Binaries". In: *Empirical Softw. Engg.* 21.3 (June 2016), pp. 778–810. ISSN: 1382-3256. DOI: 10.1007/s10664-015-9363-y. URL: http://dx.doi.org/10.1007/s10664-015-9363-y.

[38] Amir H. Ashouri et al. "MiCOMP: Mitigating the Compiler Phase-Ordering Problem Using Optimization Sub-Sequences and Machine Learning". In: *ACM Trans. Archit. Code Optim.* 14.3 (Sept. 2017), 29:1–29:28. ISSN: 1544-3566. DOI: 10.1145/3124452. URL: http://doi.acm.org/10.1145/3124452.

[39] Philip Ginsbach and Michael F. P. O'Boyle. "Discovery and Exploitation of General Reductions: A Constraint Based Approach". In: *Proceedings of the 2017 International Symposium on Code Generation and Optimization*. CGO '17. Austin, USA: IEEE Press, 2017, pp. 269–280. ISBN: 978-1-5090-4931-8. URL: http://dl.acm.org/citation.cfm?id=3049832.3049862.

[40] Nicklas Bo Jensen and Sven Karlsson. “Improving Loop Dependence Analysis”. In: *ACM Trans. Archit. Code Optim.* 14.3 (Aug. 2017), 22:1–22:24. ISSN: 1544-3566. DOI: 10.1145/3095754. URL: http://doi.acm.org/10.1145/3095754.

[41] Thomas Rupprecht et al. “DSIbin: Identifying Dynamic Data Structures in C/C++ Binaries”. In: *Proceedings of the 32Nd IEEE/ACM International Conference on Automated Software Engineering*. ASE 2017. Urbana-Champaign, IL, USA: IEEE Press, 2017, pp. 331–341. ISBN: 978-1-5386-2684-9. URL: http://dl.acm.org/citation.cfm?id=3155562.3155607.

[42] Philip Ginsbach, Lewis Crawford, and Michael F. P. O'Boyle. “CAnDL: A Domain Specific Language for Compiler Analysis”. In: *Proceedings of the 27th International Conference on Compiler Construction*. CC 2018. Vienna, Austria: ACM, 2018, pp. 151–162. ISBN: 978-1-4503-5644-2. DOI: 10.1145/3178372.3179515. URL: http://doi.acm.org/10.1145/3178372.3179515.

[43] Philip Ginsbach et al. “Automatic Matching of Legacy Code to Heterogeneous APIs: An Idiomatic Approach”. In: *SIGPLAN Not.* 53.2 (Mar. 2018), pp. 139–153. ISSN: 0362-1340. DOI: 10.1145/3296957.3173182. URL: http://doi.acm.org/10.1145/3296957.3173182.

[44] Stanislav Manilov, Christos Vasiladiotis, and Björn Franke. “Generalized Profile-guided Iterator Recognition”. In: *Proceedings of the 27th International Conference on Compiler Construction*. CC 2018. Vienna, Austria: ACM, 2018, pp. 185–195. ISBN: 978-1-4503-5644-2. DOI: 10.1145/3178372.3179511. URL: http://doi.acm.org/10.1145/3178372.3179511.

[45] Zheng Wang and Michael O’Boyle. “Machine Learning in Compiler Optimization”. English. In: *Proceedings of the IEEE* 106.11 (May 2018), pp. 1879–1901. ISSN: 0018-9219. DOI: 10.1109/JPROC.2018.2817118.

[46] Aleksandr Maramzin. *Machine Learning Based Parallelization Assistant*. The University of Edinburgh, 2019. URL: https://github.com/av-maramzin/PParMetrics.

[47] Aleksandr Maramzin et al. ““It Looks like You’Re Writing a Parallel Loop”: A Machine Learning Based Parallelization Assistant”. In: *Proceedings of the 6th ACM SIGPLAN International Workshop on AI-Inspired and Empirical Methods for Software Engineering on Parallel Computing Systems*. AI-SEPS 2019.

Athens, Greece: Association for Computing Machinery, 2019, pp. 1–10. ISBN: 9781450369831. DOI: 10.1145/3358500.3361567. URL: https://doi.org/10.1145/3358500.3361567.

[48] Aleksandr Maramzin. *Computational Frameworks*. The University of Edinburgh, 2020. URL: https://github.com/av-maramzin/DCP/abstract.

[49] Lawrence Livermore National Laboratory. *Parallel Computing Tutorials*. URL: https://computing.llnl.gov/.