Automatic Parallelization of Recursive Functions with Rewriting Rules

Rodrigo C. O. Rocha
University of Edinburgh, UK
r.rocha@ed.ac.uk

Luís F. W. Góes
PUC Minas, Brazil
lfwgoes@pucminas.br

Fernando M. Q. Pereira
UFMG, Brazil
fernando@dcc.ufmg.br

Abstract

Functional programming languages have, since their early days, being thought as the holy grail of parallelism. And, in fact, the absence of race conditions, coupled with algorithmic skeletons such as map and reduce, have given developers the opportunity to write many different techniques aimed at the automatic parallelization of programs. However, there are many functional programs that are still difficult to parallelize. This difficulty stems from many factors, including the complex syntax of recursive functions. This paper provides new equipment to deal with this problem. Such instrument consists on an observation, plus a code transformation that is enabled by said observation. Concerning the first contribution, we demonstrate that many recursive functions can be rewritten as a combination of associative operations. We group such functions into two categories, which involve monoid and semiring operations. Each of these categories admits a parallel implementation. To demonstrate the effectiveness of this idea, we have implemented an automatic code rewriting tool for Haskell, and have used it to convert six well-known recursive functions to algorithms that run in parallel. Our tool is totally automatic, and it is able to deliver non-trivial speedups onto the sequential version of the programs that it receives. As an example, we have observed speedups of almost four times in benchmarks such as the Catalan Numbers and the Horner's Method, among others.

Keywords: recursive functions, parallel computing, functional programming, algebraic framework, rewriting rules

1. Introduction

The advent of multi-core computers has greatly spread the use of parallel programming among application developers. Yet, writing code that runs in parallel is still a difficult and error-prone task. Thus, the automatic parallelization of code has surfaced as an effective alternative to the development of high-performant programs [1, 2, 3]. In this sense, functional programming languages appear as a promising alternative to the development of parallel code. They provide referential transparency, reducing shared data and eliminating side effects, which makes automatic parallelization much easier. However, in spite of years of research, automatic generation of parallelism, out of functional code, is not a solved problem [4, 5]. Testimony of this last statement is the fact that functional code is still manually parallelized, usually by means of parallel skeletons [5, 6, 7].

The main challenge faced by automatic parallelization tools in functional languages is the fact that parallelism is often hidden under the syntax of complex recursive functions. There are several techniques to discover parallelism, such as work targeting list homomorphisms [8, 9, 10, 11], or the work of Fisher and Ghuloum [12], who parallelize imperative loops that can be translated as composition of functions. Nevertheless, the programming languages community still lacks approaches to infer parallelism on recursive functions automatically. The goal of this paper is to contribute to solve this omission by extending the family of recursive functions that can be parallelized automatically.

To achieve this objective, we propose an algebraic framework for parallelizing two special classes of recursive functions. These functions need to have two core properties. First: the recursive function must contain only operations that can be used to define monoids or semirings. Second: the propagation of arguments between recursive calls has to be defined by an invertible function. The proposed framework is based on the theory introduced by Fisher and Ghuloum [12]. Those authors have designed and tested an approach to parallelize imperative loops by transforming them in recurrence relations defined by the compositions of associative functions. Yet, we go beyond our predecessors in two ways: (i) we work on recursive functions, instead of imperative loops; (ii) we provide a more general definition of parallel function composition. The key idea behind our findings is the fact that algebraic structures such as groups, monoids and semirings let us decompose recursive functions into simpler components, which are amenable to automatic parallelization.

To validate the ideas discussed in this paper, we have used them to implement a source-to-source compiler, by means of rewriting rules, that performs automatic parallelization of Haskell code. In Section 4 we show how to parallelize six different – and well-known – recursive functions. Key to efficiency is the fact that we can transform elements in the family of parallelizable functions into list homomorphisms. This transformation, which we explain in Section 3, to the best of our knowledge, is novel. Our experiments show that our technique is effective and useful. We have achieved speedups of up to 3× in a 4-core Intel i5 processor, comparing the parallelized code against its single-threaded counterpart. The parallelized code obtained impressive performance gains over the original purely recursive implementation of the benchmarks, with a minimum speedup of 1.63×. These results are even more meaningful if we consider that they have been obtained in a completely automatic way.

This paper extends our previous work [13], closing two years of investigation on the automatic parallelization of a particular class of recursive functions. This new report augments our preliminary work in three ways. First, we present a much deeper discussion of our results. At the time of our original publication, we had observed that some functions would show very little speedup when running in parallel. Careful profiling has revealed that the culprit, in this case, is the garbage collector. Second, we now implement the automatic parallelization of semiring based functions. Before, we had to parallelize those functions via manual transformations. Finally, the extra number of pages allowed us to touch more related work; hence, providing the reader with a clearer perspective on where our contribution stays, when compared with previous art.

2. Overview

We will use the well-known factorial function as an example to introduce our ideas to the reader. This function can be defined as follows:

$$f(x) \coloneqq \begin{cases} 1 & \text{if } x = 1 \\ x \cdot f(x-1) & \text{otherwise} \end{cases}$$

Factorial is a very simple function, and the reader familiar with the parallelization of reductions on commutative and associative operators will know immediately that this function has a very efficient parallel implementation. Key to perform this parallelization is the observation that factorial can be re-written as a sequence of multiplications, e.g.:$f(x) = x \cdot (x-1) \cdot \ldots \cdot 2 \cdot 1$.

A key property of multiplication – associativity – lets us solve them in a pairwise fashion. This possibility gives us the chance to run the above expression in $O(\ln x)$ time. The goal of this paper is to be able to apply this kind of parallelization automatically onto recursive functions. In order to achieve this objective, we shall be re-writing functions as a composition of simpler functions which are associative. In the case of factorial, this composition looks like:

$$f(x) = (f'_{x-1} \circ \ldots \circ f'_2 \circ f'_1)(1)$$

How do we find a suitable implementation of f'_i? We first provide this answer for a family of recursive functions which have the following format:

$$f(p) \coloneqq \begin{cases} g_0(p) & \text{if } p = p_0 \\ g_1(p) \oplus f(h(p)) \oplus g_2(p) & \text{otherwise} \end{cases} \tag{1}$$

For any function f that can be written in the format above, we show that it is possible to decompose f into a composition of functions. We first identify h, the function used to propagate the arguments of f, which we call the *hop* function. The *hop* function must be a well-defined monotonic function, which must have an inverse. For the factorial example, we have the following *hop* function h: $h(x) = x - 1$, with inverse $h^{-1}(x) = x + 1$.

The *hop* function is fundamental for generating the next arguments of the sequence of compositions. Because it has an inverse, we use it to know $x-1$, the number of functions which will constitute the sequence of compositions. We find out $x-1$ after solving the following equation: $p_0 = h^{x-1}(p)$. For the factorial example, the depth is exactly the initial argument x. After identifying the *hop* function and its inverse, we re-write f in a manner suitable for the composition of functions which does not contain a recursive call, e.g.: $f'_i(s) \coloneqq (i+1) \cdot s$, where s is the usually called accumulator parameter in funcional composition.

Now, we can write f such as $f(x) = (f'_{x-1} \circ f'_{x-2} \circ \cdots \circ f'_2 \circ f'_1)(1)$. This transformation is useful, considering that functional composition is an associative operation, which can often be parallelized if it is possible to symbolically compute and simplify intermediate compositions [12]. For instance, for $i > 1 \in \mathbb{Z}$, $(f'_i \circ f'_{i-1})(s) = (i+1) \cdot (i \cdot s)$ can be reassociated as $(f'_i \circ f'_{i-1})(s) = ((i+1) \cdot i) \cdot s$ which is essentially equivalent to the original function regarding computational complexity. Thus we can evaluate $f(x)$ by computing a reduction over a list of functions, using the functional composition operator, i.e., $f(x) = \circ/[f'_{x-1}, f'_{x-2}, \ldots, f'_1]$.

Section 3.2 generalizes the factorial example seen in this section. In Section 3.3 we extend our framework a bit further, showing that we can also parallelize functions in the format below. In this case, we consider two operators $\oplus$ and $\odot$, for which we only require that $\odot$ be associative, and $\oplus$ be commutative, in addition of being associative. Again, the hop function h must have an inverse function which can be computed efficiently.

$$f(p) \coloneqq \begin{cases} g_0(p) & \text{if } p = p_0 \\ g_1(p) \oplus (g_3(p) \odot f(h(p)) \odot g_4(p)) \oplus g_2(p) & \text{otherwise} \end{cases} \qquad (2)$$

3. Automatic Parallelization of Recursive Functions

In this section we formalize the developments earlier seen in Section 2. To this end, we provide a few basic notions in Section 3.1. In Section 3.2, we show how to parallelize functions on the format given by Equation 1. In Section 3.3, we move on to deal with functions defined by Equation 2.

3.1. Technical Background

In the rest of this paper we shall use three notions borrowed from abstract algebra: *groups*, *monoids* and *semirings*. If S is a set, then we define these algebraic structures as follows:

- A group $G = (S, +)$ is a nonempty set closed under an associative binary-operation $+$, which is associative, invertible and has a zero element $\mathbf{0}$, the identity regarding $+$. For each $a \in S$, its inverse $-a$ also belongs to S. A group need not be commutative. If a group is commutative, it is usually called an abelian group [14].

- A monoid $M = (S, +)$ is a nonempty set closed under an associative binary-operation $+$ with identity $\mathbf{0}$. A monoid need not be commutative and its elements need not have inverses within the monoid.

- A semiring $R = (S, +, \cdot)$ is a nonempty set closed under two associative binary-operations $+$ and $\cdot$, called addition and multiplication, respectively [15, 16]. A semiring satisfies the following conditions:
  - $(S, +)$ is a commutative monoid with identity element $\mathbf{0}$;
  - $(S, \cdot)$ is a monoid with identity element $\mathbf{1}$;
  - Multiplication distributes over addition from either side;
  - Multiplication by $\mathbf{0}$ annihilates R, i.e. $a \cdot \mathbf{0} = \mathbf{0} \cdot a = \mathbf{0}$, for all $a \in S$.

3.1.1. Parallelizing Functional Composition

Functional composition is an associative binary operator over functions. Previous work [12, 17] has shown that, given a family of indexed functions $\mathcal{F}$ closed under functional composition $\circ$, a function $\psi : \mathbb{Z} \times \mathbb{Z} \to \mathcal{F}$ is a composition evaluator of $\mathcal{F}$ iff $i \psi j = f_i \circ f_j$, for $f_i, f_j \in \mathcal{F}$. If each function in $\mathcal{F}$ and the composition evaluator ψ are constant-time computations, then a sequence of n compositions can be efficiently computed in $O(n/p + \ln p)$, considering p processing units. A functional composition over $\mathcal{F}$ can be evaluated by using the composition evaluator ψ. Thus, given a sequence of compositions $f_n \circ f_{n-1} \circ \cdots \circ f_1$, this sequence can be efficiently computed using a reduction operator $\circ/[f_n, f_{n-1}, \ldots, f_1]$, since the following equivalence holds: $\circ/[f_n, f_{n-1}, \ldots, f_1] = \psi/[n, n-1, \ldots, 1]$.

3.2. Monoids

In this section, we generalize the solution presented in Section 2. We provide the formal description of the mechanism used for parallelizing the factorial recursive function, regarding general algebraic structures of groups and monoids. Let S_G and S_M be sets and $M = (S_M, +)$ a monoid. Let $f : S_G \to S_M$ be a recursive function defined as:

$$f(x) := \begin{cases} g_0(x_0) & \text{if } x = x_0 \\ g_1(x) + f(h(x)) + g_2(x) & \text{otherwise} \end{cases}$$

We assume that each $g_i : S_G \to S_M$ are pure and non-recursive functions, i.e. if $a \in S_G$ then $g_i(a) \in S_M$, for $i \in [0, 2]$. Furthermore, we assume that the *hop* function $h : S_G \to S_G$ is an invertible and monotonic function over S_G.

Proposition 1. *The recursive function $f : S_G \to S_M$ can be written as a functional composition.*

Proof. We let $f'_i : S_M \to S_M$ be the following non-recursive function:

$$f'_i(s) = g_1((h^{-1})^i(x_0)) + s + g_2((h^{-1})^i(x_0))$$

In the above definition, we let $h^{-1} : S_G \to S_G$ be the inverse function of the hop h. Function $(h^{-1})^i(x_0)$ is the i-th functional power of $h^{-1} : S_G \to S_G$. To transform the recursive function into a composition, it is important to infer the depth of the recursive stack. Let $k > 0 \in \mathbb{Z}$ such that $h^k(x) = x_0$. Thus $f(x) = (f'_k \circ f'_{k-1} \circ \cdots \circ f'_2 \circ f'_1)(g_0(x_0))$ □

From Fisher and Ghuloum [12], we know that the composition of f'_i can be computed in parallel since, for $i > 1 \in \mathbb{Z}$, we have that:

$$
\begin{aligned}
&(f'_i \circ f'_{i-1})(s) \Leftrightarrow g_1((h^{-1})^i(x_0)) + f'_{i-1}(s) + g_2((h^{-1})^i(x_0)) \Leftrightarrow \\
&g_1((h^{-1})^i(x_0)) + [g_1((h^{-1})^{i-1}(x_0)) + s + g_2((h^{-1})^{i-1}(x_0))] + g_2((h^{-1})^i(x_0)) \Leftrightarrow \\
&[g_1((h^{-1})^i(x_0)) + g_1((h^{-1})^{i-1}(x_0))] + s + [g_2((h^{-1})^{i-1}(x_0)) + g_2((h^{-1})^i(x_0))]
\end{aligned}
$$

3.2.1. Defining the Hop Function

Let $G = (S_G, +)$ be a group with identity $\mathbf{0}$. If the *hop* function h is an invertible and monotonic function, we can calculate k. Let h be generally defined as $h(x) = e_1 + x + e_2$, where $e_1, e_2 \in S_G$. Then

$$
\begin{aligned}
h^k(x) &= x_0 \Leftrightarrow \\
e_1 + \cdots + e_1 + e_1 + x + e_2 + e_2 + \cdots + e_2 &= x_0 \Leftrightarrow \\
ke_1 + x + ke_2 &= x_0 \Leftrightarrow \\
x + ke_2 &= (-ke_1) + x_0 \Leftrightarrow \\
x + ke_2 - x_0 &= (-ke_1) \Leftrightarrow \\
x + (ke_2 - x_0 + ke_1) &= \mathbf{0}
\end{aligned}
$$

that is, $\exists k \in \mathbb{Z}$ such that $k > 0$ and $(ke_2 - x_0 + ke_1) \in S_G$ is the inverse of $x \in S_G$. Given the equality above, k can often be dynamically computed with a small overhead. If G is commutative, then $(ke_2 - x_0 + ke_1) = k(e_1 + e_2) - x_0$. From these notions, we define function h^{-1}, the inverse of the hopping function, as $h^{-1}(x) = (-e_1) + x + (-e_2)$. This equality is true, since:

$$
\begin{aligned}
&(h^{-1} \circ h)(x) = (-e_1) + h(x) + (-e_2) \Leftrightarrow \\
&(h^{-1} \circ h)(x) = (-e_1) + [e_1 + x + e_2] + (-e_2) \Leftrightarrow \\
&(h^{-1} \circ h)(x) = \mathbf{0} + x + \mathbf{0} = x
\end{aligned}
$$

3.2.2. Computing the Function Composition using List Homomorphism

Thus far, we have seen how to re-write a function (with certain properties) as a composition of non-recursive functions. We need now a way to implement this composition efficiently. To achieve efficiency, we use list homomorphisms. We say that a function y is a list homomorphism if we have that $y(w \mathbin{+\!\!+} z) = y(w) \mathbin{+\!\!+} y(z)$, where $\mathbin{+\!\!+}$ denotes list concatenation. In this section we derive a simple implementation for such a recursive function $f : S_G \to S_M$ by means of list homomorphism.

Proposition 2. *Let $h'(i) = (h^{-1})^i(x_0)$. Then we can write the functional composition $f(x) = (f'_k \circ f'_{k-1} \circ \cdots \circ f'_2 \circ f'_1)(g_0(x_0))$ as:*

$$f(x) = (+/(g_1 \circ h') \star [k, k-1, \ldots, 1]) + g_0(x_0) + (+/(g_2 \circ h') \star [1, 2, \ldots, k]) \,.$$

Proof. The functional composition expands as follows:

$$\begin{aligned} f(x) &= g_1'(x_0) + g_0(x_0) + g_2'(x_0), \text{where:} \\ g_1'(x_0) &= g_1((h^{-1})^k(x_0)) + g_1((h^{-1})^{k-1}(x_0)) + \cdots + g_1((h^{-1})(x_0)) \\ g_2'(x_0) &= g_2((h^{-1})(x_0)) + g_2((h^{-1})^2(x_0)) + \cdots + g_2((h^{-1})^k(x_0)) \end{aligned}$$

Since $h'(i) = (h^{-1})^i(x_0)$, then we can write:

$$\begin{aligned} g_1'(x_0) &= g_1(h'(k)) + g_1(h'(k-1)) + \cdots + g_1(h'(1)) \\ g_2'(x_0) &= g_2(h'(1)) + g_2(h'(2)) + \cdots + g_2(h'(k)) \end{aligned}$$

Therefore, it is possible to compute g_1' and g_2' using a list homomorphism, i.e.:

$$\begin{aligned} g_1'(x_0) &= +/g_1 \star (h' \star [k, k-1, \ldots, 1]) \\ g_2'(x_0) &= +/g_2 \star (h' \star [1, 2, \ldots, k]) \end{aligned}$$

Where $+/\ell$ denotes the folding of the operation $+$ onto the list ℓ, and $k \star \ell$ denotes the mapping of function k onto every element of ℓ. The above expression can then be simplified as:

$$\begin{aligned} g_1'(x_0) &= +/(g_1 \circ h') \star [k, k-1, \ldots, 1] \\ g_2'(x_0) &= +/(g_2 \circ h') \star [1, 2, \ldots, k] \end{aligned}$$

□

3.3. Semirings

We now describe a second family of recursive functions that we can parallelize automatically by rethinking them under the light of algebraic structures. Let S_G and S_R be sets and $R = (S_R, +, \cdot)$ a semiring. Let $f : S_G \to S_R$ be defined as:

$$f(x) \coloneqq \begin{cases} g_0(x_0) & \text{if } x = x_0 \\ g_1(x) + g_3(x)f(h(x))g_4(x) + g_2(x) & \text{otherwise} \end{cases}$$

where each $g_i : S_G \to S_R$ is a non-recursive function. Function $h : S_G \to S_G$ is the *hop*. Let $f_i' : S_R \to S_R$ be the following non-recursive function:

$$f_i'(s) = g_1((h^{-1})^i(x_0)) + g_3((h^{-1})^i(x_0))sg_4((h^{-1})^i(x_0)) + g_2((h^{-1})^i(x_0))$$

where $(h^{-1})^i(x_0)$ is the i-th functional power of h^{-1}, the inverse function of the *hop* function h (see Section 3.2).

In order to transform the recursive function into a composition, again we must infer the depth of the recursive stack. Let $k > 0 \in \mathbb{Z}$ such that $h^k(x) = x_0$ (see Section 3.2). Thus:

$$f(x) = (f_k' \circ f_{k-1}' \circ \cdots \circ f_2' \circ f_1')(g_0(x_0))$$

3.3.1. Computing the Function Composition using List Homomorphism

In this section we derive a simple implementation of a recursive function $f : S_G \to S_R$ by means of list homomorphism. If $h'(i) = (h^{-1})^i(x_0)$, then we can write:

$$f(x) = \phi_1(k) + \phi_3(k)g_0(x_0)\phi_4(k) + \phi_2(k)$$

where $k > 0 \in \mathbb{Z}$ such that $h^k(x) = x_0$ (see Section 3.2). We have that:

$$\begin{aligned}
\beta_1(i) &= (\cdot/(g_3 \circ h') \star [k, k-1, \ldots, i+1])\, g_1(h'(i))\, (\cdot/(g_4 \circ h') \star [i+1, i+2, \ldots, k]) \\
\phi_1(k) &= g_1(h'(k)) + (+/\beta_1 \star [k-1, k-2, \ldots, 1]) \\
\phi_3(k) &= (\cdot/(g_3 \circ h') \star [k, k-1, \ldots, 1]) \\
\phi_4(k) &= (\cdot/(g_4 \circ h') \star [1, 2, \ldots, k]) \\
\beta_2(i) &= (\cdot/(g_3 \circ h') \star [k, k-1, \ldots, i+1])\, g_2(h'(i))\, (\cdot/(g_4 \circ h') \star [i+1, i+2, \ldots, k]) \\
\phi_2(k) &= (+/\beta_2 \star [1, 2, \ldots, k-1]) + g_2(h'(k))
\end{aligned}$$

There are redundant computations in the previous definition. We can optimize the computation of β_1 and β_2 by pre-computing these redundant values using a scan operation. Considering left- and right-associative scan operations (*scanl* and *scanr*), we define lists v and w as follows:

$$\begin{aligned}
[v_1, v_2, \ldots, v_{k-1}] &= \mathit{scanr} \cdot /(g_3 \circ h') \star [k, k-1, \ldots, 2] \\
[w_1, w_2, \ldots, w_{k-1}] &= \mathit{scanl} \cdot /(g_4 \circ h') \star [2, 3, \ldots, k]
\end{aligned}$$

That is, $v_i = \cdot/(g_3 \circ h') \star [k, k-1, \ldots, k-i+1]$ and $w_i = \cdot/(g_4 \circ h') \star [k-i+1, \ldots, k-1, k]$. Once we have pre-computed v_i and w_i, we can define the following simplified construction:

$$\begin{aligned}
\beta_1(i) &= v_i \cdot g_1(h'(i)) \cdot w_{k-i} \\
\phi_1(k) &= g_1(h'(k)) + (+/\beta_1 \star [k-1, k-2, \ldots, 1]) \\
\phi_3(k) &= v_1 \cdot (g_3 \circ h')(1) \\
\phi_4(k) &= (g_4 \circ h')(1) \cdot w_{k-1} \\
\beta_2(i) &= v_i \cdot g_2(h'(i)) \cdot w_{k-i} \\
\phi_2(k) &= (+/\beta_2 \star [1, 2, \ldots, k-1]) + g_2(h'(k))
\end{aligned}$$

3.4. Examples

In this section we discuss different functions that we can parallelize automatically. We shall provide examples that we can parallelize using the monoid-based approach (Catalan Numbers and List Concatenation), and with the semiring-based approach (Financial Compound Interest, Horner's Method and Comb Filters). The actual performance of each of these examples is analyzed in Section 4.

Catalan Numbers. Catalan numbers form a sequence of positive integers that appear in the solution of several counting problems in combinatorics, including some generating functions. Catalan numbers are defined as follows:

$$C_n = \frac{2(2n-1)}{n+1} C_{n-1} \quad \text{where } C_1 = 1$$

which can be written as $f : \mathbb{Z} \to \mathbb{Z}^*$

$$f(x) \coloneqq \begin{cases} 1 & \text{if } x = 1 \\ \frac{2(2x-1)}{x+1} \cdot f(x-1) & \text{otherwise} \end{cases}$$

Similar to the factorial function seen in Section 2, the above function can be written as a composition of non-recursive functions: Let $h : \mathbb{Z} \to \mathbb{Z}$ be $h(x) = x - 1$, $g_1 : \mathbb{Z} \to \mathbb{Z}^*$ be $g_1(x) = \frac{2(2x-1)}{x+1}$ and $g_2 : \mathbb{Z} \to \mathbb{Z}^*$ be $g_2(x) = 1$.

Then, we can define a function $f'_i : \mathbb{Z}^* \to \mathbb{Z}^*$, such as

$$\begin{aligned} f'_i(s) &= g_1((h^{-1})^i(1)) \cdot s \cdot g_2((h^{-1})^i(1)) \Leftrightarrow \\ f'_i(s) &= g_1(i+1) \cdot s \cdot 1 \Leftrightarrow \\ f'_i(s) &= \frac{2(2i+1)}{i+2} \cdot s \end{aligned}$$

since $h^{-1}(x) = x + 1$. We can easily calculate that $h^k(x) = 1$ for $k = x - 1$, since $x + (0k - 1 + k(-1)) = 0 \Leftrightarrow x - k - 1 = 0$.

Therefore, $f(x) = (f'_{x-1} \circ f'_{x-2} \circ \cdots \circ f'_2 \circ f'_1)(1)$. Since, for every $i > 1 \in \mathbb{Z}$, the composition $(f'_i \circ f'_{i-1})(s)$ can be symbolicaly computed and simplified, i.e.

$$\begin{aligned} (f'_i \circ f'_{i-1})(s) &= \frac{2(2i+1)}{i+2} \cdot f'_{i-1}(s) \Leftrightarrow \\ (f'_i \circ f'_{i-1})(s) &= \frac{2(2i+1)}{i+2} \cdot \left(\frac{2(2i-1)}{i+1} \cdot s \right) \Leftrightarrow \\ (f'_i \circ f'_{i-1})(s) &= \left(\frac{2(2i+1)}{i+2} \cdot \frac{2(2i-1)}{i+1} \right) \cdot s \end{aligned}$$

Function $f(x) = (f'_x \circ f'_{x-1} \circ \cdots \circ f'_1)(1)$ can be computed in parallel as a reduction.

List Concatenation. Let $L_{\mathbb{Z}}$ be the set of lists over the set of integers $\mathbb{Z}$. A list is an ordered sequence denoted by $A = [a_1, a_2, \ldots, a_n]$, where the size of A is $\#A = n$. An empty list is denoted by $[]$ and $\#[] = 0$. Concatenation is an associative binary-operation over a set of lists. Given two lists $A = [a_1, a_2, \ldots, a_n]$ and $B = [b_1, b_2, \ldots, b_m]$, the concatenation of the lists A and B is denoted by $A + + B = [a_1, \ldots, a_n, b_1, \ldots, b_m]$. The identity element regarding concatenation is the empty list. Thus$(L_{\mathbb{Z}}, +\!\!+\ )$ is a non-commutative monoid. Let $f : \mathbb{Z} \to L_{\mathbb{Z}}$ be the following recursive function:

$$f(x) \coloneqq \begin{cases} [] & \text{if } x = 0 \\ [x] +\!\!+ f(x-1) +\!\!+ [x] & \text{otherwise} \end{cases}$$

Let $h : \mathbb{Z} \to \mathbb{Z}$ be $h(x) = x - 1$, $g_1 : \mathbb{Z} \to L_{\mathbb{Z}}$ be $g_1(x) = [x]$ and $g_2 : \mathbb{Z} \to L_{\mathbb{Z}}$ be $g_2(x) = [x]$. Then, we can define a simplified function $f'_i : L_{\mathbb{Z}} \to L_{\mathbb{Z}}$, such as:

$$\begin{aligned}
&f'_i(s) = g_1((h^{-1})^i(0)) +\!\!+\ s +\!\!+\ g_2((h^{-1})^i(0)) \Leftrightarrow\\
&f'_i(s) = g_1(i) +\!\!+\ s +\!\!+\ g_2(i) \Leftrightarrow\\
&f'_i(s) = [i] +\!\!+\ s +\!\!+\ [i]
\end{aligned}$$

since $h^{-1}(x) = x + 1$. We have that $h^k(x) = 0$ for $k = x$, since $x + (0k - 0 + k(-1)) = 0 \Leftrightarrow x - k = 0$. Therefore, $f(x) = (f'_x \circ f'_{x-1} \circ \cdots \circ f'_2 \circ f'_1)([])$. Since, for every $i > 1 \in \mathbb{Z}$, the composition $(f'_i \circ f'_{i-1})(s)$ can be symbolicaly computed and simplified, i.e.:

$$\begin{aligned}
&(f'_i \circ f'_{i-1})(s) = [i] +\!\!+\ f'_{i-1}(s) +\!\!+\ [i] \Leftrightarrow\\
&(f'_i \circ f'_{i-1})(s) = [i] +\!\!+\ ([i-1] +\!\!+\ s +\!\!+\ [i-1]) +\!\!+\ [i] \Leftrightarrow\\
&(f'_i \circ f'_{i-1})(s) = ([i] +\!\!+\ [i-1]) +\!\!+\ s +\!\!+\ ([i-1] +\!\!+\ [i]) \Leftrightarrow\\
&(f'_i \circ f'_{i-1})(s) = [i, i-1] +\!\!+\ s +\!\!+\ [i-1, i]
\end{aligned}$$

Thus $f(x) = (f'_x \circ f'_{x-1} \circ \cdots \circ f'_1)([])$ can be computed in parallel by a reduction.

Financial Compound Interest. We can define financial compound interest with periodic deposits recursively. Let $f : \mathbb{Z} \to \mathbb{R}$ be the following recursive function:

$$f(x) := \begin{cases} y_0 & \text{if } x = 0 \\ (1+r) \cdot f(x-1) + y_x & \text{otherwise} \end{cases}$$

where y_0 is the initial deposit, r is the compounded rate, and y_x is the deposit on the x-th period. Let $h : \mathbb{Z} \to \mathbb{Z}$ is $h(x) = x - 1$, $g_1 : \mathbb{Z} \to \mathbb{R}$ is $g_1(x) = 0$, $g_3 : \mathbb{Z} \to \mathbb{R}$ is $g_3(x) = (1+r)$, $g_4 : \mathbb{Z} \to \mathbb{R}$ is $g_4(x) = 1$, $g_2 : \mathbb{Z} \to \mathbb{R}$ is $g_2(x) = y_x$. From these notions, we define a simplified function $f'_i : \mathbb{R} \to \mathbb{R}$ as follows:

$$\begin{aligned}
&f'_i(s) = g_1((h^{-1})^i(0)) + g_3((h^{-1})^i(0)) s g_4((h^{-1})^i(0)) + g_2((h^{-1})^i(0)) \Leftrightarrow\\
&f'_i(s) = g_1(i) + g_3(i) s g_4(i) + g_2(i) \Leftrightarrow\\
&f'_i(s) = (1+r)s + y_i \Leftrightarrow
\end{aligned}$$

since $h^{-1}(x) = x + 1$. We have that $h^k(x) = 0$ for $k = x$, since $x + (0k - 0 + k(-1)) = 0 \Leftrightarrow x - k = 0$. Hence, $f(x) = (f'_x \circ f'_{x-1} \circ \cdots \circ f'_2 \circ f'_1)(y_0)$. Since, for every $i > 1 \in \mathbb{Z}$, the composition $(f'_i \circ f'_{i-1})(s)$ can be symbolicaly computed and simplified, i.e.:

$$\begin{aligned}
&(f'_i \circ f'_{i-1})(s) = (1+r) f'_{i-1}(s) + y_i \Leftrightarrow\\
&(f'_i \circ f'_{i-1})(s) = (1+r)((1+r)s + y_{i-1}) + y_i \Leftrightarrow\\
&(f'_i \circ f'_{i-1})(s) = (1+r)(1+r)s + (1+r)y_{i-1} + y_i \Leftrightarrow\\
&(f'_i \circ f'_{i-1})(s) = (1+r)^2 s + [(1+r)y_{i-1} + y_i]
\end{aligned}$$

Thus $f(x) = (f'_x \circ f'_{x-1} \circ \cdots \circ f'_2 \circ f'_1)(y_0)$ can be computed in parallel.

Horner's Method. Horner's method is useful to solve polynomials defined recursively. Its implementation can be parallelized as done in the previous example of financial compound interest. Let c_i be the coefficients, for $0 \leq i \leq n$. Thus a polynomial of degree n can be evaluated, for a given value of x, by the following recursive formula, as described by the Horner's method:

$$f(n) := \begin{cases} c_0 & \text{if } n = 0 \\ f(n-1) \cdot x + c_n & \text{otherwise} \end{cases}$$

Comb Filter in Signal Processing. Comb filters have several applications in signal processing [18]. The following equation represents the feedback form used by comb filters: $y_t = \alpha y_{t-T} + (1-\alpha)x_t$, where x_t is the input signal at a given time t and α controls the intensity that the delayed signal is fed back into the output y_t given a delay time T. Let $f : \mathbb{Z} \to \mathbb{R}$ be the following recursive function:

$$f(t) := \begin{cases} y_0 & \text{if } t = 0 \\ \alpha f(t-T) + (1-\alpha)x_t & \text{otherwise} \end{cases}$$

3.5. Implementation

We have used the ideas discussed in this paper to implement a source-to-source compiler based on rewriting rules and symbolic computation. Term rewriting systems are usually concerned with computing reduced forms of a given term with respect to a pre-defined set of rules. Several existing compilers also use rewriting rules as an infrastructure for implementing many specific optimizations [19, 20, 21, 22]. Compilers based on rewriting rules check that all term manipulations result in correctly typed terms, before performing the transformation specified by the appropriate rewriting rule that the matches with the given term [23].

Our source-to-source compiler uses rewriting rules in order to automatically parallelize recursive functions that match either the monoid-based or the semiring-based transformation rules. We also use symbolic computing to find the inverse of the hop function and also to infer the depth of the recursive stack (as discussed in Sections 3.2 and 3.3). Below we show the source-to-source transformation that we produce for the list concatenation benchmark:

```
# Sequential version:
f :: Integer -> [Integer]
f 0 = []
f n = [n] ++ f(n-1) ++ [n]

# Parallel version:
f_g_1 :: Integer -> [Integer]
f_g_1 i = [i]
f_g_2 :: Integer -> [Integer]
f_g_2 i = [i]
f :: Integer -> [Integer]
f n = let k = n
```

```
      in (parFoldr (++) (map f_g_1 (reverse [1..k]))) ++ [] ++
         (parFoldr (++) (map f_g_2 [1..k]))
```

In general, for the monoid-based case, we receive inputs in the format below:

```
f :: Integer -> IMGSET
f e_0 = y_0
f n = g_1(n) * f(n-e_1) * g_2(n)
```

The parallel code, automatically generated by the source-to-source compiler, consists of three new functions: f_g_1, f_g_2 and f, which have the following general format:

```
f_g_1 :: Integer -> IMGSET
f_g_1 i = g_1((i*e_1 + e_0))
f_g_2 :: Integer -> IMGSET
f_g_2 i = g_2((i*e_1 + e_0))
f :: Integer -> IMGSET
f n = let k = ((n - e_0)/e_1)
      in (parFoldr (*) (map f_g_1 (reverse [1..k]))) * y_0 *
         (parFoldr (*) (map f_g_2 [1..k]))
```

In the general case based on semirings, we receive inputs in the format below:

```
f :: Integer -> IMGSET
f e_0 = y_0
f n = g_1(n) + g_3(n) * f(n-e_1) * g_4(n) + g_2(n)
```

The generated parallel code, without using the scan-based optimization (as described in Section 3.3), has the following format:

```
f_g_1 :: Integer -> IMGSET
f_g_1 i = g_1((i*e_1 + e_0))
f_g_1 :: Integer -> IMGSET
f_g_2 i = g_2((i*e_1 + e_0))
f_g_1 :: Integer -> IMGSET
f_g_3 i = g_3((i*e_1 + e_0))
f_g_1 :: Integer -> IMGSET
f_g_4 i = g_4((i*e_1 + e_0))
f_B_1 k i = (foldr1 (*) (map f_g_3 (reverse [(i+1)..k]))) *
            (f_g_1 i) * (foldr (*) (map f_g_4 [(i+1)..k]))
f_B_2 k i = (foldr1 (*) (map f_g_3 (reverse [(i+1)..k]))) *
            (f_g_2 i) * (foldr (*) (map f_g_4 [(i+1)..k]))
f_PHI_1 k = (f_g_1 k) +
            (parFoldr (+) (map (f_B_1 k) (reverse [1..(k-1)])))
f_PHI_2 k = (parFoldr (+) (map (f_B_2 k) [1..(k-1)])) + (f_g_2 k)
f_PHI_3 k = (parFoldr (*) (map f_g_3 (reverse [1..k])))
f_PHI_4 k = (parFoldr (*) (map f_g_4 [1..k]))
f n = let k = ((n - e_0)/e_1))
      in (f_PHI_1 k) + (f_PHI_3 k)*(y_0)*(f_PHI_4 k) + (f_PHI_2 k)
```

However, this parallel code without using the scan-based optimization is not sufficiently efficient, due to many redundant computation in several calls to functions `f_B_1` and `f_B_2`. Because of that, we also implement the scan-based optimization for the recursive functions based on semirings, as described in Section 3.3. The generated parallel code, with the scan-based optimization, has the following format:

```
f_g_1 :: Integer -> IMGSET
f_g_1 i = g_1(i*e_1 + e_0)
f_g_1 :: Integer -> IMGSET
f_g_2 i = g_2(i*e_1 + e_0)
f_g_1 :: Integer -> IMGSET
f_g_3 i = g_3(i*e_1 + e_0)
f_g_1 :: Integer -> IMGSET
f_g_4 i = g_4(i*e_1 + e_0)
f_B_1 v w k i = (v!!i)*(f_g_1 i)*(w!!(k-i))
f_B_2 v w k i = (v!!i)*(f_g_2 i)*(w!!(k-i))
f_PHI_1 v w k = (f_g_1 k) +
                (parFoldr (+) (map (f_B_1 v w k) (reverse [1..(k-1)])))
f_PHI_2 v w k = (parFoldr (+) (map (f_B_2 v w k) [1..(k-1)])) + (f_g_2 k)
f_PHI_3 v w k = (v!!1) * (f_g_3 1)
f_PHI_4 v w k = (f_g_4 k) * (w!!(k-1))
f n = let k = ((n - e_0)/e_1))
          v = scanr1 (*) (map f_g_3 (reverse [2..k]))
          w = scanl1 (*) (map f_g_4 [2..k])
      in (f_PHI_1 v w k)) +
         (f_PHI_3 v w k)*(y_0)*(f_PHI_4 v w k) +
         (f_PHI_2 v w k)
```

While the scan-based optimization is general and can always be performed, there are other optimizations that could be performed on specific cases. One such optimization that can be performed on the specific case of handling numeric terms, with the addition and multiplication operators, is the following: If the functions `f_g_3` and `f_g_4` are constant numeric functions, the scan operations could be avoided by replacing the index accessing with the equivalent symbolic solution for the scan computation, e.g. `(v!!i)` could be replaced with either `(g_3 e_0)*i` or `(g_3 e_0)^i`, if the scan operation is addition or multiplication, respectively. On the case of boolean terms with the operators *or* and *and*, if `f_g_3` and `f_g_4` are constant boolean functions, the index accessing could be replaced directly with the function expression, e.g. `(v!!i)` could be replaced directly with just `(g_3 e_0)`.

In order to generate parallel code in Haskell, we use the parallel library provided by the `Strategies` package, available in the Glasgow Haskell Compiler (GHC) [24, 5, 25]. In particular, we use the following implementation for the parallel fold-right operation.

```
parFoldr _ [x] = x
parFoldr mappend xs  = (ys ‘par‘ zs) ‘pseq‘ (ys ‘mappend‘ zs) where
  len = length xs
  (ys’, zs’) = splitAt (len ‘div‘ 2) xs
  ys = parFoldr mappend ys’
  zs = parFoldr mappend zs’
```

The expression `(ys ‘par‘ zs)` *sparks* the evaluation of `ys` and `zs` in parallel, while `‘pseq‘` guarantees that they have been evaluated before evaluating `(ys ‘mappend‘ zs)`. In GHC, *sparks* are not immediately executed, instead, they are queued for execution in FIFO order. The runtime converts a *spark* into a real thread when there is an idle CPU, and then run the new thread on the idle CPU, in such a way that the available parallelism is spread amongst the CPUs. In order to provide load-balancing in the parallel implementation, each processor has a *spark pool* that supports *lock-free work-stealing* [5].

4. Evaluation

In this section, we first describe basic aspects of the Glasgow Haskell Compiler (GHC) runtime system and then we discuss our experimental results.

4.1. Experimental Results

To validate the ideas discussed in this paper, we analysed the source-to-source Haskell compiler, implemented as described in Section 3.5, with the benchmarks presented in Section 3.4. The experiments were performed in an Intel i5 quad-core processor, with 3.20 GHz of clock and 16 GiB of RAM. In this section, we show results for six benchmarks – three illustrating monoid-based parallelization (factorial, catalan and concatenation), and three illustrating semiring-based parallelization (compound interest, Horner’s method and comb filter). For the experiments with all six applications, we evaluated their speedup and scalability when varying the number of threads from 1 to 4, while fixing the input argument for each monoid-based and semiring-based benchmark in 100,000 and 10,000, respectively. We consider the average over a total of five executions.

Monoids. Figure 1 presents the results for the parallelization of recursive functions based on monoids. For the parallelization of the three monoid-based applications we implemented the construction using the list homomorphism presented in Section 3.2. Parallelization was achieved by means of a right-associative fold operator implemented using the Parallel Haskell library. We achieved a maximum speedup of 1.92×, over the execution with a single thread, in the Catalan benchmark, and an average of 1.45× speedup with four threads for all the monoid-based benchmarks. If we ignore the list concatenation benchmark, the parallelization provides an average improvement of 9.20× over the original purely recursive implementation, with a maximum speedup of 11.70× and a minimum of 6.69×. The original recursive implementation of the list concatenation benchmark has an unreasonably long execution-time for very large inputs, such as the one used in our experimental results. If we compare the number of registered garbage collection events in the profiling information, namely, the event *GC working*, the parallel implementation with four threads records 2,351 instances of such event (with an event log file of about 400 KiB), while the sequential recursive implementation incredibly records 7,232,694 instances of the

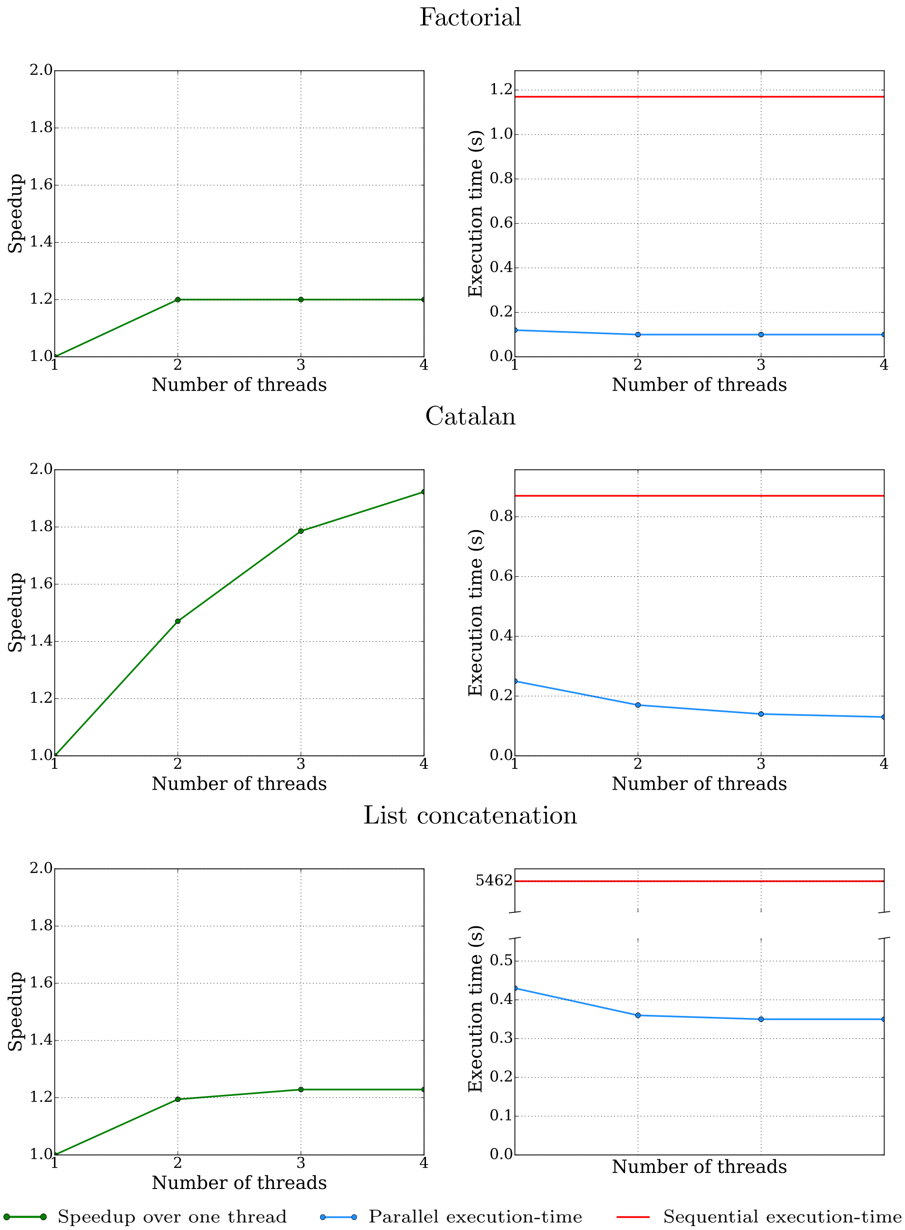


Figure 1: Analysis of the scalability (left) and the execution-time (right) of the monoid-based benchmarks with input argument fixed as 100,000. Scalability compares the runtime of the parallel implementations that we generate automatically, running with 1, 2, 3 and 4 threads.

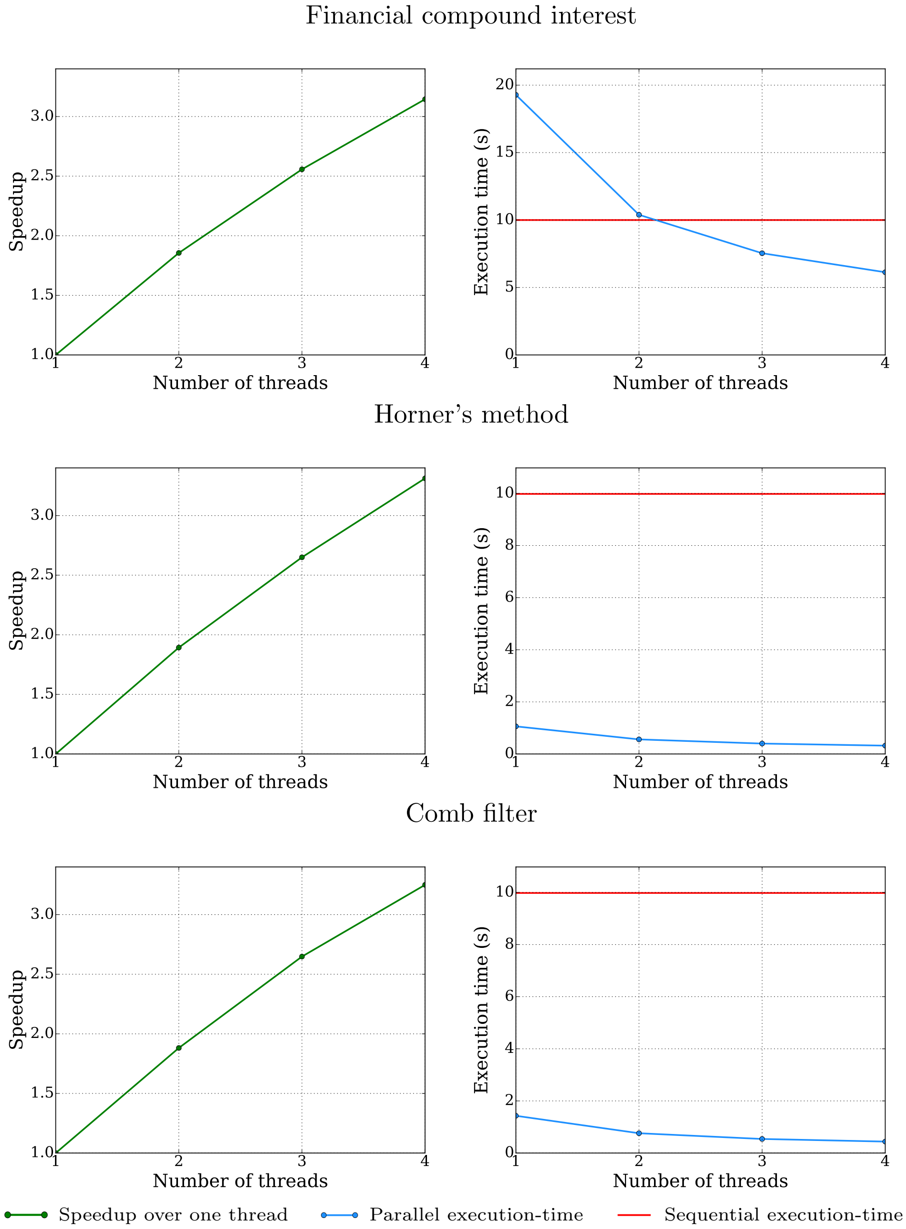


Figure 2: Analysis of the scalability (left) and the execution-time (right) of the semiring-based benchmarks with input argument fixed as 10,000.

same event (with an event log file of almost 4 GiB). Although we have not been able to precisely identify the reason for this extremely long execution-time, the garbage collector seems to be a major source of overhead. An experiment involving the execution of list concatenation with a smaller input of 20,000 cells (see Figure 3) reveals that garbage collection starts to have a large impact on its performance. A total of 72% of the execution time is spent with the GC, with a total of 134.8 GiB of allocated memory during 107.82 seconds of execution. The parallel fold implementation of list concatenation does not suffer these shortcomings. Here, we emphasize that our sequential implementation of list concatenation follows the naïve description of this algorithm, with two patterns, which can be found in any textbook on functional programming.

Semirings. Figure 2 shows the results for the parallelization of recursive functions based on semirings. For the parallelization of the three semiring-based applications we implemented the construction using the list homomorphism discussed in Section 3.3. We have used the same parallel implementation of the right-associative fold operator which we applied on the monoid-based benchmarks. Our highest speedup was 3.31× in the Horner's Method benchmark over the single threaded execution, with an average speedup of 3.23× for the executions with four threads. The parallelization provides an average improvement 18.50× over the original recursive implementation, with a maximum speedup of 31.19× and a minimum of 1.63×.

4.2. Discussion

On the Quality of Our Sequential Implementations. In both the monoid and semiring-based cases, even the transformed code executed with a single thread can present better performance than the original purely recursive implementation. This performance gain can be attributed to optimizations that can take better advantage of fold-based implementations, such as deforestation [26, 27]. We would like to emphasize that the sequential programs are naïve implementations of well-known algorithms. No attempt to optimize them has been made. We emphasize that our goal is not to speedup those functions via classic compiler optimizations. Those functions work only as starting points, from where we can derive parallel code – they have not been conceived to run fast.

On our Speedups. We have been able to observe actual speedups on the six benchmarks that we have played with. These speedups were usually sublinear, e.g., we could not observe a four-fold speedup in any of the cases. We believe that this sublinearity is due to the overhead imposed by the reduction operator required by the proposed parallel construction. Nevertheless, we would like to emphasize that all these results have been obtained by means of automatic transformations. In other words, the use of our techniques does not require any intervention from the programmer who has implemented the original version of each function that we parallelize.

Two benchmarks, namely, factorial and list concatenation, have presented limited scalability when varying the number of threads. We suggest two main

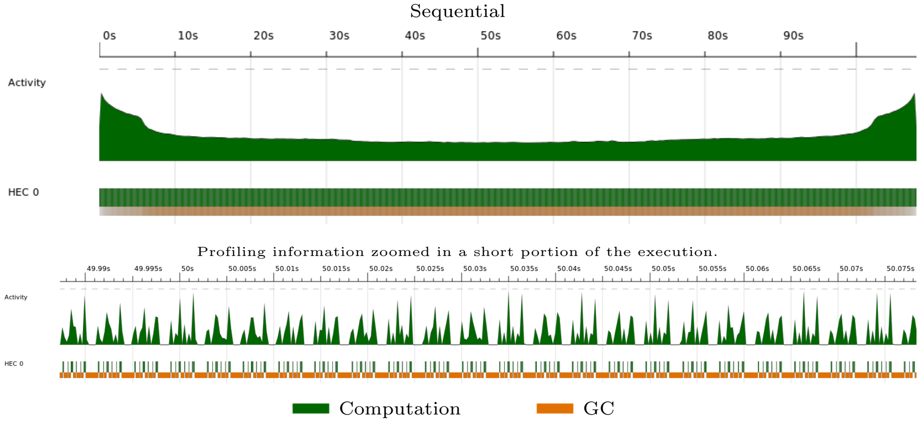


Figure 3: Profiling information for the list concatenation benchmark when concatenating two lists of 20,000 cells. Notice how the GC poses a large overhead in the execution time. About 70% of the execution time is exclusively dedicated to garbage collection.

factors that prevented better scalability: (*i*) both benchmarks produce reasonably large output, such that writing the output is a sequential portion that represents a significant percentage of the total execution time; (*ii*) garbage collection is implemented in a stop-the-world fashion. Concerning the first point, factorial and list concatenation produce reasonably large outputs, e.g. the factorial of 100,000 is a number of 456,575 digits and the output for the list concatenation is a string with 477,789 characters. In both cases, writing the output takes on average about 0.05 seconds, which represents a significant percentage of the total execution time, e.g. it is about half the execution time for the factorial benchmark. In Figure 4, the sequential portion during the second half of the execution mainly represents the processing required for writing the output. We believe that the last point – garbage collection – is sufficiently important to deserve a deeper analysis. Such analysis is the subject of the rest of this section.

4.3. The Impact of Garbage Collection in our Results

The GHC Runtime System. The Glasgow Haskell Compiler (GHC) [24, 25] runtime system consists of multiple *Haskell Execution Contexts* (HECs), which are virtual processors responsible for executing Haskell threads, i.e., the available *sparks* queued for execution. The runtime system maintains exactly one HEC for each CPU. Normally, HECs execute without any synchronization, locks, or atomic instructions. Synchronization is only required for load balancing, garbage collection, and other similar management tasks [28].

GHC implements a garbage collector (GC) that is a parallel, generational, copying collector [28], which means that it has a multi-threaded implementation, as opposed to a GC that runs concurrently with the program. Originally, the GC only run when all HECs have stopped and agreed to garbage collect, in a *stop-the-world* fashion [5]. More recently, there has been efforts into implementing a thread-local garbage collection [29], which is able to independently collect allocated space in the local heap of each thread, while the shared global heap still requires that all processors synchronize and cooperate in a parallel global garbage collection.

Profiling the Effects of Garbage Collection on Performance. In Figures 4, 5, 6, 7, and 8, the green area represents actual Haskell computation and the orange area represents garbage collection activity. At the top of each figure, the combined activity of all HECs are illustrated in a single curve, considering only the actual Haskell computation of the benchmark. Those figures also show a detailed view of the activity in each HEC at any given time during the execution of the benchmarks, where it differs Haskell computation in green and garbage collection in orange. HECs can also be idle, a state that is illustrated in white in our figures. We notice many occurrences of global garbage collection, where some HECs may be idle waiting for the GC to finish. We also notice that, in same cases, a thread-local garbage collection takes place, where at the same instant one HEC is performing Haskell computation while another HEC is performing garbage collection in its local heap [29].

From the profiling information shown in Figure 4 and Figure 6, we can conclude that the garbage collector poses a significant overhead onto factorial and list concatenation. For the factorial benchmark, the percentage of time spent with garbage collection varied from 24.4% (with a single thread) to 29.8% (with four threads), with the maximum pause for garbage collection varying from 0.0026 seconds to 0.0150 seconds. The synchronization during garbage collection, in a stop-the-world fashion, prevents the application to scale when increasing the number of threads.

Garbage collection represents a major portion of all the execution time in list concatenation, as shown in Figure 6. This benchmark has an allocation rate of more than 5 GiB/s, in respect only of the time spent in actual Haskell computation, allocating a total of about 750 MiB. Because of this particular aspect, the percentage of time spent with garbage collection is more than 65% of the total execution time. Increasing the number of threads does not affect much the garbage collection overhead, since the synchronization overheads are shadowed by performance gains from the parallel garbage collection, which is particularly beneficial for the list concatenation benchmark and its massive allocation requirements. However, the parallelizable portion of the actual application represents a very small percentage of the total execution time.

Because the Catalan benchmark has a smaller output, and a smaller memory footprint, it benefits more from parallelism, when compared to the other two monoid-based benchmarks (see Figure 5). Due to its reasonable parallelizable portion and smaller garbage collection overhead, the Catalan benchmark presents good scalability, with a maximum speedup of 1.92× with four threads, improving 6.69× over the original purely recursive implementation.

All three semiring-based benchmarks have a similar runtime behavior, as illustrated in Figures 7 and 8. In those three cases, the sequential portion that writes the output is negligible compared to the parallelizable portion of the computation. Moreover, the garbage collection also poses a negligible overhead

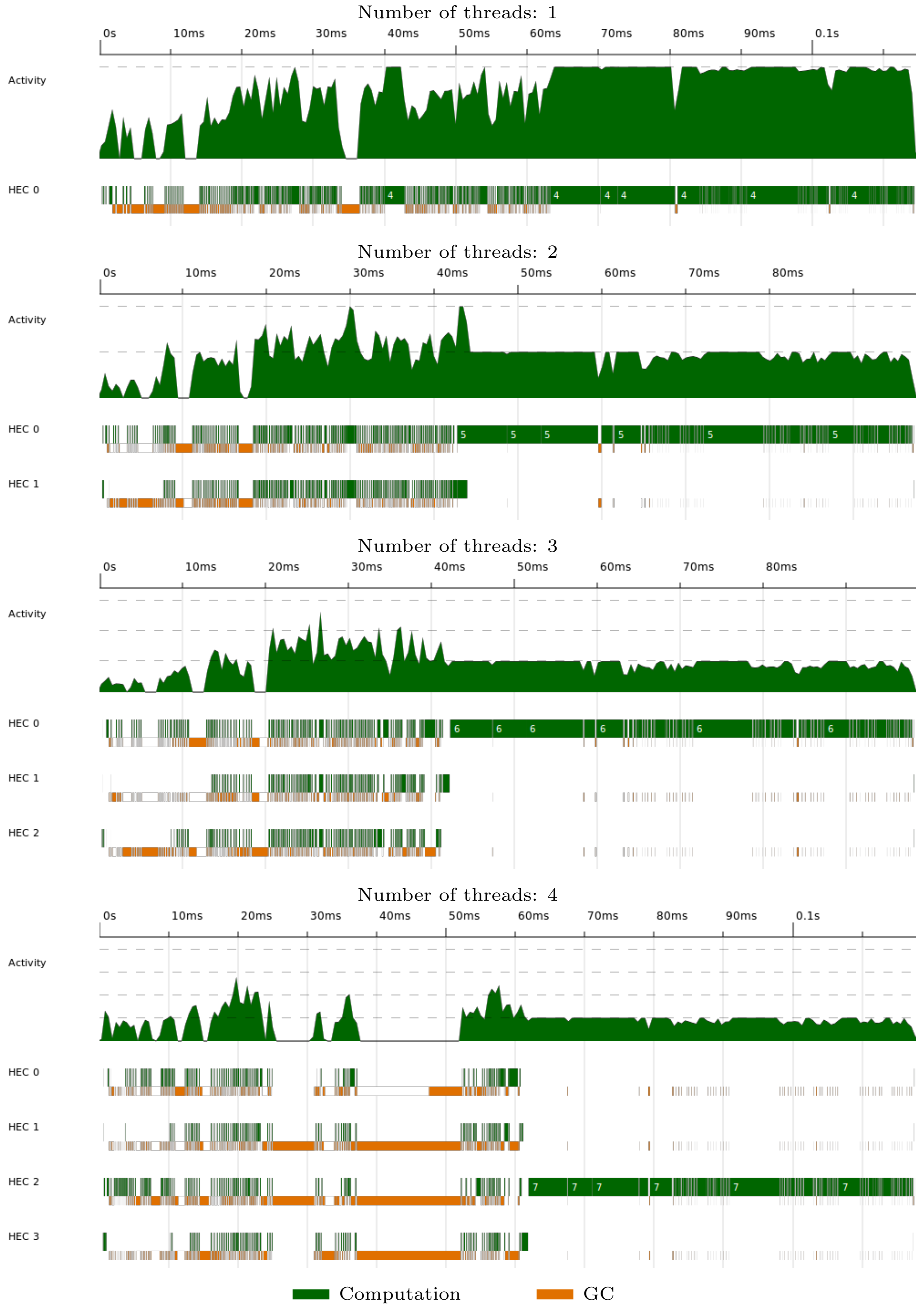


Figure 4: Profiling information for the Factorial benchmark. Notice how the second half of the execution is a uniquely sequential process, in this case, dedicated to writing the large output.

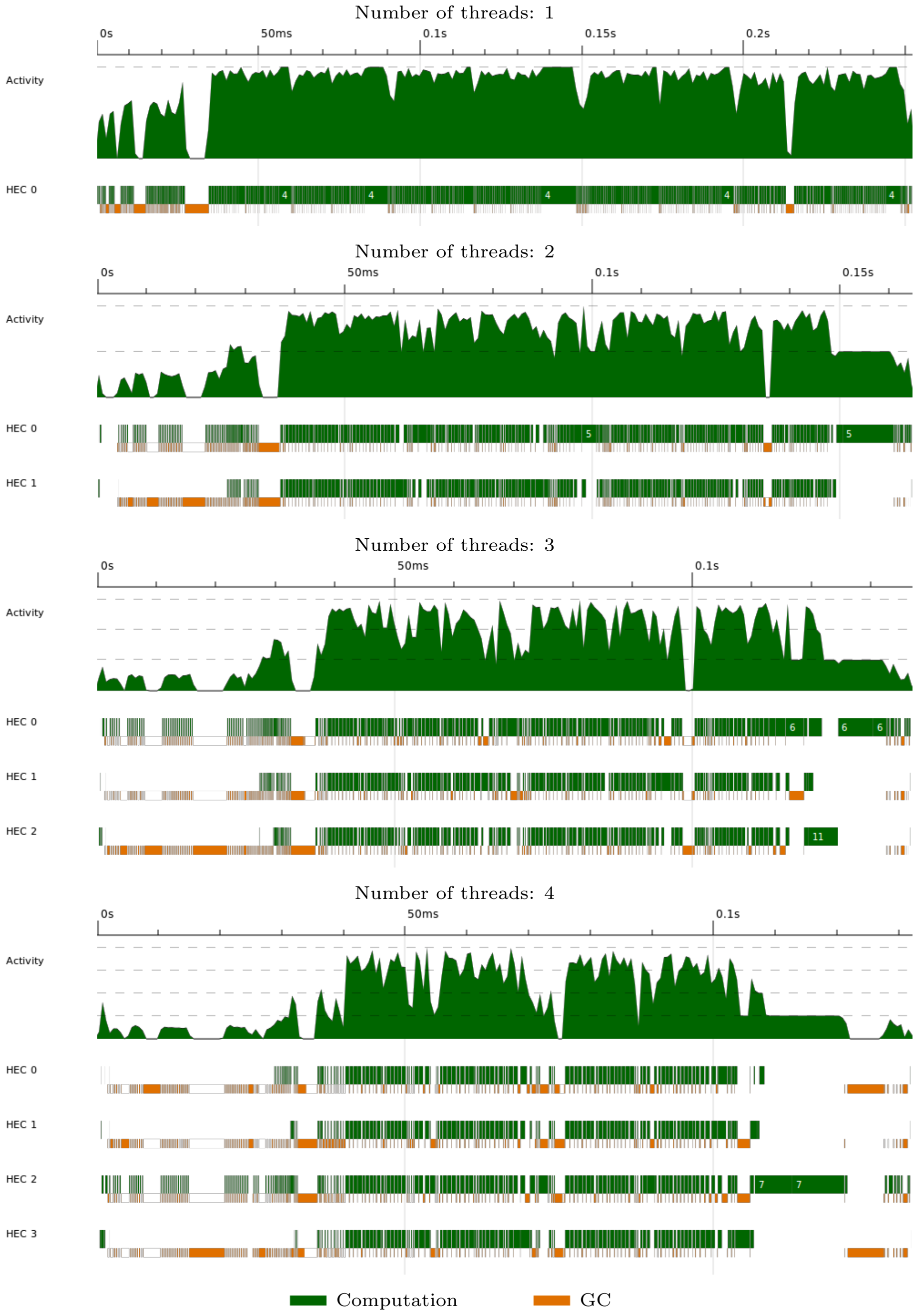


Figure 5: Profiling information for the Catalan benchmark. This benchmark has a reasonably large parallelizable portion and a fairly low garbage collection overhead.

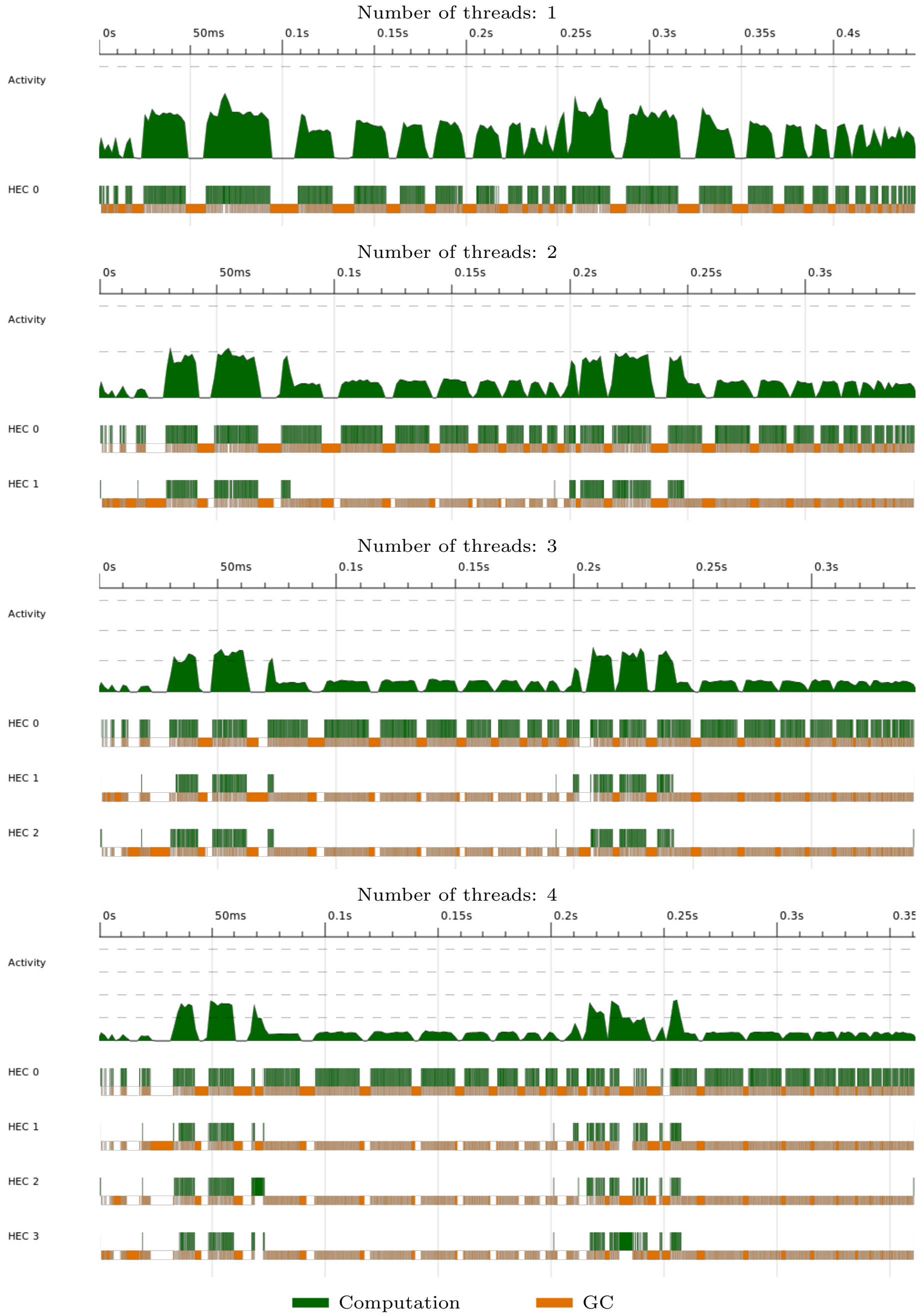


Figure 6: Profiling information for the List concatenation benchmark. The percentage of time spent in garbage collection represents more than 65% of the total execution time of this benchmark, posing a large overhead during its execution.

during runtime. Given the combination of these well behaved runtime aspects, all three benchmarks present good scalability.

5. Related Work

One of the perceived advantages of functional programming languages has always been the facility in which programs written in these languages could be parallelized. Advocates of parallelism in pure functional programming claim that programmers should not be concerned about how to parallelize their code – such task should be given to the compiler. Thus, historically, automatic parallelization has been part of the design of many purely functional programming languages [30]. Yet, we believe that an old statement, by Carriero and Gelernter is still valid: "This capability [of compilers] remains to be demonstrated, and achieving it automatically and generally strikers us as a difficult problem" [31]. This paper has presented a technique to address some of the shortcomings that Carriero and Gelernter have raised. Like us, several different researchers have proposed techniques to support the automatic parallelization of functional code.

Map-Reduce Patterns. There are several automatic parallelization techniques that, similarly to ours, seek common patterns in code. Perhaps the most widespread parallel patterns are map and reduce [32]. Map is inherently parallel, and runs in $O(1)$ on the PRAM world. Reduce can be parallelized up to $O(\ln n)$ time under that model. Thus, it comes as no surprize that both patterns can be discovered automatically, and subsequently transformed to run in parallel [10]. This kind of transformation is possible even in non-functional programs, as Mata *et al.* have demonstrated on C [33], or Morais *et al.* have demonstrated on Dinamica EGO [34, 35].

Polyhedral Patterns. Strategies based on matrix-multiplication are another well-known example of mining of parallel pattern. Kogge and Stone [36] have shown how to parallelize a recurrence equation by rewriting it in a form of matrix multiplication, also called *state-vector update* form. An expression e in a loop is a recurring expression if, and only if, e is computed from some loop-carried value. Sato and Iwasaki [37] have described a framework based on matrix multiplication for automatically parallelizing affine loops that consist of reduce or scan operations. They have also provided algorithms for recognizing the normal form and max-operators automatically. They have been able to report considerable speedups and high scalability by applying their framework onto simple benchmarks. Also along this line, Zou and Rajopadhye [38] have proposed a way to parallelize scan operations using the matrix multiplication framework with the polyhedral model [39, 40, 41]. They can handle arbitrary nested affine loops; the polyhedron model itself has already been used to parallelize different types of loops in imperative programming languages [42]. Contrary to our work, these previous approaches search for a way to deconstruct a loop as multiplication of matrices, we search for a way to deconstruct a function as a composition of monoid/semiring operations. The programs that can be parallelized by these two approaches are different.

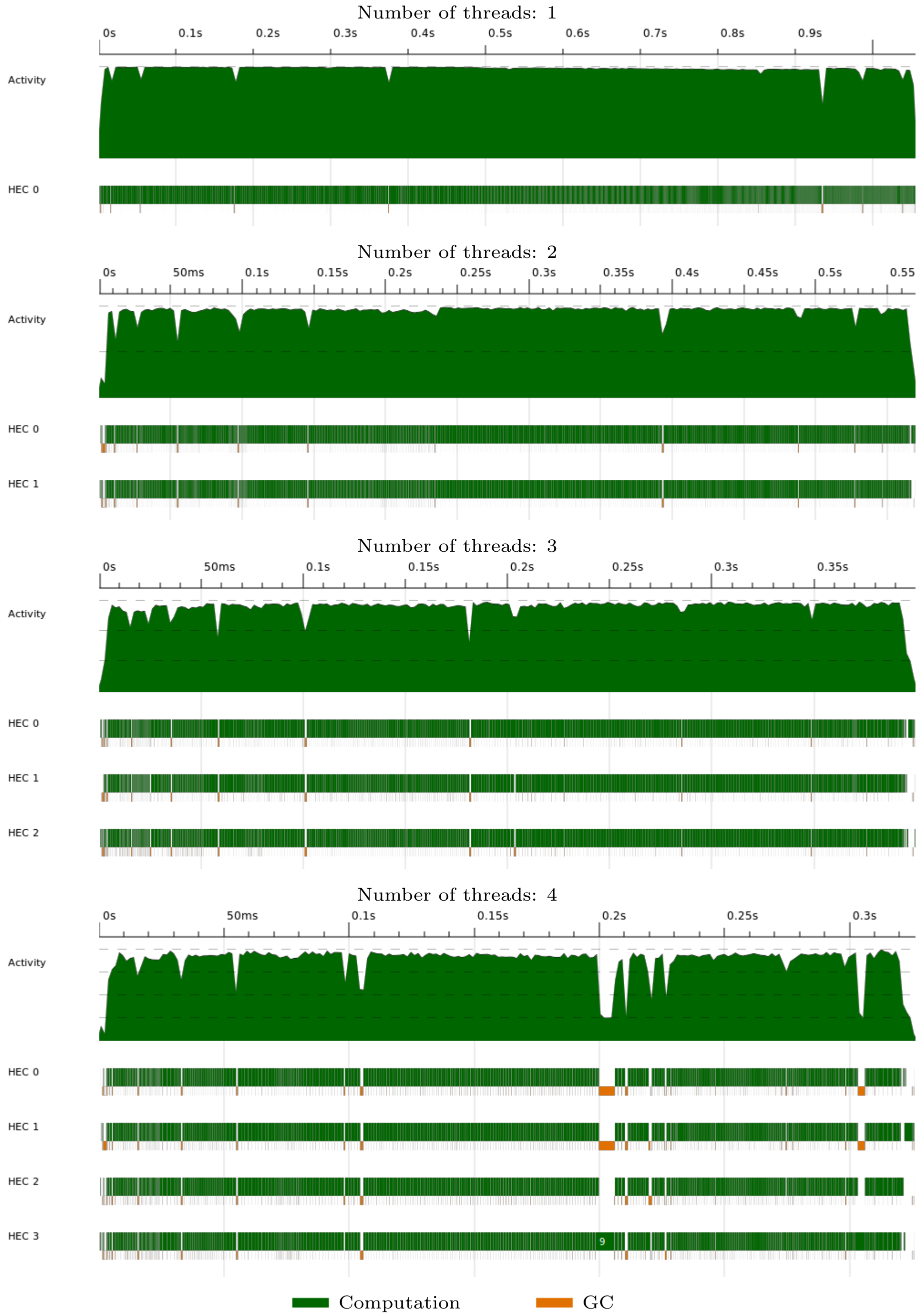
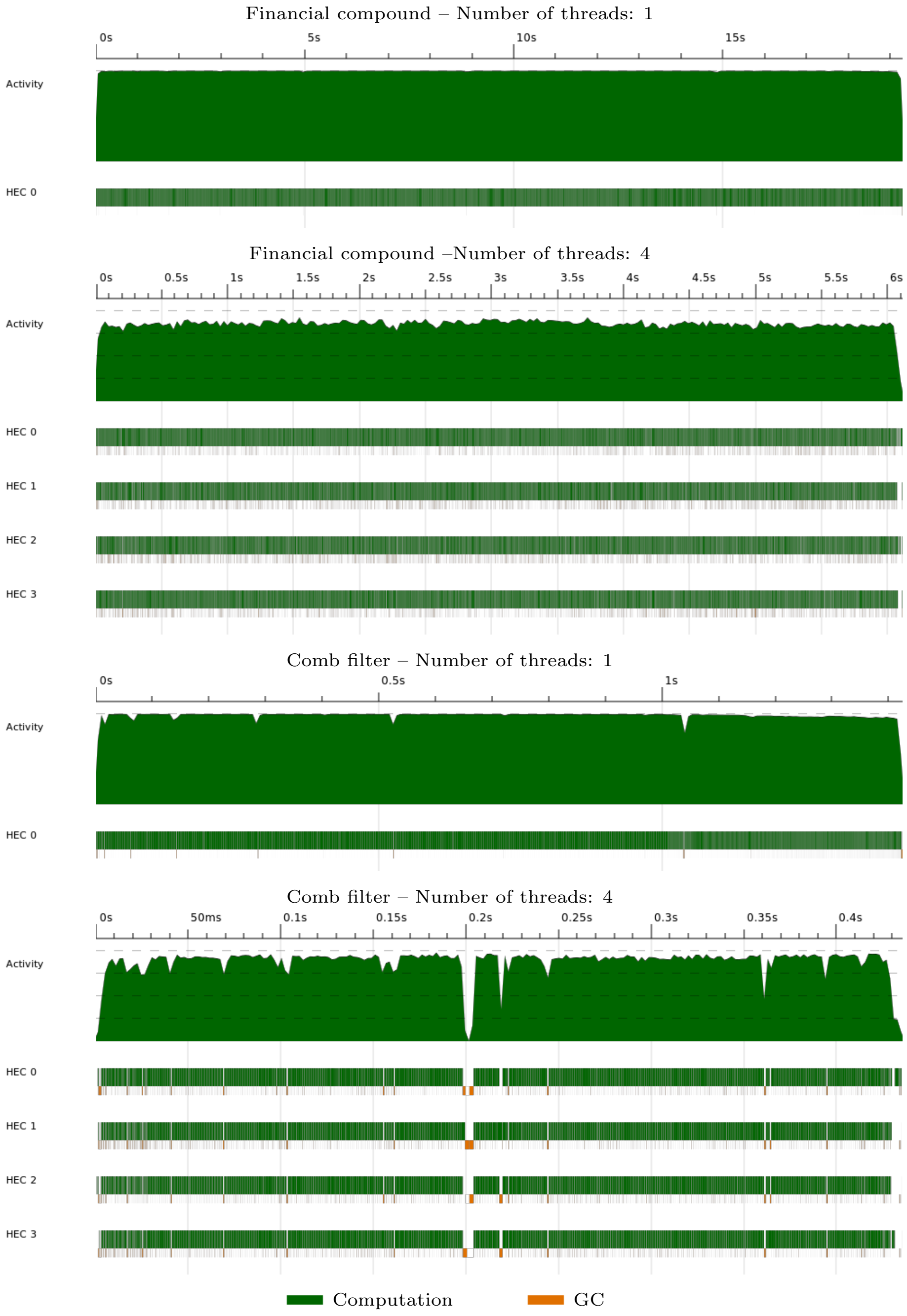


Figure 7: Profiling information for the Horner's method benchmark. Notice how the runtime system is able to leverage parallelism, with very little synchronization overhead as the number of threads increases, benefiting from load-balancing by the GHC runtime system.



Financial compound – Number of threads: 1

Financial compound –Number of threads: 4

Comb filter – Number of threads: 1

Comb filter – Number of threads: 4

Figure 8: Profiling information for the Financial and the Comb filter benchmarks. Because these benchmarks behave similarly to Horner's method, we only show executions with 1 and 4 threads.

Rewriting Rules. Rewriting rules are a well-known optimization technique, which gives programmers the ability to interact with the compiler. Such rules let the programmer specify relations that, once applied onto the program, yield better code [19]. This technique has been used with the goal to generate parallel code out of programs written in a sequential mindset. Steuwer et al. [20] use rewriting rules to perform source-to-source transformations, translating high-level functional expressions into a low-level functional representation close to OpenCL. They use a pre-defined set of rewriting rules in order to automatically search for an efficient device-specific OpenCL code. Although they provide good portable performance results [20, 21], their current implementation does not allow for automatically parallelizing recurrence expressions.

Composition Patterns. Fisher and Ghuloum [12] provide a generalized formalization for automatic parallelization of loops by extracting function composition as the main associative operator. If a function is closed under composition, its compositions can be computed efficiently. They describe loops that compute reduction or scan as the composition of its *modeling function*. For loops that fit the allowed format, they can be implemented in a manner that computes the composition of the modelling function in parallel. Our work improves on theirs, because we extend their approach to recursive functions. In fact, this is our main contribution: a general way to extract parallelism buried under the syntax of potentially convoluted recursive implementations. In addition, we also provide a more general definition of parallel code by means of algebraic structures such as monoids and semirings.

List Homomorphism Patterns. There exists vast literature about list homomorphisms [8, 9, 10, 11]. List homomorphism is a special class of natural recursive functions on lists, which has algebraic properties suitable for parallelism. These properties can also be extended to other data-structures, such as trees, as demonstrated by Morihata *et al* [43]. We rely on list homomorphisms to build efficient parallel computations of recursive functions. However, our approach does not target exclusively functions that work on lists. As many of our examples illustrate, we are able to generate parallel code for functions involving just primitive types, or even for more complex data-structures that can be processed by monoid-based operators.

6. Conclusion

This paper has presented a theoretical approach to parallelize recursive functions. This contribution is important because previous work has reported difficulties to infer parallel behavior out of recursive functions. In this case, parallelism is usually buried under heavy and convoluted syntax. We have delineated two classes of recursive functions which we can parallelize automatically. These functions have the following property: they can be re-written as the combination of themselves (through a recursive call) with non-recursive functions by means of monoid or semiring operators. There are several examples of functions that fit this framework, including typical functional implementations of algorithms to compute factorials, sum up elements of lists, concatenate lists, etc.

As future work, we intend to broaden the classes of recursive functions that our algebraic framework can automatically parallelize. We also intend to perform optimizations on the parallel code generated by our source-to-source compiler. Another important future contribution, would be to work on lowering the overhead posed by the garbage collector in parallel Haskell programs.

Acknowledgments

This work was partially supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grants EP/L01503X/1 for the University of Edinburgh, School of Informatics, Centre for Doctoral Training in Pervasive Parallelism (`http://pervasiveparallelism.inf.ed.ac.uk/`), and also by the Institute for Computing Systems Architecture (ICSA) in the School of Informatics at the University of Edinburgh. Fernando Pereira is supported by grants from CAPES, FAPEMIG and CNPq.

References

[1] R. Govindarajan, J. Anantpur, Runtime dependence computation and execution of loops on heterogeneous systems, in: Proceedings of the 2013 IEEE/ACM International Symposium on Code Generation and Optimization, 2013, pp. 31:1–31:10.

[2] S. Misailovic, D. Kim, M. Rinard, Parallelizing sequential programs with statistical accuracy tests, ACM Transactions on Embedded Computing Systems 12 (2) (2013) 88:1–88:26.

[3] Z. Wang, G. Tournavitis, B. Franke, M. F. P. O'Boyle, Integrating profile-driven parallelism detection and machine-learning-based mapping, ACM Transactions on Architecture Code Optimization 11 (1) (2014) 2:1–2:26.

[4] K. Hammond, J. Berthold, R. Loogen, Automatic skeletons in template haskell, Parallel Processing Letters 13 (03) (2003) 413–424.

[5] S. Marlow, S. Peyton Jones, S. Singh, Runtime support for multicore haskell, in: Proceedings of the 14th ACM SIGPLAN International Conference on Functional Programming, ACM, 2009, pp. 65–78.

[6] C. Brown, M. Danelutto, K. Hammond, P. Kilpatrick, A. Elliott, Cost-directed refactoring for parallel erlang programs, International Journal of Parallel Programming 42 (4) (2013) 564–582.

[7] A. Collins, D. Grewe, V. Grover, S. Lee, A. Susnea, NOVA: A functional language for data parallelism, in: Proceedings of the 2014 ACM SIGPLAN International Workshop on Libraries, Languages, and Compilers for Array Programming, 2014, pp. 8–13.

[8] M. Cole, Parallel programming with list homomorphisms, Parallel Processing Letters 5 (02) (1995) 191–203.

[9] Z. Hu, H. Iwasaki, M. Takeichi, Formal derivation of efficient parallel programs by construction of list homomorphisms, ACM Transactions on Programming Languages and Systems 19 (3) (1997) 444–461.

[10] K. Morita, A. Morihata, K. Matsuzaki, Z. Hu, M. Takeichi, Automatic inversion generates divide-and-conquer parallel programs, in: Proceedings of the ACM SIGPLAN 2007 Conference on Programming Language Design and Implementation, 2007, pp. 146–155.

[11] Y. Liu, Z. Hu, K. Matsuzaki, Towards systematic parallel programming over mapreduce, in: Proceedings of the 17th International Conference on Parallel Processing, 2011, pp. 39–50.

[12] A. L. Fisher, A. M. Ghuloum, Parallelizing complex scans and reductions, in: Proceedings of the ACM SIGPLAN 1994 Conference on Programming Language Design and Implementation (PLDI), 1994, pp. 135–146.

[13] R. C. O. Rocha, L. F. W. Góes, F. M. Q. Pereira, An algebraic framework for parallelizing recurrence in functional programming, in: SBLP, Springer, 2016, pp. 140–155.

[14] J. J. Rotman, Advanced Modern Algebra, 2nd Edition, Prentice Hall, 2003.

[15] J. S. Golan, Semirings and their Applications, 1st Edition, Springer, 1999.

[16] J. S. Golan, Power Algebras over Semirings: With Applications in Mathematics and Computer Science, 1st Edition, Mathematics and Its Applications 488, Springer, 1999.

[17] A. Morihata, K. Matsuzaki, Automatic parallelization of recursive functions using quantifier elimination, in: Proceedings of the 10th International Symposium on Functional and Logic Programming, 2010, pp. 321–336.

[18] S. J. Schlecht, E. A. P. Habets, Connections between parallel and serial combinations of comb filters and feedback delay networks, in: Proceedings of the 2012 International Workshop on Acoustic Signal Enhancement, 2012, pp. 1–4.

[19] S. P. Jones, A. Tolmach, T. Hoare, Playing by the rules: rewriting as a practical optimisation technique in ghc, in: Haskell workshop, Vol. 1, 2001, pp. 203–233.

[20] M. Steuwer, C. Fensch, S. Lindley, C. Dubach, Generating performance portable code using rewrite rules: From high-level functional expressions to high-performance opencl code, in: Proceedings of the 20th ACM SIGPLAN International Conference on Functional Programming, 2015, pp. 205–217.

[21] T. Remmelg, T. Lutz, M. Steuwer, C. Dubach, Performance portable gpu code generation for matrix multiplication, in: Proceedings of the 9th Annual Workshop on General Purpose Processing Using Graphics Processing Unit, 2016, pp. 22–31.

[22] M. Steuwer, T. Remmelg, C. Dubach, Lift: A functional data-parallel IR for high-performance GPU code generation, in: Proceedings of the 2017 International Symposium on Code Generation and Optimization, CGO '17, IEEE Press, Piscataway, NJ, USA, 2017, pp. 74–85.

[23] E. Balland, P. Brauner, R. Kopetz, P. Moreau, A. Reilles, Tom: Piggybacking rewriting on java, in: Proceedings of the 18th International Conference on Term Rewriting and Applications, 2007, pp. 36–47.

[24] J. Berthold, S. Marlow, K. Hammond, A. A. Zain, Comparing and optimising parallel haskell implementations for multicore machines, in: Proceedings of the International Conference on Parallel Processing Workshops, 2009, pp. 386–393.

[25] S. Marlow, P. Maier, H. Loidl, M. Aswad, P. W. Trinder, Seq no more: better strategies for parallel haskell, in: Proceedings of the 3rd ACM SIGPLAN Symposium on Haskell, 2010, pp. 91–102.

[26] P. Wadler, Deforestation: Transforming programs to eliminate trees, Theoretical Computer Science 73 (2) (1988) 231–248.

[27] A. J. Gill, S. L. P. Jones, Cheap deforestation in practice: An optimizer for haskell., in: Proceedings of the 13th IFIP World Computer Congress, 1994, pp. 581–586.

[28] S. Marlow, T. Harris, R. P. James, S. Peyton Jones, Parallel generational-copying garbage collection with a block-structured heap, in: Proceedings of the 7th ACM International Symposium on Memory Management, 2008, pp. 11–20.

[29] S. Marlow, S. Peyton Jones, Multicore garbage collection with local heaps, in: Proceedings of the 2011 ACM International Symposium on Memory Management, ACM, 2011, pp. 21–32.

[30] M. C. Chen, A parallel language and its compilation to multiprocessor machines or VLSI, in: Proceedings of the 13th Annual ACM Symposium on Principles of Programming Languages, 1986, pp. 131–139.

[31] N. Carriero, D. Gelernter, Linda in context, ACM Communications 32 (4) (1989) 444–458.

[32] J. Dean, S. Ghemawat, Mapreduce: Simplified data processing on large clusters, ACM Communications 51 (1) (2008) 107–113.

[33] L. L. P. Da Mata, F. M. Q. a. Pereira, R. Ferreira, Automatic parallelization of canonical loops, Science of Computer Programming 78 (8) (2013) 1193–1206.

[34] B. M. Ferreira, F. M. Q. Pereira, H. Rodrigues, B. S. Soares-Filho, Optimizing a geomodeling domain specific language, in: Proceedings of the 16th Brazilian Symposium on Programming Languages, 2012, pp. 87–101.

[35] B. M. Ferreira, B. S. Soares-Filho, F. M. Q. Pereira, The dinamica virtual machine for geosciences, in: Proceedings of the 19th Brazilian Symposium on Programming Languages, 2015, pp. 44–58.

[36] P. M. Kogge, H. S. Stone, A parallel algorithm for the efficient solution of a general class of recurrence equations, IEEE Transactions on Computers 22 (8) (1973) 786–793.

[37] S. Sato, H. Iwasaki, Automatic parallelization via matrix multiplication, in: Proceedings of the 32nd ACM SIGPLAN Conference on Programming Language Design and Implementation, 2011, pp. 470–479.

[38] Y. Zou, S. V. Rajopadhye, Scan detection and parallelization in "inherently sequential" nested loop programs, in: Proceedings of the 10th Annual IEEE/ACM International Symposium on Code Generation and Optimization, 2012, pp. 74–83.

[39] U. Bondhugula, M. M. Baskaran, S. Krishnamoorthy, J. Ramanujam, A. Rountev, P. Sadayappan, Automatic transformations for communication-minimized parallelization and locality optimization in the polyhedral model, in: Proceedings of the 17th International Conference on Compiler Construction, 2008, pp. 132–146.

[40] U. Bondhugula, A. Hartono, J. Ramanujam, P. Sadayappan, A practical automatic polyhedral parallelizer and locality optimizer, in: Proceedings of the ACM SIGPLAN 2008 Conference on Programming Language Design and Implementation, 2008, pp. 101–113.

[41] K. Trifunovic, D. Nuzman, A. Cohen, A. Zaks, I. Rosen, Polyhedral-model guided loop-nest auto-vectorization, in: Proceedings of the 18th International Conference on Parallel Architectures and Compilation Techniques, 2009, pp. 327–337.

[42] P. Feautrier, Automatic parallelization in the polytope model, in: The Data Parallel Programming Model: Foundations, HPF Realization, and Scientific Applications, 1996, pp. 79–103.

[43] A. Morihata, K. Matsuzaki, Z. Hu, M. Takeichi, The third homomorphism theorem on trees: downward & upward lead to divide-and-conquer, in: Proceedings of the 36th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, 2009, pp. 177–185.