

# Exploring the uses of Quantitative Types

## Abstract

Idris2 is a programming language featuring Quantitative type theory[1], a Type Theory centered around tracking *usage quantities* in addition to dependent types. This is the result of more than 30 years of development spawned by the work of Girard on Linear logic[2]. Until Idris2, our understanding of linear types and their uses in dependently-typed programs was hypothetical. However this changes with languages like Idris2, which allow us to explore the semantics of running programs using linear and dependent types. In this thesis I explore multiple facets of programming through the lens of quantitative programming, from ergonomics, to performance. I will present how quantitative annotations can help the programmer write program that precisely match their intension, help the compiler better analyse the program written, and help the output bytecode to run faster.

# Contents

<b>Abstract</b>	<b>1</b>
<b>1 Introduction</b>	<b>5</b>
Some vocabulary and jargon . . . . .	5
Programming recap . . . . .	6
Idris and dependent types . . . . .	7
Simple example . . . . .	8
Dependent types example . . . . .	9
Holes in Idris . . . . .	10
What's wrong with non-dependent types? . . . . .	12
The story begins . . . . .	14
<b>2 Context Review</b>	<b>15</b>
Origins . . . . .	15
Applications . . . . .	16
Cutting edge linear types . . . . .	18
Counting immutable beans . . . . .	19
Innovative uses of linear types . . . . .	20
Opening the door to new opportunities . . . . .	20
More granular dependencies . . . . .	22
Invertible functions . . . . .	23
<b>3 Idris2 and linear types</b>	<b>24</b>
Incremental steps . . . . .	25
Dropping and picking it up again . . . . .	26
Dancing around linear types . . . . .	28
Linear intuition . . . . .	30
You can't have your cake and eat it too . . . . .	30
Drawing unexpected parallels . . . . .	31
Safe inlining with 1 . . . . .	33
Erased runtime for $\emptyset$ . . . . .	34
No branching with 0 . . . . .	37
<b>4 Quantitative Type Theory in practice</b>	<b>39</b>
limitations and solutions for quantitative types . . . . .	39
Linear multiplication . . . . .	39
Permutations . . . . .	41
Levitation improvements . . . . .	46

Compile-time string concatenation . . . . .	47
<b>5 Quantitative Type Theory and Programming ergonomics</b>	<b>51</b>
Mapping primitive to data types and vice-versa . . . . .	52
Idris2 and multiplicities . . . . .	54
The problem with runtime garbage collection (WIP) . . . . .	54
What is reference counting (WIP) . . . . .	55
Alternative semirings and their semantics (WIP) . . . . .	55
Putting everything together: Linearity and reference counting (WIP) . . . . .	55
Steps toward a working implementation . . . . .	55
Next Steps . . . . .	58
Relaxing linearity inference (WIP) . . . . .	59
<b>6 Performance Improvement using linear types</b>	<b>62</b>
Linear function subtyping . . . . .	63
The nature of the changes required for linear optimisations . . . . .	65
Implementation details . . . . .	67
Mutating branches . . . . .	68
Reference nightmare . . . . .	69
Benchmarks & methodology . . . . .	73
Traditional Fibonacci . . . . .	74
Allocating Fibonacci . . . . .	75
Mutating Fibonacci . . . . .	75
Mapping lists . . . . .	76
Allocation free mapping . . . . .	76
A simple SAT solver as benchmark . . . . .	77
Measurements . . . . .	77
Idris-bench . . . . .	78
Idris-stats . . . . .	78
Expected results . . . . .	79
Running the benchmarks . . . . .	80
Results . . . . .	80
Results 1: Fibonacci . . . . .	81
Results 2: Fibonacci without startup time . . . . .	84
Results 3: Fibonacci without startup time, small loop . . . . .	86
Results 4: mapping . . . . .	87
Sat solver . . . . .	88

<b>7</b>	<b>Future work</b>	<b>89</b>
	Enlarging the scope . . . . .	89
	Making linearity easier to use . . . . .	90
	Not linear enough . . . . .	90
	Too linear now . . . . .	91
<b>8</b>	<b>Conclusion</b>	<b>94</b>
<b>9</b>	<b>Appendices: Glossary and definitions</b>	<b>96</b>
	Linearity / Quantity / Multiplicity . . . . .	96
	Linear types . . . . .	96
	Affine types . . . . .	96
	Monad . . . . .	96
	Co-monad / Comonad . . . . .	96
	Semiring . . . . .	96
	Lattice . . . . .	97
	Syntax . . . . .	97
	Semantics . . . . .	97
	Pattern matching . . . . .	98
	Implicit argument . . . . .	98
	Term/Expression/Value . . . . .	98
<b>10</b>	<b>Appendix</b>	<b>102</b>
	Sat solver benchmark input . . . . .	102
	Sat solver benchmark output . . . . .	104

# 1 Introduction

In this project I will demonstrate different uses and results stemming from a programming practice that allows us to specify how many times each variable is being used. This information is part of the type system, in our case we are tracking if a variable is allowed to be used exactly once, or if it has no restrictions. Such types system is called a “linear type system”.

As we will see there is a lot more to this story, so as part of this thesis, I will spend some time introducing dependent types as well as our host language: Idris. This warm up will be followed by a context review outlining the existing body of work about linear types and related topics. Once both the theoretical and technical landscape have been set I will introduce the concept of linear and quantitative type in order to. Which is the final piece to understand the two main topics of this thesis, the ergonomics of Quantitative Type Theory[1][3] (QTT, for short) for software development and the performance improvement we can derive from the linearity features of Idris2.

Those two chapters reflect the main goal of this thesis which is to explore the uses of quantitative types in a dependently-types setting, for real-world programs. This exploration aims to explain their limitations and propose ways of getting around them. Some of those ideas have been implemented in the Idris2 compiler and we discuss their details and consequences.

Let us begin slowly and introduce the basic concepts. The following will only make assumptions about basic understanding of imperative and functional programming.

## Some vocabulary and jargon

Technical papers are often hard to approach for the uninitiated because of their heavy use of unfamiliar vocabulary and domain-specific jargon. While jargon is useful for referencing complicated ideas succinctly, it is a double edged sword as it also tends to hinder learning by obscuring important concepts. Unfortunately, I do not have a solution for this problem, but I hope this section will help mitigate this feeling of helplessness when sentences seem to be composed of randomly generated sequences of letters rather than legitimate words.

You will find more complete definitions at the end, in the glossary.

**Type** A label associated to a collection of values. For example `String` is the type given to strings of characters for text. `Int` is the type given to integer values. `Nat` is the type given to natural numbers.

**Type-System** Set of rules the types in a program have to follow in order to be accepted by the compiler. The goal of the type-system is to catch some classes of errors while helping the programmer reach their goal more easily.

**Linear types** Types that have a usage restriction. Typically, a value labelled with a linear type can only be used once, no less, no more.

**Linearity / Quantity / Multiplicity** Used interchangeably most of the time. They refer to the number of times a variable is expected to be used.

**Syntax** The structure of some piece of information, usual in the form of *text*. Syntax itself does not convey any meaning.

**Semantics** The meaning associated to a piece of data, most often related to syntax.

**Pattern matching** Destructuring a value into its constituent parts in order to access them or understand what kind of value we are dealing with.

**Generic type / Polymorphic type / Type parameter** A type that needs a concrete type in order to be complete. For example `Maybe a` takes the single type `a` as a parameter. Once we pass it the type `Int` it becomes the complete type `Maybe Int`.

## Programming recap

If you know about programming, you’ve probably heard about types and functions. Types are ways to classify values that the computer manipulates and functions are instructions that describe how those values are changed.

In *imperative programming* functions can perform powerful operations like “malloc” and “free” for memory management or make network requests through the internet. While powerful in a practical sense, those functions are really hard to study because they are hard to define in a mathematical way. In order to make life easier we only consider functions in the *mathematical* sense of the word : A function is something that takes an input and returns an output.

---

`f : A -> B`

---

This notation tells us what type the function is ready to ingest as input and what type is expected as the output.

---

input	output
$v$	$v$
<code>f : A -&gt; B</code>	
$\wedge$	
name	

---

This simplifies our model because it forbids the complexity related to complex operations like arbitrary memory modification or network access <sup>1</sup>. Functional programming describes a programming practice centered around the use of such functions. In addition, traditional functional programming languages have a strong emphasis on their type system which allows the types to describe the structure of the values very precisely.

During the rest of this thesis we are going to talk about Idris2, a purely functional programming language featuring Quantitative Type Theory (QTT), a type theory<sup>2</sup> based around managing resources. But before we talk about QTT itself, we have to explain what is Idris and what are dependent types.

## Idris and dependent types

Before we jump into Idris2, allow me to introduce Idris[4], its predecessor. Idris is a programming language featuring dependent types.

A way to understand dependent types is to think of the programming language as having *first class types*, that is, types are values, just like any other, in the language. Types being normal values means we can create them, pass them as arguments to functions and return them to functions. This also means the difference between “type-level” and “term-level” is blurred.

---

<sup>1</sup>We can recover those features by using patterns like “monad” but it is not the topic of this brief introduction

<sup>2</sup>*Type theory* is the abstract study of type systems, most often in the context of pure, mathematical, logic. When we say “a Type Theory” we mean a specific set of logical rules that can be implemented into a *Type System*.

## Simple example

Let us start with a very simple example of a function in Idris, without dependent types. This function returns whether or not a list <sup>3</sup> is empty:

---

```
isEmpty : List a -> Bool
isEmpty [] = True
isEmpty (x :: xs) = False
```

---

Every Idris function is comprised of two parts, the *Type signature* and the *implementation* (or body) of the function. Let us start by dissecting the Type signature:

---

```
--      | Name of the function
--      |
--      |      | Type of the argument
--      |      |
--      |      |      | Return type
--      |      |      |
isEmpty : List a -> Bool
--      |
--      |      |
--      |      | Type parameter
```

---

Every signature starts with a name, followed by a colon `:` and ends with a type. In this case, the type is `List a -> Bool` which represents a function that takes a list of `a` and returns a `Bool`. Interestingly, the type `a` is a *Type parameter* and could be anything, this function will work for all lists, irrespective of their content.

As for the body, here is how it is composed:

---

<sup>3</sup>A list is defined as

“data List a = Nil | Cons a (List a)

Which mean a list is either empty (`Nil`) or non-empty (`Cons`) and will contain an element of type `a` and a reference to the tail of the list, which itself might be empty or not. It is customary to write the `Cons` case as `::`, and in fact the official implementation uses the symbol `::` instead of the word `Cons`. It is also customary to represent the empty list by `[]`.



```

--      Recall the function name
--      |
--      |
--      | Pattern matching on the empty list
--      |
--      | Return value
--      |
isEmpty [] = True
--      |
--      | Pattern matching on the non-empty list
--      |
isEmpty (x :: xs) = False
--      |
--      | Return value
--      |
--      | Tail of the list
--      |
--      | Head of the list

```

---

For the body we need to recall the name of the function for every pattern match we make. In our case we only have two cases, the empty list `[]` and the non-empty list `x :: xs`. When we match on a value we may discover that we can access new variables and we give them names. In this case the `::` case allows us to bind the head and the tail of the list to the names `x` and `xs`. Depending on whether we are dealing with an empty list or not we return `True` or `False` to indicate that the list is empty or not.

Pattern matching is particularly important in Idris because it allows the compiler to better understand how the types flow through our program. We are going to see an example of how dependent pattern matching manifests itself when we introduce *type holes*.

### Dependent types example

Now let us look at an example with dependent types.

---

```

intOrString : (b : Bool) -> if b then Int else String
intOrString True = 404
intOrString False = "we got a string"

```

---

This short snippet already shows a lot of things, but again the most important part is the type signature. Let us inspect it:

---



compiler will also report what it knows about the surrounding context to give additional insight.

If we take our example of `intOrString` and replace the implementation by a hole we have the following:

---

```
intOrString : (b : Bool) -> if b then Int else String
intOrString b = ?hole
```

---

We can then ask the compiler what is the expected type, information provided by the compiler will be shown with a column of `>` on the left side to distinguish it from code. When the compiler tells us about holes, its will display the variables available in the immediate scope of the hole above a horizontal bar, and the “goal” under the horizontal bar. Here is the type we get when asking about our hole:

---

```
> b : Bool
> -----
> hole : if b then Int else String
```

---

This information does not tell us what value we can use. However it informs us that the type of the value *depends on the value of `b`*. Therefore, pattern matching on `b` might give us more insight. It is worth noting that Idris has an interactive mode that allows the pattern matching to be done automatically by hitting a simple key stroke which will generate the code in the snippet automatically. This won’t be showcased here, as it is not an Idris development tutorial, but this approach has already been put to use and proven effective in other programming languages with proof-assistant features[5][6]. Back to the example:

---

```
intOrString : (b : Bool) -> if b then Int else String
intOrString True = ?hole1
intOrString False = ?hole2
```

---

Asking again what is in `hole1` gets us

---

```
> hole1 : Int
```

---

and `hole2` gets us

---

```
> hole2 : String
```

---

Which we can fill with literal values like 123 or "good afternoon". The complete program would look like this:

---

```
intOrString : (b : Bool) -> if b then Int else String
intOrString True = 123
intOrString False = "good afternoon"
```

---

### What's wrong with non-dependent types?

Non-dependent type systems cannot represent the behaviour described by `intOrString` and have to resort to patterns like `Either`<sup>5</sup> to encapsulate the two possible cases our program can encounter.

---

```
eitherIntOrString :: Bool -> Either Int String
eitherIntOrString True = Left 404
eitherIntOrString False = Right "we got a string"
```

---

While this is fine in principle, it comes with a set of drawbacks that cannot be solved without dependent types. In order to see this, let us place a hole in the implementation of `eitherIntOrString`:

---

```
eitherIntOrString : Bool -> Either Int String
eitherIntOrString b = ?hole
```

---

---

```
> b : Bool
> -----
> hole : Either Int String
```

---

While this type might be easier to read than `if b then Int else String` it does not tell us how to proceed in order to find a more precise type to fill. We can try pattern matching on `b`:

---

```
intOrString' : Bool -> Either Int String
intOrString' True = ?hole1
intOrString' False = ?hole2
```

---

But it does not provide any additional information about the return types to use.

---

<sup>5</sup>Defined as : “data Either a b = Left a | Right b

```
> -----  
> hole1 : Either Int String
```

---

```
> -----  
> hole2 : Either Int String
```

---

In itself using `Either` isn't a problem, however `Either`'s lack of information manifests itself in other ways during programming; take the following program:

```
checkType : Int  
checkType = let intValue = eitherIntOrString True  
            in ?hole
```

---

```
> intValue : Either Int String  
> -----  
> hole : Int
```

---

In this program we create the variable `intValue` by assigning it to the result of the function call to `eitherIntOrString` with the argument `True`. When we ask for the type of `Hole` we see that we are expected to return an `Int`. The compiler is unable to tell us if `intValue` contains an `Int` or a `String`. Despite us *knowing* that `IntOrString` returns an `Int` when passed `True`, we cannot use this fact to convince the compiler to simplify the type for us. We have to go through a runtime check to ensure that the value we are inspecting is indeed an `Int`:

```
checkType : Int  
checkType = let intValue = eitherIntOrString True in  
            case intValue of  
              (Left i) => ?hole1  
              (Right str) => ?hole2
```

---

But doing so introduces the additional problem that we now need to provide a value for an impossible case (`Right`). What do we even return? We do not have an `Int` at hand to use. Our alternatives are:

- Panic and crash the program.
- Make up a default value, silencing the error but hiding a potential bug.
- Change the return type to `Either Int String` and letting the caller deal with it.

None of which are ideal nor replicate the functionality of the dependent version

we saw before. This is why dependent types are desirable, they help:

- Avoid needless runtime checks.
- The compiler better understand the semantics of the program.
- Inform the programmer by communicating precisely which types are expected.

This concludes our short introduction to dependent types and I hope you've been convinced of their usefulness. In the next section we are going to talk about linear types.

## **The story begins**

This concludes our introductory chapter, we've laid out the ground work for what's to come with Idris and linear types. Dependent types are a thesis topic in themselves but this introduction should be enough to give you a general idea. In the next section I am going to present the existing literature surrounding the topic of this thesis: linear and quantitative types. Those research documents should steer us toward the goal of this thesis, exploring the uses of quantitative types for real-world programs, explaining their limitations and proposing ways of getting around them, and designing and implementing use cases for them within the Idris2 compiler.

I also want to stress that at the end is a glossary that lists all the important terms and concepts necessary to understand the body of this work. Please feel free to consult it if something is unclear.

## 2 Context Review

In this context review, I will comment on the existing literature about linear types and present it through three lenses: The first aims to tell the origin story of linear types and their youthful promises. The second will focus on the current understanding of their application for real-world use. And the last one will focus on the latest theoretical developments that linear types spun up.

### Origins

Linear types were first introduced by J-Y. Girard in his 1987[2] publication simply named *Linear logic*. In this text he introduces the idea of restricting the application of the weakening rule and contraction rule from intuitionistic logic in order to allow a statement to be managed as *resources*. Linear terms once used cannot be referred again, premises cannot be duplicated and contexts cannot be extended. This restriction was informed by the necessity real-world computational restriction, in particular accessing information concurrently.

One of the pain points mentioned was the inability to restrict usages to something more sophisticated than “used exactly once”. Linear variables could be promoted to their unrestricted variants with the exponential operator (!) but it removes any benefit we get from linearity. A limitation that will be revisited in the follow-up paper: Bounded linear logic.

It is worth noting that, already at this stage, memory implication were considered. Typically the exponential operator was understood as being similar to “long term storage” of a variable such that it could be reused in the future, whereas linear variables were stored in short-term memory to be reclaimed, like a register or a stack.

The follow up paper, *Bounded linear logic*[7] improves the expressivity of linear logic while keeping its benefits: intuitionistic-compatible logic that is computationally relevant. The key difference with linear logic is that weakening rules are *bounded* by a finite value such that each value can be used as many times as the bound allows. In addition, some typing rules might allow it to *waste* resources by *underusing* the variable, hinting that affine types might bring some concrete benefits to our programming model.

As before, there is no practical application of this in terms of programming language. However this brings up the first step toward managing *quantities* in the language. An idea that will be explored again later with Granule[8][9][10][11]

and Quantitative Type Theory.

## Applications

Soon after the development of linear types, they appeared in a paper aimed at optimising away redundant allocations when manipulating lists: The deforestation algorithm.

Deforestation[12] is an algorithm designed to avoid extraneous allocation when performing list operations in a programming language close to System-F. The assumption that operations on lists must be linear was made to avoid duplicating operations. If a program was non-linear, the optimisation would duplicate each of the computation associated with the non-linear variable, making the resulting program less efficient.

While deforestation itself might not be the algorithm that we want to implement today, it is likely we can come up with a similar set of optimisation rules in Idris2 that rely on linearity.

In this case linearity avoid duplicating computation, this idea was again investigated in *Once upon a Type*[13] which formalises the detection of linear variables and uses this information for safe inlining. Indeed arbitrarily inlining functions might result in duplicated computation. We’ve seen an example of this earlier:

---

```
let x = f y in
(x, x)
```

---

Where inlining into  $(f\ x, f\ x)$  duplicates the work done by  $f$ .

Beside inlining and mutation, another way to use linear types for performance is memory space mutation. *Linear types can change the world*[14] show that Linear types can be used for in-place update and mutation instead of relying on copying. They provide a safe API to update values by relying on the linearity of the argument being used.

However the weakness of this result is that the API exposed to the programmer relies on nested continuations when traversing nested data types. Which is largely seen as undesirable user experience. *Is there a use for linear logic*[15] makes similar claims and introduces a concept of types with a *unique* pointer to them. This restriction for unique pointers is something we are going to encounter again during this thesis.



The memory management benefits from linear types do not stop here, in *Reference counting as a computational interpretation of linear logic* [16], linear types are shown to correspond to  $\lambda_{\text{RC}}$ , this paper shows that a simple calculus augmented with memory management primitives can make use of linearity in order to control memory allocation and deallocation using linear types. While the calculus itself isn't very ergonomic to use for programming, one could imagine using it as a backend for a reference counted runtime.

*Practical affine types* [17] aims to answer “What does it mean to have access to linear and affine types *in practice*”? Indeed, most the results we've talked about develop a theory for linear types using a variant of linear logic, and then present a toy language to showcase their contribution. However this does not teach us how they would interact and manifest in existing programs or in existing software engineering workflows. Do we see emerging new programming patterns? Is the user experience improved or diminished? In what regards is the code different to read and write? How is code organised and shared? All those questions can only be answered by a fully fledged implementation of a programming language equipped to interact with existing systems.

Practical affine types show that their implementation for linear+affine types allow to express common operations in concurrent programs without any risk of data races. They note that typical stateful protocols should also be implementable since their language is a strict superset of other which already provided protocol implementations. Those two results hint at us that linear types in a commercially-relevant programming language would provide us with additional guarantees without impeding on the existing writing or reading experience of programs.

Haskell already benefits from a plethora of big and small extensions, they are so powerful that combining them results in some extremely useful, though quite unergonomic, implementation of dependent types [18]. Linear Haskell [19] is notable in that it extends the type system to allow linear functions to be defined. It introduces the linear arrow `->#` which declares a function to be linear. Because of Haskell's laziness, linearity doesn't mean “will be used exactly once” but rather “*if it is used*, then it will be used exactly once”.

This addition to the language was motivated by a concern for safe APIs, typically when dealing with unsafe or low-level code. Linear types allow to expose an API that cannot be misused while keeping the same level of expressivity and being completely backwards compatible. This backward compatibility is in part

allowed thanks to parametric linearity, the ability to abstract over linearity annotations. And in part due to the approach of making *function arrows* linear instead of *types* themselves.

Linearity on the type has the benefit that values themselves can be marked linear, which would enforce their uniqueness. But often comes at the cost of splitting our context in two[20] between linear values and non-linear values which makes is unsuitable for a dependent type theory in which both types and terms live in the same context.

## Cutting edge linear types

Granule is a language that features *quantitative reasoning via graded modal types*. They even have indexed types to boot! This effort is the result of years of research in the domain of effect, co-effect, resource-aware calculus and co-monadic computation. Granule itself makes heavy use of *graded monads* (Katsuma, Orchard et al. 2016) which allow to precisely annotate co-effects in the type system. This enables the program to model *resource management* at the type-level. What's more, graded monads provide an algebraic structure to *combine and compose* those co-effects. This way, linearity can not only be modelled but also *mixed-in* with other interpretations of resources. While this multiplies the opportunities in resource tracking, this approach hasn't received the treatment it deserves with regards to performance and tracking runtime complexity.

Up until now we have not addressed the main requirement of our programming language: We intend to use *both* dependent types *and* linear types within the same language. However, no satisfying theory came until\_I got plentty o nuttin\_[3] and its descendant, *Quantitative type theory* [1], came to fill the gap. Previous attempt by noted that linear types should be unrestricted in intuitionistic declarations but kept the original design to split linear and intuitionistic contexts. So in order to allow full dependent types, two changes were made: - Dependent typing can only use *erased* variables - Multiplicities are tracked on the *binder* <sup>6</sup>rather than being a feature of type of a variable. While this elegantly merges the capabilities of a Martin-Löf-style type theory[21] and Linear Logic[2], the proposed result does not touch upon potential performance improvement that such a language could feature[22]. However it has the potential to bring together Linear types and dependent types in a way that allows precise resource tracking and strong performance guarantees. Resource tracking

---

<sup>6</sup>A Binder associates a value to a name. `let n = 3 in ...` binds the value 3 to the name n. Named arguments are also *binders*.

through semirings is not a novel idea, but has been explored through the lens of and explicit comonadic computation [23][24][9][25][26] instead of dependent linear type systems.

Unfortunately, this approach does not solve every long standing problems of computer science like having the univalence axioms be a theorem in the language like cubical agda[27] does. Interestingly enough, cubical type theory and homotopy type theory display a problem that could be solved in Idris2 very easily: They are non-computational. Which means they simply could be restricted to *erased* multiplicity such that they are enforce by the compiler to never appear at runtime.

Since this approach has already been tried in Agda[28], maybe Idris2 should be extended with Observational Type Theory [29] in order to also solve the problem of functional extensionality. Which would greatly improve the range of programs Idris2 can write as we see with Idris-CT[30] which require to postulate function extensionality for a lot of their proofs.

### Counting immutable beans

As we’ve seen, linearity has strong ties with resource and memory management, including reference counting. Though *Counting immutable beans* does not concern itself with linearity per se, it mentions the benefits of *reference counting* as a memory management strategy for purely functional programs. Indeed, while reference counting has, for a long time, been disregarded in favor of runtime garbage collectors, it now has proven to be commercially viable in languages like Swift or Python. The specific results presented here are focused on the performance benefits in avoiding unnecessary copies and reducing the amount of increment and decrement operation when manipulating the reference count at runtime. It turns out the concept of “borrowing” a value without changing its reference count closely matches a linear type system with graded modalities. Indeed as long as the modality is finite and greater than 1 there is no need to decrement the reference count. Here is an illustration of this idea

---

```
f : (2 v : Int) -> Int
f v = let 1 val1 = operation v -- operation borrow v, no need for dec
      1 val2 = operation v -- we ran out of ses for v, dec here
      in val1 + val2
```

---

In our example, since `v` could be shared prior to the calling of `f` we cannot prove that `v` can be freed, we can only decrement its reference count. However, by

inspecting the reference count we could in addition reuse our assumption about “mutating unique linear variables” and either reclaim the space or reuse it in-place.

## Innovative uses of linear types

In this final section of the context review I will present some interesting uses for linear types in a bit more detail.

### Opening the door to new opportunities

Protocol descriptions and dependent types work marvellously well. State machines can be represented by data types and their index can ensure we compose them in ways that make sense.

As a reminder of how state machines can be encoded in Idris1, here is a simple door protocol in Idris 1 using *indexed monads* [31] where the indices behave like a Hoare Triple. An approach that’s been put to practice in *State machines all the way down*[32] and gives rise to our example:

---

```
data DoorState = Open | Closed

data Door : Type -> DoorState -> DoorState -> Type where
  Open : Door () Closed Open
  Close : Door () Open Closed
  Play : Door () Open Opened
  Pure : ty -> Door ty state state
  (>=>) : Door a state1 state2 ->
    (a -> Door b state2 state3) ->
    Door b state1 state3
```

---

This allows the door protocol to be expressed with a function with the following type:

---

```
doorProtocol : Door () Closed Closed
```

---

This indicates that the door starts closed and ends closed. One issue with this approach is that it constrains us to a free-monadic style implementation where we need to write an interpreter for our monadic program and we cannot mix other protocols within it without changing the `Door` data type.

Implementing the same protocol using linear type is much easier:

---

```

data DoorState = Open | Closed

data Door : (state : DoorState) -> Type where
  MkDoor : Door Closed
  Open   : (1 _ : Door Closed) -> Door Open
  Close  : (1 _ : Door Open) -> Door Closed
  Play   : (1 _ : Door Open) -> Door Open

operateDoor : (op : (1 _ : Door Closed) -> Door Closed)
              -> Door Closed
operateDoor op = op MkDoor

```

---

In this example, instead of monadic composition we use plain *function composition* in order to ensure that our protocol obeys the rules of our protocol, that is: The door starts open and ends closed.

This verifies the protocol because it forces the `op` function to make use of the provided door. The client of the API has no choice but to manipulate the existing door until the protocol is over.

---

```

-- doesn't typecheck because door is ignored.
operateDoor (\door => MkDoor)

```

---

We can now define program simply by combining our door operations:

---

```

infixr 1 &>

-- Sequential linear function composition.
(&>) : ((1 _ : a) -> b) ->
      ((1 _ : b) -> c) ->
      (1 _ : a) -> c
(&>) f g x = g (f x)

connectDoor : (1 _ : Door Closed) -> Door Closed
connectDoor = Play &> Close &> Open &> Play &> Play &> Close

```

---

This also makes combining other protocols trivial, assume we have another protocol to connect to the internet and is characterised by the signature `connect : (1 _ : Socket Open) -> Socked Closed` we can take the product of the two connect functions and interleave their arguments:

---

```

connectProd : (1 _ : (Socket Open, Door Open))
             -> (Socket Closed, Door Closed)
connectProd (socket, door) =
  let openDoor = Open door
      socket' = send "message" socket
      door' = Enjoy openDoor in
    (close socket', Close door')

```

---

Which is much more natural than having to rewrite our door protocol and rewrite our interpreter for the operations of the protocol.

### More granular dependencies

While our previous example has only be mildly successful, there exist a language that can express our idea, and that is *Granule*.

Granule is a programming language with *graded modal types*, types which rely on *graded modalities*. Those are annotation that span a *range* of different values in order to describe each type. Those values can themselves be paired up together and combined in order to represent even more complex behaviour than our linear multiplication. For now, let us stick to multiplication and see what a future version of Idris supporting graded modalities could look like.

Granule's syntax is very close to Idris, Agda and Haskell, however, linearity in *Granule* is the default so there is nothing to specify for a linear variable. In addition, `Nat` is not a `Type` in *Granule* but a *modality*, which means, in order to work with `Nat` and write dependencies between them we will create a data type indexed on `Nat`:

```

data INat (n : Nat) where
  Z : INat 0;
  S : INat n -> INat (n + 1)

```

---

This allows us to write the add function as follows:

```

linearAdd : forall {n m : Nat} . INat n -> INat m -> INat (n + m)
linearAdd Z m = m;
linearAdd (S n) m = S (linearAdd n m)

```

---

If we were to omit the `m` in the first branch and write `linearAdd Z m = Z` we would get the error:

```
> Linearity error: multiplication.gr:
> Linear variable `m` is never used.
```

---

Which is what we expect.

Now that we have indexed `Nat` we can try again our multiplication function:

```
multiplication : forall {n m : Nat} . INat n -> (INat m) [n] -> INat (n * m)
multiplication Z [m] = Z;
multiplication (S n) [m] = linearAdd m (multiplication n m)
```

---

As you can see, we annotate the second argument of the type signature with `[n]` which indicates that the modality of the second argument depends on the value of the first argument. This syntax repeats in the implementation where the second argument `m` has to be “unboxed” using the `[m]` syntax which will tell the compiler to correctly infer the usage allowed by the indexed modality. In the first branch there are 0 uses available, and in the second there are  $n + 1$  uses available.

While *Granule* doesn’t have dependent types, indexed types are enough to implement interesting programs such as multiplication. More recent developments have made progress toward implementing full type dependency between quantities and terms in the language.

## Invertible functions

Yet another use of linearity that I have not seen mentioned very often is to use them for *invertible functions*. That is functions that have a counterpart that can undo their actions. Such functions are extremely common in practice but aren’t usually described in terms of their ability to be undone. Here are a couple example

- Addition and subtraction
- `::` and `tail`
- serialisation/deserialisation

The paper about Sparcl[33] goes into details about how to implement a language that features invertible functions, they introduce a new (postscript) type constructor `• : Type -> Type` that indicate that the type in argument is invertible. Invertible functions are declared as linear functions  $A \bullet \multimap B$ <sup>7</sup>. Invertible

---

<sup>7</sup>The lollipop arrow  $a \multimap b$  is equivalent to our linear functions  $(1 \_ : a) \rightarrow b$

functions can be called to make progress one way or the other given some data using the `fwd` and `bwd` primitives:

---

```
fwd : (A• -o B•) -> A -> B
bwd : (A• -o B•) -> B -> A
```

---

Invertible functions aren't necessarily total, For example `bwd (+ 1) Z` will result in a runtime error. This is because of the nature of invertible functions: the `+ 1` functions effectively adds a `S` layer to the given data. In order to undo this operation we need to *peel off* a `S` from the data. But `Z` doesn't have a `S` constructor surrounding it, resulting in an error.

Those type of runtime errors can be avoided in Idris by adding a new implicit predicate that ensure the data is of the correct format:

---

```
bwd : (f : (1 _ : A•) -> B•) -> (v : B) -> {prf : v = fwd f x} -> A
```

---

This ensures that we only take values of `B` that come from a `fwd` operation, that is, it only accepts data that has been correctly build instead of abitrary data. If we were to translate this into our `nat` example it would look like this:

---

```
undo+1 : (n : Nat) -> {prf : n = S k} -> Nat
```

---

Which ensures that the argument is a `S` of `k` for any `k`.

### 3 Idris2 and linear types

We've seen in the introduction how Idris features dependent types and some basic demonstrations about how to use them. Idris2 takes things further and introduces *linear types* in its type system, allowing us to define how many times a variable will be used. Three different quantities exist in Idris2 : `0`, `1` and `ω`. `0` means the value cannot be used in the body of a function, `1` means it has to be used exactly once, no less, no more. `ω` means the variable isn't subject to any usage restrictions, just like other (non-linear) programming languages. In Idris2, `ω` is implicit, when no quantity is specified and left blank, it is assumed to be unrestricted. This is why we have not seen any usage quantities during the introduction.

We are going to revisit every usage quantities later as there are more subtleties, especially about the `0` usage. For now we are going to explore some examples



of linear functions and linear types, starting with very simple functions such as incrementing numbers.

## Incremental steps

When designing examples, natural numbers are a straightforward data type to turn to. Their definition is extremely simple: `data Nat = Z | S Nat` which means that a `Nat` is either zero `Z` or the successor of another natural number, `S`.

Given this definition let us look at the function that increments a natural number:

---

```
increment : Nat -> Nat
increment n = S n
```

---

As we've seen before with our `intOrString` function we can name our arguments in order to refer to them later in the type signature. We can do the same here even if we do not use the argument in a dependent type. Here we are going to name our first argument `n`.

---

```
increment : (n : Nat) -> Nat
increment n = S n
```

---

In this case, the name `n` doesn't serve any other purpose than documentation, but our implementation of linear types has one particularity: quantities have to be assigned to a *name* (something we will explore in further details in a later section). Since the argument of `increment` is used exactly once in the body of the function we can update our type signature to assign the quantity 1 to the argument `n`:

---

```
--           ⌈ We declare `n` to be linear
--           ▼
increment : (1 n : Nat) -> Nat
increment n = S n
--           ▲
--           |
--           ⌋ We use n once here
```

---

The compiler will now check that `n` is used exactly once.

Additionally, Idris2 features pattern matching and the rules of linearity also apply to each variable that was bound when matching on it. That is, if the value

we are matching is linear then we need to use the pattern variables linearly.

```
sum : (1 n : Nat) -> (1 m : Nat) -> Nat
sum Z m = m
-- ▲
-- | We match on the argument here
sum (S n) m = S (sum n m)
-- └───┐
--      ▲
--      | We use `n`
--      |
--      | We match on the argument and bind `n`
```

In this last example we match on `S` and bind the argument of `S` (the predecessor) to the variable `n`. Since the original argument was linear, `n` is linear too and is indeed used later with `sum n m`.

## Dropping and picking it up again

Obviously this programming discipline does not allow us to express every program the same way as before. Here are two typical examples that cannot be expressed :

```
drop : (1 v : a) -> ()  
  
copy : (1 v : a) -> (a, a)
```

Here, the first function aims to ignore the argument and the second duplicates its argument and packages it in a pair.

We can explore what is wrong with those functions by trying to implement them and making use of holes.

```
drop : (1 v : a) -> ()
drop v = ?drop_rhs
```

```
> 0 a : Type
> 1 v : a
> -----
> drop_rhs : ()
```

As you can see, each variable is annotated with an additional number on its left, **0** or **1**, that informs us how many times each variable has to be used (If there

is no restriction, the usage number is simply left blank, just like our previous examples didn't show any usage numbers).

As you can see we need to use `v` (since it is marked with `1`) but we are only allowed to return `()`. This would be solved if we had a function of type `(1 v : a) -> ()` to consume the value and return `()`, but this is exactly the signature of the function we are trying to implement!

If we try to implement the function by returning `()` directly we get the following:

---

```
drop : (1 v : a) -> ()
drop v = ()
```

---

---

```
> There are 0 uses of linear variable v
```

---

Which indicates that `v` is supposed to be used but no uses have been found.

Similarly for `copy` we have:

---

```
copy : (1 v : a) -> (a, a)
copy v = ?hole
```

---

---

```
> 0 a : Type
> 1 v : a
> -----
> hole : (a, a)
```

---

In which we need to use `v` twice but we're only allow to use it once. Using it twice result in this program with this error:

---

```
copy : (1 v : a) -> (a, a)
copy v = (v, v)
```

---

---

```
> There are 2 uses of linear variable v
```

---

Interestingly enough, partially implementing our program with a hole gives us an amazing insight:

---

```
copy : (1 v : a) -> (a, a)
copy v = (v, ?hole)
```

---

---

```
> 0 a : Type
> 0 v : a
> -----
> hole : a
```

---

The hole has been updated to reflect the fact that though `v` is in scope, no uses of it are available. Despite that we still need to make up a value of type `a` out of thin air, which is impossible <sup>8</sup>.

While there are experimental ideas that allow us to recover those capabilities they are not currently present in Idris2. We will talk about those limitations and how to overcome them in the “limitations and solutions for quantitative types” section.

## Dancing around linear types

Linear values cannot be duplicated or ignored, but provided we know how the values are constructed, we can work hard enough to tear them apart and build them back up. This allows us to dance around the issue of ignoring and duplicating linear variables by exploiting pattern matching in order to implement functions such as `copy : (1 v : a) -> (a, a)` and `drop : (1 v : a) -> ()`.

The next snippet show how to implement those for `Nat`:

---

```
dropNat : (1 v : Nat) -> ()
dropNat Z = ()
dropNat (S n) = dropNat n

copyNat : (1 v : Nat) -> (Nat, Nat)
copyNat Z = (Z, Z)
copyNat (S n) = let (a, b) = copyNat n in
                (S a, S b)
```

---

It is worth noticing that `dropNat` effectively spends  $O(n)$  doing nothing, while `copyNat` *simulates* allocation by constructing a new value that is identical and takes the same space as the original one (albeit very inefficiently, one would expect a `memcpy` to be  $O(1)$ , not  $O(n)$  in the size of the input).

We can encapsulate those functions in the following interfaces:

---

<sup>8</sup>This is because we have no information about the type parameter `a`, as we will see later, if we are given more information about how to construct and destructure values, we *can* implement `copy` and `drop`.

---

```
interface Drop a where
  drop : (1 v : a) -> ()

interface Copy a where
  copy : (1 v : a) -> (a, a)
```

---

Since we know how to implement those for `Nat` we can ask ourselves how we could give implementation of those interfaces to additional types.

---

```
module Main

import Data.List

data Tree a = Leaf a | Branch a (Tree a) (Tree a)

interface Drop a where
  drop : (1 v : a) -> ()

interface Copy a where
  copy : (1 v : a) -> (a, a)

Drop a => Drop (List a) where
  copy ls = ?drop_list_impl

Copy a => Copy (List a) where
  copy ls = ?copy_list_impl

Drop a => Drop (Tree a) where
  copy tree = ?drop_Tree_impl

Copy a => Copy (Tree a) where
  copy tree = ?copy_tree_impl
```

---

Going through this exercise would show that this process of tearing apart values and building them back up would be very similar than what we've done for `Nat`, but there is an additional caveat to this approach: primitive types do not have constructors like `Nat` and `Tree` do. `String`, `Int`, `Char`, etc, are all assumed to exist in the compiler and are not declared using the typical `data` syntax. This poses a problem for our implementation of `Copy` and `Drop` since we do not have access to their constructors nor structure. A solution to this problem will be provided in the section “QTT ergonomics”

## Linear intuition

One key tool in understanding this thesis is to develop an intuition for linear types. Since they are (mostly) not present in other programming languages, it is important to build up this intuition especially in the absence of familiarity with the practice of linear types.

To that end we are going to explore a number of examples that illustrate how linearity can be understood.

### You can't have your cake and eat it too

Imagine this procedure

---

```
eat : (1 cake : Cake) -> Full
eat (MkCake ingredients) = digest ingredients
```

---

This allows you to eat your cake. Now imagine this other one:

---

```
keep : (1 cake : Cake) -> Cake
keep cake = cake
```

---

This allows you to keep your cake to yourself. Given those two definitions, you can't both *have* your cake *and eat it too*:

---

```
notPossible : (1 cake : Cake) -> (Full, Cake)
notPossible cake = (eat cake, keep cake)
```

---

This fails with the error

---

```
> Error: While processing right hand side of notPossible.
>   There are 2 uses of linear name cake.
>
>   |
>   | notPossible cake = (eat cake, keep cake)
>   |                               ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
>
> Suggestion: linearly bounded variables must be used exactly once.
```

---

A linear variable must be used exactly once, therefore, you must choose between having it, or eating it, but not both.

Note that removing every linearity annotation makes the program compile:

---

```
eat : (cake : Cake) -> Full
eat (MkCake ingredients) = digest ingredients
```

```
keep : (cake : Cake) -> Cake
keep cake = cake
```

```
nowPossible : (cake : Cake) -> (Full, Cake)
nowPossible cake = (eat cake, keep cake)
```

---

Since we have not said anything about `Cake` it means there are no restriction in usage, and we can have as much cake as we want and keep it too.

### **Drawing unexpected parallels**

Take the following picture:



This is a simple *connect the dots* game. If you have this text printed out and have a pencil handy, I encourage you to try it out and discover the picture that hides behind those dots.



The point of this exercise is to show what happens to a linear variable once it's consumed: It cannot be reused anymore. The dot pattern is still visible, the numbers can still instruct you how to connect them. But since the drawing has already been carried out, there is no possible way to repeat the experience of connecting the dots.

Linear variables are the same, once they are used, they are “spent”. This is shown explicitly in Idris2's type system by using holes:

---

```
let 1 dots = MkDots
    1 drawing = connect dots in
    ?rest
```

---

inspecting the hole we get:

---

```
> 0 dots : Graph
> 1 drawing : Graph
> -----
> rest : Fun
```

---

Which indicates that, while we can still *see* the dots, we cannot do anything with them, they have linearity 0. However, we ended up with a **drawing** that we can now use! <sup>9</sup>

### Safe inlining with 1

Linear variables have to be used exactly once, no less, no more. An extremely nice property this gives us can be summarised with the following statement:

A linear variable can always be safely inlined

*Inlining* refers to the ability of a compiler to replace a function call by the implementation of the function. This is a typical optimisation technique aimed at reducing the cost of function calls<sup>10</sup> as well as enabling further optimisations on the resulting program.

*Safely inlined* means that the inlining process will not result in a bigger and less efficient program. Take the following example:

---

<sup>9</sup>You will notice that the program asks us to return a value of type `Fun` this is because the goal of this exercise is to have fun.

<sup>10</sup>In low level programming, procedure calls require a context change where all the existing variables are stored in long term storage, then the procedure is called, then the result is stored somewhere, and finally the previous context is finally restored.

---

```
let x = f y in
(x, x)
```

---

After inlining `x`, that is, replace every occurrence of `x` by its definition, we obtain:

---

```
(f y, f y)
```

---

Which is less efficient than the original program. Indeed, imagine that `f` is a function that takes 5 days to run. The first case calls `f` once and duplicates its result, which would take 5 days. But the second case calls `f` twice, which would take 10 days in total.

If `x` were to be *linear* this problem would be caught:

---

```
let 1 x = f y in
(x, x)
```

---

---

```
> There are 2 uses of linear variable x
```

---

Conversely, if a program typechecks while using linear variable, then all linear variables can be inlined without loss of performance. What's more, inlining can provide further opportunities for optimisations down the line. In the following example, while `y` cannot be inlined, `x` can be.

---

```
let 1 x = 1 + 3 in
  y = x + 10 in
(y, y)
```

---

The result of inlining `x` would be as follows <sup>11</sup>:

---

```
let y = 1 + 3 + 10 in
(y, y)
```

---

### Erased runtime for `0`

In Idris2, variables can also be annotated with linearity `0`, this means that the value is *inaccessible* and cannot be used. But if that were truly the case, what would be the use of such a variable?

---

<sup>11</sup>This example does not quite work in Idris2 as it stands, we will talk about this topic in the “QTT ergonomics” section.

Those variables are particularly useful in a dependently-typed programming language because, while they cannot be used in the body of our program, they can be used in type signatures. Take this example with vector:

---

```
length : Vect n a -> Nat
length [] = Z
length (_ :: xs) = S (length xs)
```

---

The length of the vector is computed by pattern matching on the vector and recursively counting the length of the tail of the vector and adding `+1` to it (recall the `S` constructor for `Nat`). If the vector is empty, the length returned is zero (`Z`).

Another way to implement the same function in `Idris1` (without linear types) was to do the following:

---

```
-- This works in Idris1
length : Vect n a -> Nat
length _ {n} = n
```

---

the `{n}` syntax would bring the value from the *type level* to the *term level*, effectively making the type of the vector a value that can be used within the program. However doing the same in `Idris2` is forbidden:

---

```
> Error: While processing right hand side of length.
>   n is not accessible in this context.
>
>   |
>   | length _ {n} = n
>   |               ^
```

---

It is hard to understand why this is the case just by looking at the type signature `Vect n a -> Nat` and this is because it is not complete. Behind the scenes, the `Idris2` compiler is adding implicit arguments <sup>12</sup> for `n` and `a` and automatically gives them linearity `0`. The full signature looks like this:

---

<sup>12</sup>Implicit arguments are arguments to function that are not given by the programmer, but rather are filled in by the compiler automatically. Implicit arguments are extremely important in dependently-types languages because without them every type signature would be extremely heavy. Moreover, since the distinction between types and terms is blurry, the mechanism to infer *types* is the same as the mechanism to infer *terms* which is how the compiler can infer which value to insert whenever a function require an implicit argument.

```
length : {0 n : Nat} -> {0 a : Type} -> Vect n a -> Nat
length _ {n} = n
```

The `0` means we cannot use the variable outside of type signatures, but how come we can use them in the type signature and not to implement our `length` function?

The difference is that linearity **0** variables are available *at compile time* and are forbidden to appear *at runtime*. The compiler can use them, compute types with them, but they cannot be allocated and used during execution of the program we generate.

This is why linearity **0** variables are also called **erased** variables because they are removed from the execution of the program. We can use them to convince the compiler that some invariants hold, but we cannot allocate any memory for them during the execution of our program.

**Arguments annotated with `0`** Finally, another subtlety is that erased variable can actually appear inside the body of function, but only in position where they are allowed are arguments to functions with `0` usage. Such functions are:

```
toNat : (0 n : Nat) -> INat n -> Nat
toNat Z Zero = Z
toNat (S n) (Succ m) = S (toNat n m)
```

--           ▲   ▲  
--           |   |  
--           └── Used even if erased  
--           └── Bound with linearity 0

Here the recursive call uses `n` which has linearity `0`, but this is allowed because the first argument of `toNat` takes an argument of linearity `0`. In other words, `n` cannot be consumed, but `toNat` does not consume its first argument anyways, so all is good.

## Rewrites

```
sym : (0 prf : x = y) -> y = x
sym prf = rewrite prf in Refl
```

Rewriting a type does not consume the proof.

## Type signatures

---

```
--      ↑ `n` is erased
--      ▼
reverse' : {0 n : Nat} -> (1 vs : Vect n Nat) -> Vect n Nat
reverse' vs = let v2 : Vect n Nat = reverse vs in v2
--
--      ▲
--      | `n` appears here
```

---

Even if  $n$  appears in the body of the function, appearing in a type signature does not count as a use.

## No branching with 0

In general, we cannot match on erased variables, there is however an exception to this rule. Whenever matching on a variable *does not* result in additional branching, then we are allowed to match on this variable, even if it erased. Such matches are called *uninformative*, and they are characterised by the fact that they do not generate new codepaths.

No new codepaths means that, whether we match or not, the output bytecode would be the same. Except that matching on those variable would inform us of very important properties from our types. Just like the `intOrString` example informed us of the return type of our function, an uninformative match can reveal useful information to both the programmer and the compiler.

A good example of an uninformative match is `Refl` which has only one constructor:

---

```
data (=) : (a, b : Type) -> Type where
  Refl : (a : Type) -> a = a
```

---

This suggests that, even if our equality proof has linearity 0, we can match on it, since there is only 1 constructor we are never going to generate new branches.

But an uninformative match can also happen on types with multiple constructors. Take this indexed <sup>13</sup> `Nat` type:

---

<sup>13</sup>A *type parameter* that changes with the values that inhabit the type. For example `["a", "b", "c"] : Vect 3 String` has index 3 and a type parameter `String`, because it has 3 elements and the elements are Strings.

```
data INat : Nat -> Type where
  IZ : INat Z
  IS : INat n -> INat (S n)
```

---

This is simply a duplicate for `Nat` but carries its own value as index. Now let us write a function to recover the original `Nat` from an `INat`:

```
toNat : (0 n : Nat) -> (1 m : INat n) -> Nat
toNat Z IZ = Z
toNat (S n) (IS m) = S (toNat n m)
```

---

Even if we annotated `n` with linearity `0` we are allowed to match on it. To understand why, let us add some holes and remove the matching:

```
toNat : (0 n : Nat) -> (1 m : INat n) -> Nat
toNat n IZ = ?branch
toNat (S n) (IS m) = S (toNat n m)
```

---

Idris2 will not allow this program to compile and will fail with the following error:

```
> Error: While processing left hand side of toNat.
> When unifying INat 0 and INat ?n.
> Pattern variable n unifies with: 0.
>
> |
> |   IZ : INat Z
> |   ^
> |   IS : INat n -> INat (S n)
> |   toNat n IZ = ?branch
> |   ^
>
> Suggestion: Use the same name for both pattern variables, since they
> unify.
```

---

It tells us that `n` unifies with `Z` and forces the user to spell out the match. Effectively forcing uninformative matches to be made. A similar error appears if we try the same thing on the second branch, trying to remove `S n`.

## 4 Quantitative Type Theory in practice

QTT is still a recent development and because of its young age, it has not seen wide spread use in commercial applications. The Idris2 compiler itself stands as the most popular example of a complex program that showcases uses for QTT and quantitative types. Linear types already benefit from a body of work that showcase their uses [34][14][35][36][37][38][15][13][12], but one of the goals of this thesis was to list and discover some new and innovative uses for linear types and QTT.

### limitations and solutions for quantitative types

We've seen how we can write addition of natural numbers using linear types. But can we write a multiplication algorithm using linear types? Let us inspect the traditional multiplication algorithm and see if we can update it with linear types.

#### Linear multiplication

Here is a multiplication function without any linear variables

---

```
multiplication : (n : Nat) -> (m : Nat) -> Nat
multiplication Z m = Z
multiplication (S n) m = m + (multiplication n m)
```

---

Just like with addition, we notice that some variables are only used once, but some aren't. `n` is used exactly once in both branches, but `m` is not used in one branch, and used twice in the other. Which leads to the following program:

---

```
multiplication : (1 n : Nat) -> (m : Nat) -> Nat
multiplication Z m = Z
multiplication (S n) m = m + (multiplication n m)
```

---

Which compiles correctly, but how could we go about implementing a completely linear version of `multiplication`?

indeed, writing `multiplication : (1 n : Nat) -> (0 m : Nat) -> Nat` gets us the error:

Error: While processing right hand side of multiplication. m is not accessible in this context.

```
|
| multiplication (S n) m = m + (multiplication n m)
|                               ^
```

---

Which catches the fact that m is use twice in the second branch (but the first branch is fine).

Ideally we would like to write this program:

```
--           The multiplicity depends on the first arugment
--                               v
multiplication : (1 n : Nat) -> (n m : Nat) -> Nat
multiplication Z m = Z
multiplication (S n) m = m + (multiplication n m)
```

---

However Idris2 and QTT do not support *dependent linearities* or *first class linearity* where linearity annotations are values within the language.

We can however attempt to replicate this behaviour with different proxies:

```
provide : Copy t => Drop t => (1 n : Nat) -> (1 v : t) -> (DPair Nat (\x => n = x), Vect n t)
provide 0 v = let () = drop v in (MkDPair Z Refl, [])
provide (S k) v = let (v1, v2) = copy v
                    (MkDPair n prf, vs) = provide k v1 in MkDPair (S n) (cong S prf), v2 :: vs)

multiplication : (1 n, m : Nat) -> Nat
multiplication n m = let (MkDPair n' Refl, ms) = provide n m in mult n' ms
  where
    mult : (1 n : Nat) -> (1 vs : Vect n Nat) -> Nat
    mult 0 [] = 0
    mult (S k) (m :: x) = m + (mult k x)
```

---

This program attempts to simulate the previous signature by creating a dependency between n and a vector of length n containing copies of the variable m<sup>14</sup> with the type `mult : (1 n : Nat) -> (1 _ : Vect n Nat) -> Nat`.

<sup>14</sup>For the purposes of this example there is no proof that the vector *actually* contains only copies of m but this is an invariant that could be implemented at the type level with a data structure like the following:

```
data NCopies : (n : Nat) -> (t : Type) -> (v : t) -> Type where Empty : NCopies Z t v
-- More : (1 v : t) -> (1 vs : NCopies n t v) -> NCopies (S n) t v
```

The example given uses Vect for brevity and readability, the code quickly becomes unwieldy with equality proofs everywhere, which aren't the point of the example.



As we’ve demonstrated, we technically can express more complex relationship between linear types provided they implement our interfaces `Drop` and `Copy`. However, the extra work to make the dependency explicit in the type isn’t worth the effort. Indeed, giving up this dependency allows us to write the following program:

---

```

lmult : (l n, m : Nat) -> Nat
lmult 0 m = let () = drop m in Z
lmult (S k) m = let (a, b) = copy m in a + lmult k b

```

---

Which is a lot simpler and achieves the same goal, it even has the same performance characteristics.

## Permutations

The following example is certainly my favourite since it combines both dependent types and linear types in a way that wasn’t possible before. It has been done in the context of my work for the company Statebox strongly relies on dependent types to write formally verified software.

One of Statebox’s project is a validator for petri-nets[39] and petri-net executions: FSM-oracle. While the technical details of this projects are outside the scope of this text, there is one aspect of it that is fundamentally linked with linear types, and that is the concept of permutation.

FSM-Oracle describes petri-nets using *hypergraphs* [40] those hypergraphs have a concept of *permutation* that allows to move wires around. This concept is key in a correct and proven implementation of hypergraphs. However, permutations also turn out to be extremely complex to implement as can attest the files trying to fit their definition into a Category.

The relevant bit about permutation can be found in the two files `Permutations.idr` and `SwapDown.idr`, a permutation relies on a list of “swaps” that describe the operation that we can do on a list to generate a new permutation of the same list.

---

```

data Perm : {o : Type} -> List o -> List o -> Type where
  Nil : Perm [] []
  Ins : Perm xs ys -> SwapDown (a::ys) zs -> Perm (a::xs) zs

data SwapDown : List t -> List t -> Type where
  HereS : SwapDown (a::as) (a::as)
  ThereS : SwapDown (a::as) bs -> SwapDown (a::b::as) (b::bs)

```

---

If you don't have access to the internet to witness the proofs in their full glory, here is an example of what we are dealing with:

```

permAssoc : (ab : Perm aas bbs) -> (bc : Perm bbs ccs) -> (cd : Perm ccs dds)
            -> permComp ab (permComp bc cd) = permComp (permComp ab bc) cd
permAssoc Nil bc cd = Refl
permAssoc (Ins {xs=as} {ys=bs} ab' abb) bc cd with (shuffle abb (permComp bc cd)) proof bdPrf
| Ins {ys=ds} bd' add with (shuffle abb bc) proof bcPrf
| Ins {ys=cs} bc' acc with (shuffle acc cd) proof cdPrf
| Ins {ys=ds'} cd' ad'd =
  let (Refl, Refl, Refl) = shuffleComp abb bc cd bcPrf cdPrf bdPrf in
  insCong5 Refl Refl Refl (permAssoc ab' bc' cd') Refl

```

---

This function ensures that composition of permutation is associative. It relies on helper functions such as `shuffleComp` which are not even implemented in `Idris1`, because it is too hard to convince the compiler.

Linear types can thankfully ease the pain by providing a very simple representation of permutations :

```

Permutation : Type -> Type
Permutation a = (l : List a) -> List a

```

---

That is, a `Permutation` parameterised over a type `a` is a linear function from `List a` to `List a`<sup>15</sup>.

This definition works because no elements from the input list can be omitted or reused for the output list. *Every single element* from the argument has to find a new spot in the output list. Additionally, since the type `a` is unknown, no special value can be inserted in advance. Indeed, the only way to achieve this effect would be to pattern match on `a` and create values once `a` is known, but this would require `a` to be bound with a multiplicity greater than 0:

---

<sup>15</sup>This has already been formally proven by Bob Atkey <https://github.com/bobatkey/sorting-types/blob/master/agda/Linear.agda>

---

```

fakePermutation : {a : Type} -> (1 _ : List a) -> List a
fakePermutatoin {a = Int} ls = 42 :: ls
fakePermutation {a = _} ls = reverse ls

```

---

In this example, `a` is bound with *unrestricted* multiplicity, which give us the hint that it *is* inspected and the permutation might not be a legitimate permutation.

What's more, viewing permutations as a function gives it extremely simple categorical semantics: It is just an instance of the category of types with linear functions as morphisms.

Assuming `Category` is defined this way:

---

```

-- operator for composition
infix 2 .*.
-- operator for morphisms
infixr 1 ~>

record Category (obj : Type) where
  constructor MkCategory
  (~>)      : obj -> obj -> Type -- morphism
  identity  : {0 a : obj} -> a ~> a
  (.*.)     : {0 a, b, c : obj}
             -> (a ~> b)
             -> (b ~> c)
             -> (a ~> c)
  leftIdentity : {0 a, b : obj}
             -> (f : a ~> b)
             -> identity .*. f = f
  rightIdentity : {0 a, b : obj}
             -> (f : a ~> b)
             -> f .*. identity = f
  associativity : {0 a, b, c, d : obj}
             -> (f : a ~> b)
             -> (g : b ~> c)
             -> (h : c ~> d)
             -> f .*. (g .*. h) = (f .*. g) .*. h

```

---

We can write and instance of `Category` for `List o`:

---

```
Permutation : List o -> List o -> Type
Permutation a b = Same a b

permutationCategory : Category (List o)
permutationCategory = MkCategory
  Permutation
  sid
  linCompose
  linLeftIdentity
  linRightIdentity
  linAssoc
```

---

Using the definitions and lemmas for `Same` which is a data type that represents a linear function between two values of the same type:

---

```

-- a linear function between two values of the same type
record Same {o : Type} (input, output : o) where
  constructor MkSame
  func : LinearFn o o
  -- check the codomain of the function is correct
  check : (func `lapp` input) = output

sid : Same a a
sid = MkSame lid Refl

linCompose : {o : Type}
  -> {o a, b, c : o}
  -> Same a b
  -> Same b c
  -> Same a c
linCompose (MkSame fn Refl) (MkSame gn Refl)
  = MkSame (lcomp fn gn) Refl

linRightIdentity : {o : Type}
  -> {o a, b : o}
  -> (f : Same a b)
  -> linCompose f (MkSame Main.lid Refl) = f
linRightIdentity (MkSame (MkLin fn) Refl) = Refl

linLeftIdentity : {o : Type}
  -> {o a, b : o}
  -> (f : Same a b)
  -> linCompose (MkSame Main.lid Refl) f = f
linLeftIdentity (MkSame (MkLin fn) Refl) = Refl

linAssoc : (f : Same a b) ->
  (g : Same b c) ->
  (h : Same c d) ->
    linCompose f (linCompose g h) = linCompose (linCompose f g) h
linAssoc (MkSame (MkLin fn) Refl)
  (MkSame (MkLin gn) Refl)
  (MkSame (MkLin hn) Refl) = Refl

```

---

LinearFunction is defined as follows:

---

```

record LinearFn (a, b : Type) where
  constructor MkLin
  fn : (1 _ : a) -> b

lid : LinearFn a a
lid = MkLin (\1 x => x)

lapp : LinearFn a b -> (1 _ : a) -> b
lapp f a = f.fn a

lcomp : LinearFn a b -> LinearFn b c -> LinearFn a c
lcomp f g = MkLin (\1 x => g.fn (f.fn x))

```

---

While this looks like a lot of code, the entire definition holds within 100 lines (including the `Category` definition), and a lot of the definitions like `LinearFn` and `Same` are generic enough to be reused in other modules.

Additionally, this approach is extremely straightforward. So much that in the future, it wouldn't seem extravagant to have the type-system automatically generate the code as part of a derived interface.

Finally and most importantly, this program could not exist without using both dependent types to declare the necessary proofs to formally verified our Categorical structure, and linear types to implement `Permutation`. It is a beautiful example where the two features meet and merge in a way that is both practical and elegant.

## Levitation improvements

*The gentle art of levitation*[41] shows that a dependently typed language has enough resources to describe all indexed data types with only a few constructors. The ability to define types as a language library rather than a language features allows a great deal of introspection which in turns allows a realm of possibilities. Indeed, we can now define operations on those data types that will preserve the semantics of the type but make the representation more efficient. We can generate interface implementations for them automatically. And we can use them to construct new types out of existing ones[42][43][44], including representations for primitive types such as `Int` or `String`, an approach that's been proven effective with `Typedefs`[45].

*The practical guide to levitation*[41] shows that those features are plagued by multiple shortcomings: the verbosity of the definitions not only make the data

declaration hard to write and read, it makes the compiler spend a lot of time constructing and checking those terms and it has trouble identifying what is a type parameter or what is an index.

Thankfully, the performance inefficiency from levitation can be alleviated by a smart use of erasure. In his thesis, Ahmad Salim relies on erasing terms with the `.` (dot) syntax, which does its best but cannot enforce erasure of terms.

While the following has not been implemented, it shows Idris2 has a lot of promise in lifting previous performance limitations. This is because, in Idris2, we can perform and enforce erasure by annotating well-formedness proofs with `0` and use data types such as `Exists` instead of `DPair`.

More challenges arise when we try to use levitation to represent Idris data definitions. Indeed, linear and erased variables in constructors cannot be represented. We cannot represent the following constructor.

---

```
(::) : {n : Nat} -> {0 a : Type} -> a -> Vect n a -> Vect (S n) a
```

---

This suggests that levitation could be extended to support constructor with linear and erased arguments, but the it's unknown if *levitation* itself (defining the description of linear data types in terms of itself) would be achievable.

Interestingly enough, encoding linearity in levitated description might also help fix one of the shortcoming of levitation in idris: automatically discerning between type parameters and type indices. Combined with the ability to pattern match on types, in turn would allow the Idris2 compiler to generate definitions for interfaces such as `Functor` and `Applicative`.

## Compile-time string concatenation

Strings are ubiquitous in programming. That is why a lot of programming languages have spent a considerable effort in optimising string usage and string API ergonomics. Most famously, Perl is notorious for its extensive and powerful string manipulation API including first-class regex support with more recent additions including built-in support for grammars.

One very popular feature to ease the ergonomics of string literals is *string interpolation*. String interpolation allows you to avoid this situation

---

```
show (MyData arg1 arg2 arg3 arg4) = "MyData (" ++ show arg1 ++ " " ++ show arg2 ++ " " ++ show arg3 ++
```

---

by allowing string literal to include expressions *inline* and leave the compiler to build the expected string concatenation. One example of string interpolation syntax would look like this:

---

```
show (MyData arg1 arg2 arg3 arg4) = "MyData ({arg1} {arg2} {arg3} {arg4})"
```

---

The benefits are numerous but I won't dwell on them here. One of them however is quite unexpected: Predict compile-time concatenation with linear types.

As mentioned before, the intuition to understand the *erased linearity*  $\emptyset$  is to consider those terms absent at runtime but available at compile-time. In the case of string interpolation, this intuition becomes useful in informing the programmer when the compiler is able to perform compile-time concatenation.

---

```
let name = "Susan"
    greeting = "hello {name}" in
    putStrLn greeting
```

---

In the above example, it would be reasonable to expect the compiler to notice that the variable `name` is a string literal and that, because it is only used in a string interpolation statement, it can be concatenated at compile time. Effectively being equivalent to the following:

---

```
let greeting = "hello Susan" in
    putStrLn greeting
```

---

But those kind of translations can lead to very misleading beliefs about String interpolation and its performance implications. In this following example the compiler would *not* be able to perform the concatenation at compile time:

---

```
do name <- readLine
    putStrLn "hello {name}"
```

---

Because the string comes from the *runtime*. Indeed static strings can be inserted at compile-time while strings from the runtime need to be concatenated. This means the following program typechecks:

---

```
let  $\emptyset$  name = "Susan"
    1 greeting = "hello {name}" in
    putStrLn greeting
```

---



Since the variable `name` has linearity `0`, it cannot appear at runtime, which means it cannot be concatenated with the string `"hello "`, which means the only way this program compiles is if the string `"Susan"` is inlined with the string `"hello "` at compile-time.

Using holes we can describe exactly what would happen in different circumstances. As a rule, string interpolation would do its best to avoid allocating memory and performing operations at runtime. Much like our previous optimisation, it would look for values which are constructed in scope and simply concatenate the string without counting it as a use.

---

```
let 1 name = "Susan"
  1 greeting = "hello {name}" in
  putStrLn greeting
```

---

Would result in the compile error

---

```
There are 0 uses of linear variable name
```

---

Adding a hole at the end would show.

---

```
let 1 name = "Susan"
  1 greeting = "hello {name}" in
  ?interpolation
```

---



---

```
1 name : String
1 greeting : String
-----
interpolation : String
```

---

As you can see, the variable `name` has not been consumed by the string interpolation since this transformation happens at compile time.

Having the string come from a function call however means we do not know if it has been shared before or not, which means we cannot guarantee (unless we restrict our programming language) that the string was not shared before, therefore the string cannot be replaced at compile time.

---

```
greet : (1 n : String) -> String
greet name = let 1 greeting = "hello {name}" in ?consumed
```

---

```
0 name : String  
1 greeting : String
```

```
-----  
consumed : String
```

---

The string `name` has been consumed and the core will therefore perform a runtime concatenation.

## 5 Quantitative Type Theory and Programming ergonomics

Idris2 features both linear and dependent types, providing programmers with the tools to write elegant and correct software[5], while ensuring performance. In this thesis I am going to reuse the intuition originally suggested by [13][14]\cite{linear\_types\_update[15]} and rephrase it as

If you own a linear variable and it is not shared, you can freely mutate it

This somewhat mimics Rust’s [46] model of ownership where variables are free to be mutated as long as they are uniquely owned. But differs in that it does not make use of *uniqueness* but rather uses linearity as a proxy for it.

This idea can be illustrated with the following example:

---

```
let 1 v : Int = 3
    1 v' : Int = v + 17
in print v'
```

---

The bound variable `v` was created locally and is *linear* advertising a use of `1`. Then, the function `+ 17` makes use of it and puts the result in a variable `v'`. This new variable does not need to allocate more memory and can simply reuse the one allocated for `v`, effectively mutating the memory space occupied by `v'` without losing the immutable semantics we expect.

As one can see this does away with the continuation while reaping the benefits of in-place mutation awarded by our linear property. In addition, this innocuous could be expanded to more complex examples like

---

```
let 1 array = ...
    1 array' = map f array
    1 array'' = map g array'
in sum array''
```

---

Which could be executed in *constant space complexity* effectively duplicating the results from deforestation, but in a more general setting since this approach does not make any assumption about the type of the data handled, only that operations are linear and mutations are constant in space.

## Mapping primitive to data types and vice-versa

At the end of the introduction we saw how primitive types are a problem for linearity, because we have no way of constructing them with classical constructors.

As a reminder, we were trying to implement the functions `copy : (1 _ : a) -> (a, a)` and `drop : (1 _ : a) -> ()` for primitive types such as `String` or `Int`.

Those types are not defined using the `data` syntax used for other user-defined types. Rather, they are assumed to exist by the compiler and are built using custom functions which are different for each backend. If only `String` and `Int` were defined as plain data types, we could implement functions such as `copy` and `drop`.

It turns out this is possible, we can use a clever encoding that maps plain data types to primitive types and have the compiler “pretend” until the codegen phase where they are substituted by their primitive variants. Just like in *Haskell*, `String` could be represented as a `List` of `Char`. `Char` itself is a primitive type, but the key insight here is to see that itself can be represented as `Vect 8 Bool`. Using both those definitions our primitive types have now become plain data with regular constructors:

---

```
Bit : Type
Bit = Bool

Int32 : Type
Int32 = Vect 32 Bit

Char : Type
Char = Vect 8 Bit

String : Type
String = List Char
```

---

This allows us to implement `Copy` and `Drop` by inheriting the instance from `List` and `Vect`. Additionally, since the procedure to implement those instances is very mechanical, it could almost certainly be automatically derived.

This approach is reminiscent of projects like *practical levitation*[47] or *Type-defs*<sup>16</sup>, which describe data types as data structures. And the benefits of both

---

<sup>16</sup><http://typedefs.com>

would translate quite well to our situation, beyond the ability to dance around linearity with primitive types.

Indeed, in both instance, once our data type is represented we can use its structure to infer its properties. For example, if a data type has *type parameters* then we can generate instance for `Functor`, `Applicative`, etc. If the data type resembles `List Char` then we can replace it by the primitive type `String`. Finally, we can use semantic-preserving operations on our data types in order to optimise their representations in the generated code. For example `data Options = Folder Int | Directory Int` is equivalent to `Vect 33 Bool` which can be represented as an unboxed `Int64`. Even more optimisations could be performed by mapping `Vect` of constant length to buffers of memory of constant size and index through them in  $O(1)$  instead of  $O(n)$ .

As the cherry on top, this mapping would allow the coverage checker to infer missing cases accurately for primitive types. Indeed this example should work but the Idris2 compiler is unable to check the coverage of all `Strings`:

---

```
data IsTrueOrFalse : String -> Type where
  IsTrue  : IsTrueOrFalse "True"
  IsFalse : IsTrueOrFalse "False"

fromString : (str : String) -> IsTrueOrFalse str => Bool
fromString "True"  @{{IsTrue}} = True
fromString "False" @{{IsFalse}} = False
```

---

However, the following works correctly, suggesting that the special treatment of primitives is the culprit.

---

```

data Bit = I | 0

MyChar : Type
MyChar = Vect 8 Bit

MyString : Type
MyString = List MyChar

data IsTrueOrFalse : MyString -> Type where
  IsTrue : IsTrueOrFalse [[0,0,0,0,0,0,0,0]]
  IsFalse : IsTrueOrFalse [[0,0,0,0,0,0,I,0]]

fromString : (str : MyString) -> IsTrueOrFalse str => Bool
fromString [[0,0,0,0,0,0,0,0]] @{IsTrue} = True
fromString [[0,0,0,0,0,0,I,0]] @{IsFalse} = False

```

---

## Idris2 and multiplicities

Linear types allow us to declare variable to be use either 0, 1 or any number of times. However, we've seen this approach is pretty restrictive, we've also seen that this limitation has been caught on early on with Bounded Linear Logic. In our case, we're interested in making program more efficient by avoiding extraneous memory allocation to take place.

As we've seen in the previous chapter, though those optimisations are promising, they are not necessarily significant for every program. I posit that in order to bring about *significant* change, and equally significant (but not unreasonable) change must be made to the computational model: Changing the memory model to reference counting rather than runtime garbage collection.

In this section I will explain the premise that lead me to this statement and follow it up with preliminary work done to achieve this goal.

### The problem with runtime garbage collection (WIP)

*Garbage collection* generally refers to algorithm of automatic memory management such as *mark & sweep* or to runtimes using them such as the *Boehm-GC*.

The greatest benefit of garbage collection is that the burden of memory management is completely lifted from the programmer. But this feature is not always good news, the trade-off comes at the expense of predictability and control.

Since the garbage collector runs automatically without any input from the pro-

grammer, there is no way to tell how often the garbage collector will run and for how long, just by looking at a program. What’s more, garbage collection does not happen instantly, some algorithm stop the execution of the program for an unknown amount of time in order to reclaim memory. And this process cannot be

## What is reference counting (WIP)

### Alternative semirings and their semantics (WIP)

Quantitative type theory track usage through a generic semirings, any implementation of QTT can use any semiring for it to function. Idris2 uses the following semiring:

---

```
data ZeroOneOmega = Rig0 | Rig1 | RigW

rigPlus : ZeroOneOmega -> ZeroOneOmega -> ZeroOneOmega
rigPlus Rig0 a = a
rigPlus a Rig0 = a
rigPlus Rig1 a = RigW
rigPlus a Rig1 = RigW
rigPlus RigW RigW = RigW

rigMult : ZeroOneOmega -> ZeroOneOmega -> ZeroOneOmega
rigMult Rig0 _ = Rig0
rigMult _ Rig0 = Rig0
rigMult Rig1 a = a
rigMult a Rig1 = a
rigMult RigW RigW = RigW
```

---

### Putting everything together: Linearity and reference counting (WIP)

The idea of mapping linear usage to reference counting isn’t new [16], but with QTT, we can have a much greater control over the semantics of the semiring we use in order to make it correspond precisely with the behaviour of our runtime.

### Steps toward a working implementation

The first and easier step toward a working implementation of this idea is to extend the Idris2 compiler to support alternative multiplicities. The current Idris2 compiler uses `data RigCount = Rig0 | Rig1 | RigW` for linearity `0`, `1`, `ω` respectively. In order to support a different semiring we need to abstract over

the operations of semirings and update the existing codebase to make use of them instead of relying on our definition of `RigCount`.

A `Semiring` is defined as follows:

---

```
interface Semiring a where
  (|+|) : a -> a -> a
  plusNeutral : a
  (|*|) : a -> a -> a
  timesNeutral : a
```

---

Where `plusNeutral` is the neutral element for `|+|` and `timesNeutral` is the neutral element for `|*|`. `|+|` and `|*|` are binary functions which distribute as expected.

This change is quite disruptive since it forbids us to pattern match on `RigCount` and instead forces us to be entirely generic on the multiplicity we use. Signatures like:

---

```
linearCheck : RigCount -> RigCount -> Bool
```

---

Become:

---

```
linearCheck : Semiring r => r -> r -> Bool
```

---

In order to recover the functionality from pattern matching we introduce an eliminator for `Semiring`:

---

```
elimSemi : (Semiring a, Eq a) => (zero : b) -> (one : b) -> (a -> b) -> a -> b
elimSemi zero one other r {a} =
  if r == Semiring.plusNeutral {a}
  then zero
  else if r == Semiring.timesNeutral {a}
    then one
    else other r
```

---

As well as a series of helper functions:

---



```

isErased : (Semiring a, Eq a) => a -> Bool
isErased = elimSemi True False (const False)

isLinear : (Semiring a, Eq a) => a -> Bool
isLinear = elimSemi False True (const False)

isRigOther : (Semiring a, Eq a) => a -> Bool
isRigOther = elimSemi False False (const True)

branchZero : (Semiring a, Eq a) => Lazy b -> Lazy b -> a -> b
branchZero yes no rig = if isErased rig then yes else no

branchOne : (Semiring a, Eq a) => Lazy b -> Lazy b -> a -> b
branchOne yes no rig = if isLinear rig then yes else no

branchVal : (Semiring a, Eq a) => Lazy b -> Lazy b -> a -> b
branchVal yes no rig = if isRigOther rig then yes else no

```

---

But this is still not enough since the compiler also makes assumptions about *ordering* within the multiplicity. In the following program, we check that multiplicities are compatible by checking if the left one is smaller than the right one.

---

```

rigSafe : RigCount -> RigCount -> Core ()
rigSafe Rig1 RigW = throw
  (LinearMisuse fc !(getFullName x) Rig1 RigW)
  -- `x` comes from the enclosing scope
rigSafe Rig0 RigW = throw
  (LinearMisuse fc !(getFullName x) Rig0 RigW)
rigSafe Rig0 Rig1 = throw
  (LinearMisuse fc !(getFullName x) Rig0 Rig1)
rigSafe _ _ = pure ()

```

---

The same program can be refactored using an ordering on the semiring we use:

---

```

rigSafe : Semiring r => r -> r -> Core ()
rigSafe lhs rhs =
  when (lhs < rhs)
    (throw (LinearMisuse fc !(getFullName x) lhs rhs))

```

---

Semirings in general do not enforce an order and we do not want to add too many restrictions to our multiplicity types to avoid making it less flexible than it

needs. Typically, if affine types were to be implemented using intervals, we could not rely on a *total order*<sup>17</sup> since some intervals cannot be compared with each others. For example `[3..5]` and `[0..4]` are not directly related to each other. For this reason we are going to use a **Preorder** which only requires equality to be valid for some pairs of values and requires transitivity and associativity to hold. In Idris2 this is formulated as:

---

```
interface Preorder a where
  (<=) : a -> a -> Maybe Bool
  preorderRefl : {x : a} -> x <= x = Just True
  preorderTrans : {x, y, z : a} -> x <= y = Just True -> y <= z = Just True -> x <= z = Just True
```

---

For engineering reasons this is not exactly the version used in the Idris2 compiler but this should not detract from the observation that pure semirings are not enough.

Finally, there is one last missing piece and that is a default value. In a lot of the existing implementation **RigW** is used as a default value when generating new values with unknown linearity. This is because Idris2 is *unrestricted by default* instead of *linear by default* in our case, the default value has a very neat property: it is greater or equal to every other values of the semiring. In technical term, in addition to having a semiring and a preorder we also need our multiplicity type to have a *top* value, which means we are actually dealing with a *meet semi-lattice*.

Top is easily defined as follows:

---

```
interface Preorder a => Top a where
  top : a
  -- top absolves everything else
  topAbs : {x : a} -> x <= top = Just True
```

---

Now that we finally have all the pieces required by the Idris2 compiler we can replace every occurrence of **Rig0**, **Rig1** and **RigW** by our generic implementation of semirings.

## Next Steps

This exercise was particularly informative in teaching us what is *really needed* in practice to build up a programming language based on QTT. *Meet semi-lattice*

---

<sup>17</sup>A *total order* is an ordering for which every value is related to every other value.

aren't mentioned in Quantitative Type Theory[1] but they appear in Granule [8] while this is not entirely surprising it is noteworthy that both approach reach the same conclusion: ordered semirings are the foundation of a sound implantation of a calculus with quantities into a useful programming language.

This implementation is only the first step into extended the language for full reference counting without garbage collection, in the meantime, there are a lot of intermediate steps. And additionally, now that the seal is broken on multiple types of multiplicities, nothing stops us from implementing different multiplicities for Idris2 and see what happens. We can now discover semantics for semirings, rather than use different semirings to reach different semantics. Granule already has products, intervals, access variables and natural numbers in addition to linearity as their multiplicity. It would be interesting to see what matrices, or polynomials offer in terms of semantics as a resource tracking semiring.

The immediate next step for this project is to repeat this experiment (generalizing some aspect of the compiler) but with linearity checking rather than semirings. Indeed given an implementation of infinite Nats as a semiring, we should be able to replace our previous semiring by the new one and enjoy the benefits in terms of added functionality.

Unfortunately, while the compiler accepts Nat-multiplicity programs, it keeps the semantics of  $0$ ,  $1$ ,  $\omega$  because everything bigger than  $1$  is considered unrestricted In the linearity checker.

## Relaxing linearity inference (WIP)

problem: I want to do this

---

```
linearProof : (1 v : Nat) -> ProofOnNat v
linearProof Z = Refl
linearProof (S n) = let 0 prf = linearProof n in
                    rewrite prf in Refl
```

---

It doesn't work because `prf` is inferred to have linearity `1`.

This should be legal because `n` has been consumed, it just turns out that it produces noting but tha'ts fine.

Similarly, this should work

---

```

let 1 x = 1 + 3
      ω y = f x in
      (y, y)

```

---

Because we use  $x$  only once, it turns out binding to another linearity shouldn't be a problem.

Even the contrived example

```

let 1 x = 1 + 3
      ω y = (let ω z = f x in \k => z + 2) in
      (y 3, y 4)

```

---

Should work because we only access  $x$  once, even if we bind its result to a captured variable that will be accessed twice.

This last example it taken from *Once Upon a Type* which focuses on linearity inference and how to improve it in order to maximise occurrences of safe inlining. More recently, work on linearity inference[48] has resumed, motivated by its use in linear haskell [19]

In the introduction we used *safe inlining* as an intuition to understand linear variables. We saw that for a variable to be linear means that it can be inlined without any loss of performance. This inlining in turn can be used to perform further optimisations, a-la deforestation.

The rule seem to be that linear variable cannot be *captured* in the body of a lambda. Indeed if the lambda is bound with linearity other than 1 then the inlined value is

```

linearProof : (1 v : Nat) -> ProofOnNat v linearProof Z = Refl linearProof (S
n) = let proof = linearProof n in rewrite proof in Refl

```

This doesn't work because proof is automatically bound with linearity 1, however in that case I don't want to use it, at all.

I've been playing a lot with that a couple months ago and couldn't find the rules in the QTT paper that justified this behaviour. Someone asked on the slack why it was like this and I remember you said it was because "if we allow linearity unrestricted, then  $n$  (in this case) could be used more than once.

But I don't think this is true, at least in the way I understand it. It is clear if we use the "safe inlining" as an intuition for linearity that we can allow unrestricted

binding from a linear function. However, it means it cannot be inlined (because that would cause duplicate use of linear variable). Take this example

let  $x = 1 + 2$  in  $y = x + 3$  in  $y + y$

This is safe as long as  $y$  is not inlined but  $x$  can be inlined into  $y$ .

if  $x$  is inlined we get

let  $y = 1 + 2 + 3$  in  $y + y$

which is safe. However if we inline  $y$  we get

let  $x = 1 + 2$  in  $x + 3 + x + 3$

which duplicates  $x$ . So while binding  $y$  as linear forcefully deals with that, it is unnecessarily restrictive. The semantics of binding the result of a linear function to an unrestricted variable could just be “program is less efficient because we can’t prove the variable is safely inlinable”

Did I miss anything or does that sound reasonable?

## 6 Performance Improvement using linear types

This section aims to answer the question “what performance improvement can we get by using linear types *today?*”

Idris2 already allows to define linear functions, but despite the extra information that the argument cannot be used twice, the compiler does not perform any special optimisation for linear types. One obvious observation one can make is that once a linear value has been consumed, its memory space can be reclaimed, or reused. In practice this manifests this way:

---

```
isTrue : (1 _ : a) -> (f : (1 _ : a) -> Bool) -> String
isTrue arg predicate
--
--           └ We can free `arg` after this
--           ▼
  = if predicate arg then "It's true!"
    else "Fake news."
```

---

The value `arg` of type `a` can be freed immediately after being called by `predicate`. Indeed, it’s been used, therefore it cannot be used again, which means its memory space won’t ever be accessed again. Freeing memory might not sound very exciting but the main benefit is that we do not need to wait for the garbage collector to notice our value can be freed. Removing the reliance on garbage collection is a huge step toward predictable performance.

The safe-update mechanism is similar:

---

```
data Numbers = Inc Nat | Dec Nat

update : (1 _ : Numbers) -> Numbers
update (Inc n) = Inc (S n)
update (Dec Z) = Dec Z
update (Dec (S n)) = Dec n
```

---

We have data type with two cases, and update checks which case we have and then incrementing or decrementing the `Nat` value.

This function should perform in constant space ( $O(1)$ ) but currently, the Idris compiler always allocates new values when a constructor is called. In addition, the old value is now ready to be freed but we have to wait on the garbage collector to catch it.

Unfortunately, we cannot simply implement the free and update feature in order

to see what kind of performance improvement we can get out of linear types. This is because a linear variable is not guaranteed to be unique when it is called on a linear function. One example of this sharing mechanism can be seen with function subtyping.

## Linear function subtyping

In order to make the interplay between unrestricted and linear functions easier, Idris2 features subtyping on functions. As a refresher, subtyping allows a function to accept another type than the one that it has been specified with, as long as the other type is a subtype of the original one. In the following example we are going to assume we have access to two types, **A** and **B**, and that **B** is a subtype of **A**, noted with **B <: A**

---

```
f : A -> A

g : B -> B

-- Assume we have B <: A

let a : A = ...
    b : B = ...
    y1 = f a -- Expected
    y2 = f b -- b is a subtype, it's valid
    no = g a -- a is a supertype of b, invalid
in ?rest
```

---

In Idris2, types cannot have a subtyping relation, except function types. This is useful for higher-order functions such as `map` defined non-linearly:

---

```
map : (f : a -> b) -> List a -> List b
linearMap : (f : (1 _ : a) -> b) -> (1 _ : List a) -> List b
```

---

Those functions do the same things but won't typecheck the same way. Given an unrestricted `f` the second function will refuse to typecheck. Whereas feeding a linear function to `map` will compile correctly.

---

```

inc : Nat -> Nat

linInc : (1 _ : Nat) -> Nat

-- Success everything unrestricted
map inc [1,2,3]

-- Success because got linear and expected unrestricted
map linInc [1,2,3]

-- Fail because inc is unrestricted but linMap expected linear
linearMap inc [1,2,3]

-- Success because everything is linear
linearMap linInc [1,2,3]

```

---

This is because linear functions are considered to be a subtype of unrestricted functions. In formal notation it means  $((1 \_ : a) \rightarrow b) <: (a \rightarrow b)$ . This details is extremely relevant for our optimisation because it means that *linearity does not guarantee uniqueness*, which is the property we are interested in when performing those optimisations. In the following example we show how this breaks down our assumption of safe updates

---

```

update : (1 _ : Nat) -> Nat

do let list1 = [1,2,3]
   let list2 = list1
   println $ map update list1
   println list2

```

---

This program typechecks, even if we are using a linear function in an unrestricted setting. If the update was performed we should expect the output to be:

---

```

> [2, 3, 4]
> [2, 3, 4]

```

---

Instead of the (correct) output:

---

```

> [2, 3, 4]
> [1, 2, 3]

```

---

This suggests that our intuition that linear functions can be use for safe updates and safe memory frees is not restrictive enough. We need an additional level of



restriction to ensure *uniqueness* of the values instead of linearity.

## The nature of the changes required for linear optimisations

Since linearity is not enough to ensure uniqueness, we are going to add an additional restriction: We can only mutate a variable if it has been instantiated within in the immediate surrounding scope of the program. In the following example, `update` will perform an in-place mutation because it has been called with a variable that and been created in scope.

---

```
update : (1 _ : a) -> a
```

```
let 1 v = a :: b in
  update v
```

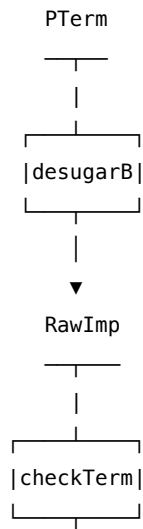
---

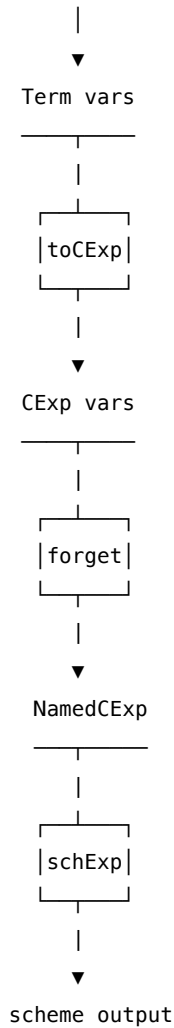
This ensure uniqueness because the linearity of the function prevents the variable from having been shared before it's been used with `update`.

The changes to the Idris AST are as follow:

- Allow constructor to know if they are bound linearly
- Allow linear functions to mutate constructors that are bound linearly
- Add a new instruction to the backend AST in order to mutate values

Conceptually the compiler architecture is quite simple, in the following illustration we link trees transformation with their corresponding functions. Each tree is it's own data declaration:





`P`Term represents the surface level language, it is desugared into `RawImp` which is a front end for our typed terms `Term`. `Term` is indexed over the list of variables in context, which helps avoid indexing errors (specially when manipulating de-brujin indices) and makes explicit the rules of context extension, for example when binding arguments under a lambda. `CExp` is a tree of compiled expressions ready for codegen, it keeps the index for variables from `Term` and the index is then dropped when variables indices are replaced by their names in `NamedCExp`. Which is the same but with all variable substitutions performed.

Since our changes do not touch the type checker or elaborator, they do not change the `checkTerm` stage. However, since we do need information from the

control flow and the access the case tree we need to perform our changes between Term and CExp.

## Implementation details

The goal is to be able to write this program

---

```
%mutating
update : Ty -> Ty
update (ValOnce v) = ValOnce (S v)
update (ValTwice v w) = ValTwice (S v) (S (S w))
```

---

Where the `%mutating` annotation indicates that the value manipulated will be subject to mutation rather than construction.

If we were to write this code in a low-level c-like syntax we would like to go from the non-mutating version here:

---

```
void * update(v * void) {
    Ty* newv;
    if v->tag == 0 {
        newv = malloc(sizeof(ValOnce)); // Allocation here
        newv->tag = 0;
        newv->val1 = 1 + v->val1;
    } else {
        newv = malloc(sizeof(ValTwice)); // Allocation here
        newv->tag = 1;
        newv->val1 = 1 + v->val1;
        newv->val2 = 1 + 1 + v->val2;
    }
    return newv;
}
```

---

To the more efficient mutating version here:

---

```
void * update(v * void) {
    if v->tag == 0 {
        v->val1 = 1 + v->val1;
    } else {
        v->val1 = 1 + v->val1;
        v->val2 = 1 + 1 + v->val2;
    }
    return v; // Return the mutated argument
}
```

---

The two programs are very similar but the second one mutates the argument directly instead of mutating a new copy of it.

There is however a very important limitation:

**We only mutate uses of the constructor we are matching on** The following program would see no mutation

---

```
%mutating
update : Ty -> Ty
update (ValTwice v w) = ValOnce (S v)
update (ValOnce v) = ValTwice (S (S v)) (S (S w))
```

---

Since the constructor we are matching on the left side of the clause does not appear on the right.

This is to avoid implicit allocation when we mutate a constructor which has more fields than the one we are given. Imagine representing data as a records:

---

```
ValOnce = { tag : Int , val1 : Int }
ValTwice = { tag : Int , val1 : Int val2 : Int }
```

---

if we are given a value `ValOnce` and asked to mutate it into a value `ValTwice` we would have to allocate more space to accomodate for the extra `val2` field.

Similarly if we are given a `ValTwice` and are asked to mutate it into a value `ValOnce` we would have to carry over extra memory space that will remain unused. In the worst case, reckless access to the fields in memory would result in accessing out of bound memory and provoke either segfault or memory corruption. For example while attempting to mutate the second argument of a constructor that has only one.

Ideally our compiler would be able to identify data declaration that share the same layout and replace allocation for them by mutation, but for the purpose of this thesis we will ignore this optimisation and carefully design our benchmarks to make use of it. Which brings us to the next section

### Mutating branches

For this to work we need to add a new constructor to the AST that represents *compiled* programs `CExp`. We add the constructor `CMut`:

---

```
CMut : (ref : Name) -> (args : List (CExp vars)) -> CExp vars
```

---

It represents mutation of a variable identified by its `Name` in context and using the argument list to modify each of its fields.

Once this change reached the code generator it needs to output a `mutation` instruction rather than an allocation operation. Here is the code for the scheme backend:

---

```
||| Mutates the given vector at the given index
mutateValue : (ref : String) -> (index : Nat) -> String -> String
mutateValue ref idx newVal =
  "(vector-set! " ++ ref ++ " " ++ show idx ++ " " ++ newVal ++ ")"

||| Mutate all fields of a given constructor
||| We skip the first value in the vector because we do not change
||| @vecRef : The vector to update
||| @args : The list of new arguments
schMutate : (ref : String) -> (args : List String) -> String
schMutate ref args =
  -- we start indexing at 1 since 0 is the tag and doesn't change
  let indices = [1 .. (length args + 1)]
      zipped : List (Nat, String) = zip indices args
      mutation = (showSep " " (map (uncurry $ mutateValue ref) zipped)) in
    "(begin " ++ mutation ++ " " ++ ref ++ ")"
```

---

As you can see we generate one instruction per field to mutate as well and a final instruction to *return* the value passed in argument, this to keep the semantics of the existing assumption about constructing new values. The function `schMutate` is then called to output the scheme program from a `NamedCExp`:

---

```
schExp i (NmMut fc ref args)
  = schMutate (schName ref) <$> (traverse (schExp i) args)
```

---

## Reference nightmare

There is however an additional details that hasn't been expressed and that is how to *construct* a value of `CMut`. `CMut` is only constructed in the following circumstances:

- The function is linear.
- The function is annotated with `%mutating`.

- The function use the same constructor on the left hand side and the right hand side of its definition.
- The value passed in argument is *unique*.

Even though the first one is obvious, it will be left out for trivial reasons: lots of the standard library is not defined linearly such that the following will not compile:

---

```
lolInts : (1 _ : Int) -> Int
lolInts n = n + 1
```

---

This is because `+` as a foreign function is not declared linearly. This could be solved with an alternative standard library which is purely linear.

The second condition is easy to implement as a boolean check, but requires a bit of thought about *where* to put it. Since the next condition requires us to change the *case tree* of our function definition we are going to add the boolean check in `src/TTImp/ProcessDef.idr` which will process every definition of our program. We are going to short-circuit the case-tree generation by replacing the compiled one by our modified version of it:

---

```
(rargs ** (tree_rt', _)) <- getPMDDef ...
tree_rt <- if Mutating `elem` flags gdef
           then makeMutating tree_rt'
           else pure tree_rt'
```

---

The third constraint is the most important one without which we corrupt the memory of our program as we've seen earlier.

Let's look at our `update` function to understand the nature of the changes and change it slightly:

---

```
%mutating
update : (1 _ : Ty) -> Ty
update arg = case arg of
              ValTwice v => ValTwice (S (S v))
              ValOnce v  => ValOnce (S v)
```

---

This version makes use of a temporary variable `arg` before matching on the function argument directly. The calls to constructors `ValTwice` and `ValOnce` allocate memory but we want to replace them by `CMut` which will reuse memory. In order to reuse the memory space taken by `arg` they need to store a pointer

to it. This is reflected in the signature of `CMut : FC -> (ref : Name) -> List (CExp vars) -> CExp vars` where the second argument is the reference to reuse.

---

```

case arg of
  ValTwice v w => -- access `arg` and mutate it with S (S v)
                  ValTwice (S (S v)) (S (S w))
  ValOnce v => -- access `arg` and mutate it with S v
               ValOnce (S v)

```

---

However looking at the AST for pattern matching clauses we see that it does not carry any information about the original value that was matched:

---

```

ConCase : Name -> (tag : Int) -> (args : List Name) ->
               CaseTree (args ++ vars) -> CaseAlt vars

```

---

Indeed, the `Name` we see refers to the constructor we're matching on, and not to the variable name we are inspecting.

Thankfully this reference can be found earlier in `CaseTree` :

---

```

public export
data CaseTree : List Name -> Type where

  Case : {name, vars : _} ->
        (idx : Nat) ->
        (0 p : IsVar name idx vars) ->
        (scTy : Term vars) -> List (CaseAlt vars) ->
        CaseTree vars
  STerm : Int -> Term vars -> CaseTree vars

  Unmatched : (msg : String) -> CaseTree vars

  Impossible : CaseTree vars

```

---

The reference is carried by `idx` which is the index of the variable in its context, as supported by the proof `p`. Once we get a hold of this reference we can move on to the actual tree transformation. The transformation itself is pretty simple and can be summarised with this snippet:

```

replaceConstructor : (cName : Name) -> (tag : Int) ->
  (rhs : Term vars) ->
    Core (Term vars)
replaceConstructor cName tag
  (App fc (Ref fc' (DataCon nref t arity) nm) arg) =
  if cName == nm
  then pure (App fc (Ref fc'
    --
    --
    (DataCon (Just ref) t arity) nm) arg)
  else App fc (Ref fc' (DataCon nref t arity) nm)
  <$> replaceConstructor cName tag arg

```

---

Which will replace every application of a data constructor by one which has a reference to the variable we mutate `ref` as indicated by the `Just ref`.

This refers to another AST change we haven't mentioned yet. And that is to track which data constructors are allowed to mutate a reference instead of constructing a new value. We change `DataCon` into `DataCon : (ref : Maybe Name) -> (tag : Int) -> (arity : Nat) -> NameType` where the first argument represents the variable we can mutate, if there is one. `DataCon` is a field of `Ref`, it indicates what kind of data the `Ref` is pointing to.

The last step of this chain of changes is to inspect our `DataCon` constructor during compilation and emit the correct `CMut` instruction. We do this in `toCExp`:

```

toCExpTm n (Ref fc (DataCon (Just ref) tag arity) fn)
  = pure $ CMut fc ref []

```

---

The final piece of the puzzle is to ensure variables are *unique*, and for this we are going to use the property that linear variables that have been constructed in scope are unique:

```

let 1 v = MkValue 3 in
  f v -- v is unique

```

---

In order to track this change we are going to update our `DataCon` once again to carry the linearity information necessary `DataCon : (rig : RigCount) -> (ref : Maybe Name) -> (tag : Int) -> (arity : Nat) -> NameType`. This step is done when checking let-bindings in `checkLet`. We simply look for this pattern:

---



```

lineariseDataCon : RigCount -> Term vars
                -> Maybe (Term vars)
lineariseDataCon rig
  (App fc (Ref fc' (DataCon r ref tag ary) name) arg) =
    toMaybe
      (rig /= r)
      (App fc (Ref fc' (DataCon rig ref tag ary) name) arg)
lineariseDataCon _ _ = Nothing

```

---

And call the function:

```

-- from the explicit linearity annotation ↴
let newVal = fromMaybe valv --
                      (lineariseDataCon rigb valv)

...

-- ↵ Return the updated constructor
pure (Bind fc n -- ↵
      (Let rigb newVal tyv) scopev,
      gnf env (Bind fc n (Let rigb newVal' tyv) scopet))

```

---

An astute reader might notice that this implementation has multiple shortcomings:

- The unique tag isn't removed from `DataCon` once they leave the `let` scope. Allowing the mutating mechanism to trigger on values that have been shared.
- The `%mutating` flag does not emit compile errors when used on functions called with non-unique variables.
- The `%mutating` flag does not emit compile errors when used on non-linear functions or when the optimisation does not run.
- The optimisation is incompatible with “newtype-optimisation” which removes needless pattern matching and constructing of values when they have a single constructor.

Thankfully, most of those limitations are a matter of user ergonomics and not program performance. Which means we are now ready to benchmark linear programs!

## Benchmarks & methodology

In order to test our performance hypothesis I am going to use a series of programs and run them multiple times under different conditions in order to mea-

sure different aspects of performance. Typically, observing how memory usage and runtime varies depending on the optimisation we use.

Each benchmark will be compared to its control which will be the original compiler without any custom optimisation. The variable we are going to introduce is the new mutation instruction.

It is important to stress that those are *synthetic benchmarks* designed to showcase our improvements, and do not necessarily represent their effect in production software. This has often been the source of many misleading claims about performance and is the topic of endless discussion. The `nofib`[49] project (<https://gitlab.haskell.org/ghc/nofib>) makes the effort to distribute a set of programs more representative of real-world production software, especially for functional languages. In our case, synthetic benchmarks will be enough to tell if linear types provide any benefit worth pursuing.

Additionally, the Idris2 compiler itself was intended to be used as a benchmark by measuring the time it takes to compile itself. However, due to the nature of our optimisation there are not enough opportunities to make noticeable performance improvement, due to the predominantly non-linear style of writing in the compiler.

### Traditional Fibonacci

Our first benchmark is the traditional fibonacci exercise. This will serve as a baseline for our optimisation. We are then going to update this version to include a redundant allocation and attempt to eliminate it using our optimisation. This first version is the one you would expect from a traditional implementation in a functional programming language:

---

```
tailRecFib : Nat -> Int
tailRecFib Z = 1
tailRecFib (S Z) = 1
tailRecFib (S (S k)) = rec 1 1 k
  where
    rec : Int -> Int -> Nat -> Int
    rec prev curr Z = prev + curr
    rec prev curr (S j) = rec curr (prev + curr) j
```

---

As you can see it does not perform any extraneous allocation since it only makes use of primitive values like `Int` which are not heap-allocated. If our optimisation works perfectly, we expect to reach the same performance signature as this

implementation.

### Allocating Fibonacci

This version of Fibonacci does allocate a new value for each call of the `update` function. We expect this version to perform worse than the previous one, both in memory and runtime, because those objects are allocated on the heap (unlike ints), and allocating and reclaiming storage takes more time than mutating values.

---

```
data FibState : Type where
  MkFibState : (prev, curr : Int) -> FibState

next : FibState -> FibState
next (MkFibState prev curr) = MkFibState curr (prev + curr)

rec : FibState -> Nat -> Int
rec (MkFibState prev curr) Z = prev + curr
rec state (S j) = rec (next state) j

tailRecFib : Nat -> Int
tailRecFib Z = 1
tailRecFib (S Z) = 1
tailRecFib (S (S k)) = rec (MkFibState 1 1) k
```

---

### Mutating Fibonacci

This third version is almost the same as the previous one except our `update` function should now avoid allocating any memory, while this adds a function call compared to the first version we do expect this version to have a similar performance profile as the first one

---

```

import Data.List
import Data.Nat

data FibState : Type where
  MkFibState : (prev, curr : Int) -> FibState

%mutating
next : (1 _ : FibState) -> FibState
next (MkFibState prev curr) = MkFibState curr (prev + curr)

tailRecFib : Nat -> Int
tailRecFib Z = 1
tailRecFib (S Z) = 1
tailRecFib (S (S k)) = rec (MkFibState 1 1) k
  where
    rec : FibState -> Nat -> Int
    rec (MkFibState prev curr) Z = prev + curr
    rec state (S j) = rec (next state) j

```

---

## Mapping lists

We know that we can remove the need for intermediate data-structures when mapping them using linear operations[12], but can we reproduce this result here? This benchmark aims to simply map a list from `Int` to `Int`.

```

mapList : (1 _ : List Int) -> List Int
mapList [] = []
mapList (x :: xs) = x + 1 :: mapList xs

main : IO ()
main = println (length (mapList [0 .. 9999]))

```

---

## Allocation free mapping

Similarly to the deforestation algorithm we are going to see if we can improve the performance of linear functions across data structures like lists. When mapping across a list with a function of type `a -> a` that only mutates the input we can avoid allocating an

```

%mutating
mapList : (List Int) -> List Int
mapList [] = []
mapList (x :: xs) = x + 1 :: mapList xs

main : IO ()
main = println (length (mapList [0 .. 9999]))

```

---

A small detail to notice here is that the function is non-linear, indeed using the signature `mapList : (1 _ : List Int) -> List Int` will not compile because `n + 1` does not consume `n` linearly.

### A simple SAT solver as benchmark

Sat solvers themselves aren't necessarily considered "real-world" programs in the same sense that compilers or servers are. However they have two benefits:

- You can make them arbitrarily slow to make the performance improvement very obvious by increasing the size of the problem to solve.
- They still represent a real-life case study where a program need to be fast and where traditional functional programming has fallen short compared to imperative programs.

In this case we are interested in solving the core loop of the state monad that threads through the entire program:

---

```

(>=>) : (1 _ : LState a) -> (1 f : ((_ : a) -> LState b))
      -> LState b
(>=>) (MkLState f) c = MkLState
  (\x => let (res # s') = f x in
        let (MkLState r2) = c res in
        r2 s')

```

---

Our optimisation is subsumed by the "newtype-optimisation" which erases any boxing around types with a single constructor. We can however safely inline this function since `(>=>)` will only be composed with itself, since it takes linear arguments the inlining when composed is safe.

### Measurements

The benchmarks were run with Idris-bench, a command link script written in idris itself which takes a source folder and recursively traverses it in order to find programs to execute and measure their runtime.

The analysis of the result was performed with Idris-stat, another command line script which reads the output of our benchmark program and output data like minimum, maximum, mean and variance along with arrays for plotting our results on a graph.

## Idris-bench

Idris-bench is our benchmarking program and can be found at <https://github.com/andrevidela/idris-bench>.

Idris-bench takes the following arguments:

- `-p | --path IDRI_PATH` the path to the idris2 compiler we want to use to compile our tests. This is used to test the difference between different compiler versions. Typically running the benchmarks with our optimized compiler and running the benchmarks without our optimisation can be done by calling the program with two different versions of the idris2 compiler.
- `-t | --testPath TEST_PATH` The path to the root folder containing the test files.
- `-o | --output FILE_OUTPUT` The location and name of the CSV file that will be written with our results.
  - Alternatively `--stdout` can be given in order to print out the results on the standard output.
- `-c | --count` The number of times each file has to be benchmarked. This is to get multiple results and avoid lucky/unlucky variations.
- `--node` If the node backend should be used instead. If this flag is absent, the Chez backend will be used instead.

## Idris-stats

Once our results have been generated with idris-bench, we are going to analyse them by computing the minimum time, maximum time, mean and variance of the collection of benchmark results. Idris-stats will take our CSV output and compute the data we need and output them as another CSV file. Again the code can be found here <https://github.com/andrevidela/idris-bench/blob/master/script/idris-stats.idr>.

The program takes the file to analyse as single argument. The file is expected to be a CSV file with the following format:

name of first benchmark, result1, result2, result3, etc

name of second benchmark, result1, result2, result3, etc

...

name of final benchmark, result1, result2, result3, etc

The first column is ignored for the purpose of data analysis.

For each row we compute the minimum value, the maximum value, the mean and the variance. As a reminder, the variance is computed as the sum of the square difference with the mean divided by the number of samples.

$$\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$$

### Expected results

Our expectation is that our programs will run faster for 2 reasons:

- Allocation is slower than mutation
- Mutation avoids short lived variables that need to be garbage collected

Indeed allocation will always be slower than simply updating parts of the memory. Memory allocation requires finding a new memory spot that is big enough, writing to it, and then returning the pointer to that new memory address. Sometimes, allocation of big buffers will trigger a reshuffling of the memory layout because the available memory is so fragmented that a single continuous buffer of memory of the right size isn't available.

Obviously all those behaviours are hidden from the programmer through *virtual memory* which allows to completely ignore the details of how memory is actually laid out and shared between processes. Operating systems do a great job at sandboxing memory space and avoid unsafe memory operations. Still, those operations happen, and make the performance of a program a lot less consistent than if we did not have to deal with it.

In addition, creating lots of short lived objects in memory will create memory pressure and trigger garbage collection during the runtime of our program. A consequence of automatic garbage collection is that memory management is now outside the control of the programmer and can trigger at times that are undesirable for the purpose of the program. Real-time applications in particular suffer from garbage collection because it makes the performance of the program hard to predict, an unacceptable trade-off when execution needs to be guaranteed to run within a small time-frame.

## Running the benchmarks

All the benchmarks were run on a laptop with the following specs:

- Intel core-i5 8257U (8th gen), 256KB L2 cache, 6MB L3 cache
- 16Gb of ram at 2133Mhz

While this computer has a base clock of 1.4Ghz, it features a boost clock of 3.9Ghz (a feature of modern CPUs called “turbo-boost”) which is particularly useful for single-core application like ours. However, turbo-boost might introduce an uncontrollable level of variance in the results since it triggers based on a number of parameters that aren’t all under control (like ambient temperature, other programs running, etc). Because of this I’ve disabled turbo boost on this machine and run all benchmarks at a steady 1.4Ghz.

## Results

In this section I will present the results obtained from the compiler optimisation. The methodology and the nature of the benchmarks is explained in the “Benchmarks & methodology” section.

Our first test suite runs the benchmark on our 3 fibonacci variants. As a refresher they are as follow:

- The first one is implemented traditionally, carrying at all times 2 Ints representing the last 2 fibonacci numbers and computing the next one
- Second one boxes those Ints into a datatype that will be allocated every time it is changed
- The Third one will make use of our optimisation and mutate the boxes values instead of discarding the old one and allocating a new one.

The last one is an example of mapping across a list, a very common operation in functional programming languages. Our last one is an example of inlining in a program making use of a linear state monad.

The hypothesis is as follows: Chez is a very smart and efficient runtime, and our example is small and simple. Because of this, we expect a small difference in runtime between those three versions. However, the memory pressure incurred in the second example will trigger the garbage collector to interfere with execution and introduce uncertainty in the runtime of the program. This should translate in our statistical model as a greater variance in the results rather than a strictly smaller mean.



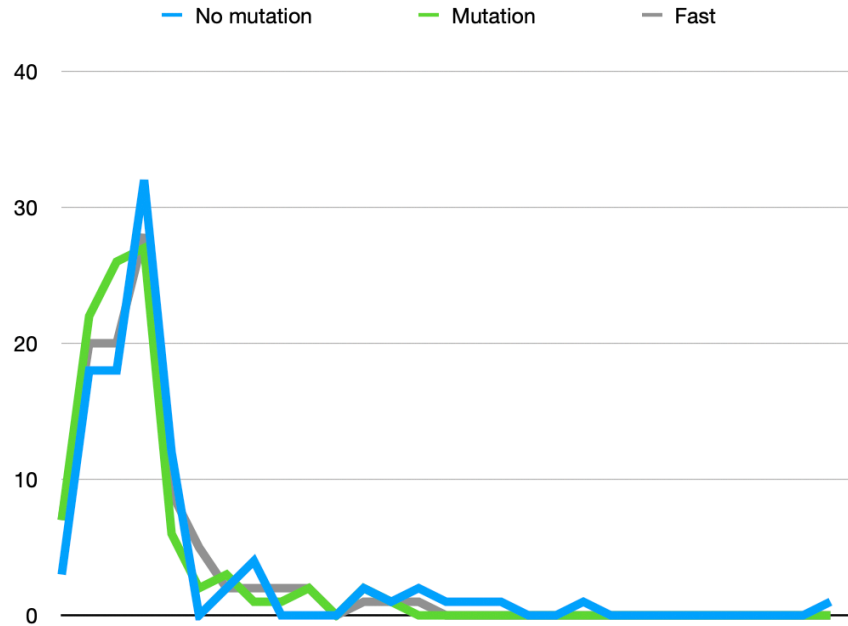
## Results 1: Fibonacci

Here are there results of running our benchmarks 100 times in a row:

---

```
../Idris2_fib_benchmark/fibTestNoMutation.idr,  
5.76e-4,0.00119,  
6.733999999999996e-4,  
1.081849999999999e-8  
  
../Idris2_fib_benchmark/fibTest.idr,  
5.84e-4,  
8.41e-4,  
6.449e-4,  
2.379789999999993e-9  
  
../Idris2_fib_benchmark/fibTailRec.idr,  
5.88e-4,  
8.53e-4,  
6.499100000000001e-4,  
2.692301899999999e-9  
  
[3, 18, 18, 32, 12, 0, 2, 4, 0, 0, 0, 2, 1, 2,  
1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1]  
[7, 22, 26, 27, 6, 2, 3, 1, 1, 2, 0, 2, 1, 0, 0,  
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]  
[7, 20, 20, 28, 9, 5, 2, 2, 2, 2, 0, 1, 1, 1, 0,  
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
```

---



This is the result of calling our data analysis program on the csv file generated by our benchmarking program. The benchmarking program was called with those options

---

```
build/exec/benchmarks -d ../idris2-fib-benchmarks
                      -o results.csv -p idris2dev
                      -c 99
```

---

And the statistical analysis program with no options except for the file:

---

```
build/exec/stats results_mutation_100_attempts_with_startup.csv
```

---

The results are in the following format:

name of benchmark, minimum, maximum, average, variance

The three arrays at the end correspond to an aggregation of the data in “buckets”. Our statistical tool makes 30 “buckets” which represent the different time slots that each benchmark result falls into. The first bucket is the minimum time measured across all benchmarks and the last bucket is the maximum time measured across all benchmarks. There are 28 other buckets in between those two extremities. The array represents the number of results that land for each

bucket.

As you can see the results are pretty consistent with our predictions but the values themselves aren't statistically significant. In order to get a better picture we are going to run the same benchmark 1000 times instead of 100.

---

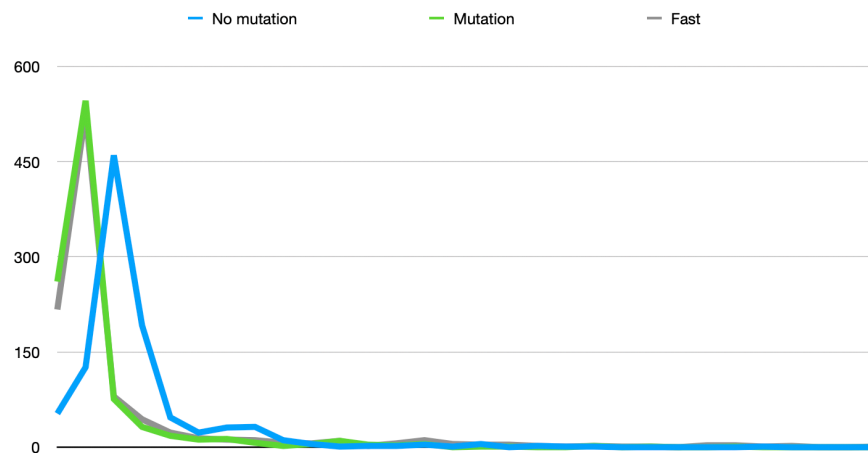
```
../Idris2_fib_benchmark/fibTestNoMutation.idr,
5.08e-4,0.001795,
6.561829999999996e-4,
1.4789385511000005e-8

../Idris2_fib_benchmark/fibTest.idr,
5.08e-4,
0.001753,
5.882930000000001e-4,
1.5392219150999998e-8

../Idris2_fib_benchmark/fibTailRec.idr,
4.89e-4,0.001974,
6.241300000000006e-4,
3.8718697099999986e-8

[53, 126, 460, 192, 47, 23, 31, 32, 11, 5, 1, 2, 2, 4, 1, 5,
0, 2, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
[261, 546, 76, 32, 18, 12, 13, 7, 2, 5, 10, 4, 3, 5, 0,
1, 1, 0, 0, 2, 0, 1, 0, 0, 1, 0, 0, 0, 0]
[217, 527, 80, 44, 23, 14, 12, 11, 7, 6, 10, 2, 6, 11, 5,
4, 4, 2, 1, 2, 1, 1, 0, 3, 3, 1, 2, 0, 0, 1]
```

---



The results are pretty similar which gives us a greater confidence in their

accuracy.

There is however something we can do to improve our measurement and that is to subtract the startup time of the scheme runtime. Indeed every program is measured using the difference between the time it started and the time it ended. But this time also includes the time it takes to launch scheme and then execute a program on it. Indeed the following program:

---

```
main : IO ()  
main = pure ()
```

---

Takes 0.13 seconds to run despite doing nothing. This time is the startup time and can go up to 0.3 seconds.

## Results 2: Fibonacci without startup time

In order to remove the startup time we are going to change the emitted bytecode to wrap our main function inside a time-measuring function (`time main`). Since the timer won't start until the program is ready to run the startup time will be eliminated. Running our empty program we get the following (expected) result:

```
> 0.000000000s elapsed cpu time
```

This time we will run our benchmarks 1000 times using the same command as before. We expect to see the same results but the difference should give us a greater interval of confidence. Running our statistical analysis gives us those results

---

```

../Idris2_fib_benchmark/fibTestNoMutation.idr,
1.696760896,
1.977075901,
1.7447377120060026,
9.18303221644259e-4

```

```

../Idris2_fib_benchmark/fibTest.idr,
1.734708117,
2.152951106,
1.786299514231,
0.002247060292357963

```

```

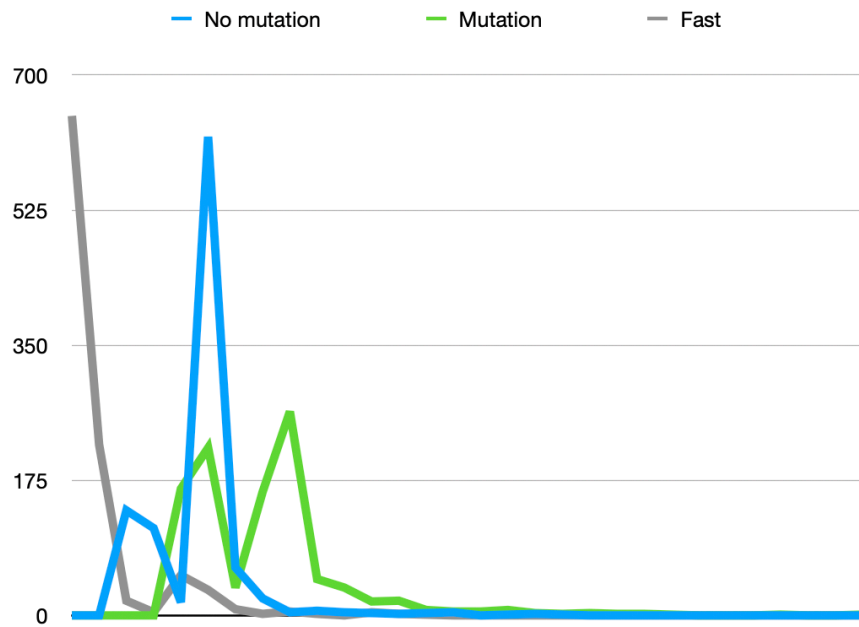
../Idris2_fib_benchmark/fibTailRec.idr,
1.65627213,
1.881412963,
1.6768551703740004,
9.142437018832634e-4

```

```

[0, 0, 136, 113, 17, 620, 62, 22, 4, 6, 4, 3, 2, 3, 4,
0, 1, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
[0, 0, 0, 0, 164, 217, 36, 160, 264, 47, 36, 18, 19, 7, 5,
5, 7, 3, 2, 3, 2, 2, 1, 0, 0, 0, 1, 0, 0, 1]
[647, 221, 19, 4, 52, 33, 8, 2, 5, 2, 0, 4, 2, 1, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

```



As you can see the results aren't exactly as expected, both our *average* and our *variance* is higher than without any optimisation. A result that we definitely did not anticipate and goes against the belief that founded our hypothesis.

One possible explanation is that scheme performs JIT compilation and correctly identifies the hot-loop in our unoptimised example but is unable to perform such optimisation with our mix of pure and mutating code.

### Results 3: Fibonacci without startup time, small loop

In order to test the JIT hypothesis we are going to run the same test, *without* startup time but with a much smaller loop so that the results are measured in milliseconds rather than seconds. This should be enough to prevent the runtime from identifying the loop and performing its optimisation.

In order to reduce the running time from seconds to milliseconds we simply change the loop count from  $8 \cdot 10^6$  to  $8 \cdot 10^4$  reducing it by two orders of magnitude reduces the running time accordingly.

---

```

./Idris2_fib_benchmark/fibTestNoMutation.idr,
0.007216185,0.008861124,
0.007520532116999987,
3.5827851856347296e-8

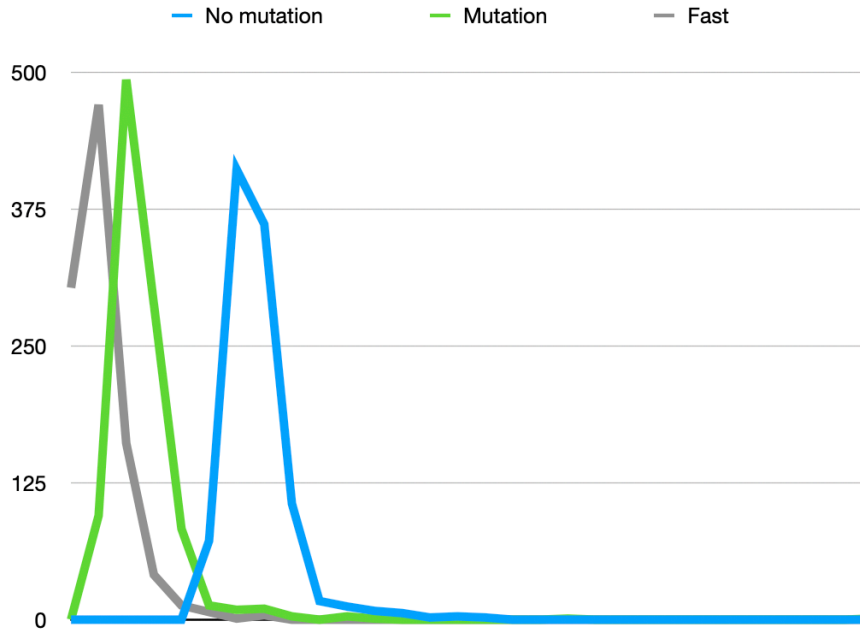
./Idris2_fib_benchmark/fibTest.idr,0.006543267,
0.010942671,
0.006867369243000004,
5.3106313711037986e-8

./Idris2_fib_benchmark/fibTailRec.idr,
0.006385357,
0.007625528,
0.006624731209000001,
2.4002041892751334e-8

[0, 0, 0, 0, 0, 72, 411, 361, 106, 17, 12, 8, 6, 2, 3, 2,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
[0, 95, 493, 288, 83, 13, 9, 10, 3, 0, 3, 1, 0, 0, 0, 0,
0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]
[303, 470, 161, 41, 13, 7, 1, 4, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

```

---



And indeed this data does not disprove our JIT hypothesis (but it does not confirm it either). However, since this thesis is not about the intricacies of the scheme runtime we are going to let the issue rest for now.

Those results showcase two things: That our optimisation works, and that it is not significant enough to be strictly superior to other forms of optimisations. Ideally the best way to test our optimisation would be to write our own runtime which runs on *bare metal* or some approximation of it (WASM/LLVM) which would (probably) be even faster than scheme, and give us more control over which optimisations play nicely together and which ones are redundant or even harmful to performance.

Idris2 has an alternative javascript backend, however, the javascript code generation is unable to translate our programs into plain loops free of recursion. Because of this, our benchmark exceeds the maximum stack size and the program aborts. When the stack size is increased the program segfaults. This could be solved with the use of trampolines in the runtime, instead of plain recursion.

#### Results 4: mapping

In this exercise we are going to aggregate the result of our control and our experimental run of `map`; which increments a list of number and returns its size.

We ran each program a 100 times using `idris-bench`. As before we expect a negligible average speedup but a significant decrease in variance.

---

```

./map_test/map.idr,
8.73337e-4,
0.005067636,
0.0011461037300000004,
2.211481014400371e-7
./map_test_mutate/mapMutation.idr,
5.23816e-4,
0.001308711,
6.850426700000001e-4,
1.9851180862141108e-8

[0, 0, 47, 29, 6, 4, 4, 6, 2, 1, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]
[63, 26, 5, 5, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

```

---

This time our performance improvements are more noticeable in average than before. Tough it might be because the benchmark is measured in tens of microseconds.

## Sat solver

For this benchmark we didn't use `idris-bench` rather we compiled two versions of `isat` (<https://git.sr.ht/~cypheon/idris-minisat>), one with the inlining optimisation, and one without. And ran each version 10 times on the same input (in appendix) :

sat	average	variance
normal	384008.545454545	337134927.872727
mutating	371382.454545455	264041625.672727

The results show an improvement of about 3.3% in average which is encouraging but not enough to warrant further inspection. Despite this small improvement, we could take things further and implement *safe inlining by default* on every variable that is linearly bound, since it does not hurt performance. And hope that the inlined function will be subject to further optimisations after being inlined.



## 7 Future work

Work never ends and progress instils progress, while a lot of ground has been covered, exploring the forest of all knowledge not only yields results, it uncovers new paths to explore. This section outlines additional work that could be done in order to ease the use of linear types as well as make use of them for additional functionality.

### Enlarging the scope

Currently we only look at the *immediate* scope of let bindings. Technically speaking there is nothing preventing our optimisation from working with a more indirect scoping mechanism. Indeed the following should trigger our optimisation

---

```
defaultVal : MyData
defaultVal = MkDefault 3

update : (1 rec : MyData) -> MyData
update (MkDefault n) = MkDefault (S n)
update (MkOther n) = MkOther (S (S n))

operate : MyData
operate = let 1 def = defaultVal
          1 newVal = update def in
          update newVal
```

---

But it will not because `defaultVal` is not a constructor, it's a function call that returns itself a constructor.

One implementation strategy would be to wait for the compiler to inline those definitions and then run our optimiser without further changes.

Another optimisation would be to follow references to see if they result in plain data constructors and replace the entire call chain by the constructor itself, and then run our optimisation.

While both those strategies are valid they incur a cost in terms of complexity and compile time that may not be worth the effort in terms of performance results. They could be hidden behind a `-O3` flag, but that kind of effort is probably better spend in making the ergonomics of linear types more streamlined, which would help make those optimisations more commonplace. Which is the topic of

the next section

## Making linearity easier to use

There are multiple barriers that make linearity harder to use than one might expect. They roughly end up in two buckets:

- I want to use linearity but I cannot
- I have a linear variable and that's actually annoying

### Not linear enough

The first one appears when the programmer tries to make thoughtful usage of linear and erased annotation but finds that other parts of existing libraries do not support linearity. The following program:

---

```
operate : (1 n : Nat) -> (1 m : Nat) -> Int
operate n m = n + m
```

---

Gives the error:

```
> Trying to use linear name n in non-linear context
```

Because the + interface is defined as:

---

```
interface Num ty where
  (+) : ty -> ty -> ty
```

---

Despite addition on Nat being defined linearly:

---

```
plus : (1 n : Nat) -> (1 m : Nat) -> Nat
plus Z m = m
plus (S n) m = S (plus n m)
```

---

A similar problem occurs with interfaces

---

```
interface Monad (Type -> Type) where
  ...

data MyData : (0 ty : Type) -> Type where
  ...

instance Monad MyData where
  ...
```

---

---

```
> Expected Type -> Type
> got (0 ty : Type) -> Type
```

One way to solve those issues would be to have linearity polymorphism and be able to abstract over linearity annotations. For example the map function could be written as

---

```
map : forall l . ((l v : a) -> b) -> (l ls : List a) -> List b
map f [] = []
map f (x :: xs) = f x :: map f xs
```

---

That is, the list is linearly consumed iff the higher order function is linear. What it means for our interface problem is that it could be rewritten as

---

```
interface forall l . Functor (m : (l _ : Type) -> Type) where
  ...
interface forall l . Functor {l} m => Applicative {l} m where
  ...
interface forall l . Applicative {l} m => Monad {l} m where
  ...
```

---

A similar solution could be provided for Num

---

```
interface Num ty where
  (+) : forall l. (l n : ty) -> (l m : ty) -> ty
```

---

So that it can be used with both linear and non-linear variables.

### Too linear now

We've already mentioned before how beneficial it would be for our optimisation strategy to be *100%* linear in every aspect. We also mentioned how this is a problem to implement basic functionality like `drop` and `copy`, but those are artificial examples, rarely does a programmer need to call `copy` or `drop` in industrial applications. In the following example I will show how linear types hold back programming and how QTT can fix it.

A common scenario when debugging effectful code is sprinkling around log statements hoping that running the program will give insight into how it's running.

---

```

do datas <- getData arg1 arg2
  Just success <- trySomething datas (options.memoized)
  | _ => pure $ returnError "couldn't make it work"
  case !(check_timestamp success) of
    Safe t v => functionCall t v
    Unsafe t => trySomethingElse t
    UnSynchronized v => functionCall 0 v
    Invalid => pure $ returnError "failed to check"

```

---

Assuming everything is linear, there is no possible way to add a new print statement without getting a linearity error:

```

do datas <- getData arg1 arg2
  Just success <- trySomething datas (options.memoized)
  | _ => pure $ returnError "couldn't make it work"
  putStrLn $ show success -- <- one use here
  --
  -- And one use there ----|
  --                         v
  case !(check_timestamp success) of
    Safe t v => functionCall t v
    Unsafe t => trySomethingElse t
    UnSynchronized v => functionCall 0 v
    Invalid => pure $ returnError "failed to check"

```

---

There are two solutions for this problem that are worth exploring. The first one is to introduce `Nat` as the multiplicity semiring and write:

```

do datas <- getData arg1 arg2
  2 (Just success) <- trySomething datas (options.memoized)
  | _ => pure $ returnError "couldn't make it work"
  putStrLn $ show success
  case !(check_timestamp success) of
    ...

```

---

Which will allow `success` to be bound with linearity 2 and therefor used twice. We've already seen how multiplicities are useful for memory management, but more flexibility also results in better ergonomics.

The other solutions to this problem would be to allow *borrowing* of read-only variables:

```
do datas <- getData arg1 arg2
  Just success <- trySomething datas (options.memoized)
  | _ => pure $ returnError "couldn't make it work"
  putStrLn $ show &success -- this doesn't count as a use
  case !(check_timestamp success) of
```

---

We've seen that read-only variables have been a topic for linearly typed system with *Linear Types can change the world*[14], where linear read-only variables are allowed within some restriction on their scope. It would be interesting to implement a similar set of rules, adapted for QTT, that allows multiple use of linear types as long as they are read-only (Maybe even reuse our `%mutating` implementation to differentiate between read-only and mutating functions).

## 8 Conclusion

I have learned a lot during this master project. Not only about linear types, quantitative types, graded modal types, containers, categories, optimisation techniques, benchmarking, program safety, compiler design and implementation, etc.

Most importantly I learned about myself and how I best approach research topics. Research is not an activity to carry out alone, it is not a mindless job, and it is not limited to the ivory tower that is higher education. I learned that I do my best research work when I explore the paths that opens in the forest of knowledge after clearing one of them. I would never have understood graded modalities if I wasn't distracted by semirings. I would never have understood semirings if I wasn't frustrated porting some code to Idris2 and noticing the limitations of  $\emptyset$ ,  $1$  and  $\omega$ . I would have never started porting code to Idris2 if I wasn't looking for an escape from writing test for my performance improvements.

While this behaviour is dangerous, in that it can result in nothing of value produced, it also taught me the importance of having a strong and reliable support group. Friends, both in research and otherwise, focus the mind, which results in a more focused work. Finally, research is not something one can *tune out of*, thoughts linger, ideas sprawl out of control and new connections are drawn during every event of daily life. I have discovered that sharing my research through teaching has been an extremely effective mean to contextualise my work and to put boundaries between personal and research life.

Quantitative types, just like me, still have a long way to go. The work displayed here is only a minuscule sliver of what can be done, and needs to be done, for them to become commercially significant. In line with the two main topics of this thesis, I think the areas quantitative types need to improve are ergonomics and performance. They are two pain points that functional programming has failed to fix, while other programming language have been successful basing their entire value proposition upon them (Rust, Javascript, Python, etc).

Ergonomics have made huge strides forward with interactive development environment like we find in Idris, Agda or Coq. But their gimmicks cannot sustain the behemoth of features that commercially successful programming languages showcase. Linear types won't turn the balance around, but they will help in some important aspects: ensuring APIs contracts and protocols are respected, showing those information in the type, and therefore in the documentation, help the compiler generate more helpful error messages, and help guide new users

navigate complex programs by being explicit about usage rules that were left unwritten before.

Until the recent development in resource tracking technologies, it wasn't clear how the concept of performance could be approached and studied in functional programming. We did not have a way to get a hold of it. But just like we did not have a way to get a hold onto side-effect until monads, I strongly suspect quantitative types will allow us to turn performance from an abstract concept into a concrete term in a programming language, just like `IO` became a concrete term in functional programming languages.

I hope you enjoyed this report on linear and quantitative types as much as I did writing it. I wish I explored the paths of knowledge that I discovered with more depth. Yet I am satisfied with the result of this year of exploration. I come out of it with a strong intuition about quantitative types, how to use them and how to improve them. I strongly believe this thesis, as the first detailed exploration of quantitative types used in practice with dependent types, will be the start of a series of great discoveries.

## 9 Appendices: Glossary and definitions

While the initial list of fancy words in the introduction is nice it suffers from being superficial and therefore incomplete. These are more details definitions using examples and imagery

### Linearity / Quantity / Multiplicity

Used interchangeably most of the time. They refer the the number of type a variable is expected to be used.

### Linear types

Linear types describe values that can be used exactly 0 times, exactly 1 time or have no restriction put on them

### Affine types

Affine types describe values that can be used at most 0 times, at most 1 times or at most infinitely many times (aka no restrictions)

### Monad

A mathematical structure that allows to encapsulate *change in a context*. For example `Maybe` is a Monad because it creates a context in which the values we are manipulating might be absent.

### Co-monad / Comonad

A mathematical structure that allows to encapsulate *access to a context*. For example `List` is a Comonad because it allows us to work in a context were the value we manipulate is one out of many available to us, those other values available to us are the other values of the list.

### Semiring

A mathematical structure that requires its values to be combined with `+` and `*` in the ways you expect from natural numbers



## Lattice

A mathematical structure that relates values to one another in a way that doesn't allow arbitrary comparison between two arbitrary values. Here is a pretty picture of one:

As you can see we can't really tell what's going on between X and Y, they aren't related directly, but we can tell that they are both smaller than W and greater than Z

## Syntax

The structure of some piece of information, usual in the form of *text*. Syntax itself does not convey any meaning. Imagine this piece of data

*picture of a circle*

We can define a syntactic rules that allow us to express this circle, here is one: all shapes that you can draw without lifting your pen or making angles. From this definition lots of values are allowed, including |, -, O but not + for example because there is a 90° angle between two bars. Is it supposed to be the letter "O", the number "0" the silhouette of a planet? the back of the head of a stick figure?

## Semantics

The meaning associated to a piece of data, most often related to syntax. From the *syntax* definition if we have

*picture of 10*

we can deduce that the circle means "the second digit of the number 10" which is the number "0". We were able to infer semantics from context. Similarly

*picture of :)*

we can deduce that the meaning of the circle was to represent the head of a stick figure, this time from the front.

**Pattern matching**

**Implicit argument**

**Term/Expression/Value**

## References

- [1] R. Atkey, “Syntax and semantics of quantitative type theory,” in *Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science*, LICS '18, p. 56–65, Association for Computing Machinery, Jul 2018.
- [2] J.-Y. Girard, “Theoretical computer science,” in *Linear logic*, vol. 50, pp. 1–102, 1987.
- [3] C. McBride, *I Got Plenty o’ Nuttin’*, vol. 9600 of *Lecture Notes in Computer Science*, p. 207–233. Springer International Publishing, 2016.
- [4] E. Brady, “Idris, a general-purpose dependently typed programming language: Design and implementation,” *Journal of Functional Programming*, vol. 23, p. 552–593, Sep 2013.
- [5] E. Brady, *Type-driven development with Idris*. Manning, 2017.
- [6] U. Norell, “Interactive programming with dependent types,” in *Proceedings of the 18th ACM SIGPLAN international conference on Functional programming - ICFP '13*, p. 1, ACM Press, 2013.
- [7] J.-Y. Girard, A. Scedrov, and P. J. Scott, “Bounded linear logic: a modular approach to polynomial-time computability,” *Theoretical computer science*, vol. 97, no. 1, pp. 1–66, 1992.
- [8] D. Orchard, V.-B. Liepelt, and H. Eades III, “Quantitative program reasoning with graded modal types,” *Proceedings of the ACM on Programming Languages*, vol. 3, p. 1–30, Jul 2019.
- [9] M. Gaboardi, S.-y. Katsumata, D. Orchard, F. Breuvar, and T. Uustalu, “Combining effects and coeffects via grading,” *ACM SIGPLAN Notices*, vol. 51, p. 476–489, Dec 2016.
- [10] A. L. Smirnov, “Graded monads and rings of polynomials,” *Journal of Mathematical Sciences*, vol. 151, p. 3032–3051, Jun 2008.

- [11] D. Orchard, T. Petricek, and A. Mycroft, “The semantic marriage of monads and effects,” *arXiv:1401.5391 [cs]*, Jan 2014. arXiv: 1401.5391.
- [12] P. Wadler, “Deforestation: Transforming programs to eliminate trees,” in *European Symposium on Programming*, pp. 344–358, Springer, 1988.
- [13] D. N. Turner, P. Wadler, and C. Mossin, “Once upon a type,” in *Proceedings of the seventh international conference on Functional programming languages and computer architecture*, pp. 1–11, 1995.
- [14] P. Wadler, “Linear types can change the world!,” 1990.
- [15] P. Wadler, “Is there a use for linear logic?,” May 1991.
- [16] J. Chirimar, C. A. Gunter, and J. G. Riecke, “Reference counting as a computational interpretation of linear logic,” *Journal of Functional Programming*, vol. 6, p. 195–244, Mar 1996.
- [17] J. A. Tov and R. Pucella, “Practical affine types,” Jan 2011.
- [18] S. Lindley and C. McBride, “Hasochism: the pleasure and pain of dependently typed haskell programming,” in *Proceedings of the 2013 ACM SIGPLAN symposium on Haskell*, Haskell ’13, p. 81–92, Association for Computing Machinery, Sep 2013.
- [19] J.-P. Bernardy, M. Boespflug, R. R. Newton, S. Peyton Jones, and A. Spiwack, “Linear haskell: practical linearity in a higher-order polymorphic language,” Dec 2017.
- [20] K. Mazurak, J. Zhao, and S. Zdancewic, “Lightweight linear types in system  $\mathbb{F}^*$ ,” in *Proceedings of the 5th ACM SIGPLAN workshop on Types in language design and implementation - TLDI ’10*, p. 77, ACM Press, 2010.
- [21] P. Martin-Löf, *An intuitionistic theory of types*.
- [22] E. Brady, C. McBride, and J. McKinna, *Inductive Families Need Not Store Their Indices*, vol. 3085 of *Lecture Notes in Computer Science*, p. 115–129. Springer Berlin Heidelberg, 2004.
- [23] A. Brunel, M. Gaboardi, D. Mazza, and S. Zdancewic, *A Core Quantitative Coeffect Calculus*, vol. 8410 of *Lecture Notes in Computer Science*, p. 351–370. Springer Berlin Heidelberg, 2014.
- [24] D. Orchard, “Should i use a monad or a comonad,” 2012.

- [25] S.-y. Katsumata, “Parametric effect monads and semantics of effect systems,” in *Proceedings of the 41st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages - POPL ’14*, p. 633–645, ACM Press, 2014.
- [26] T. Petricek, D. Orchard, and A. Mycroft, “Coeffects: a calculus of context-dependent computation,” *ACM SIGPLAN Notices*, vol. 49, p. 123–135, Nov 2014.
- [27] A. Vezzosi, A. Mörtberg, and A. Abel, “Cubical agda: a dependently typed programming language with univalence and higher inductive types,” Jul 2019.
- [28] U. Norell, *Towards a practical programming language based on dependent type theory*. Doktorsavhandlingar vid Chalmers Tekniska Högskola, Chalmers Univ. of Technology, 2007.
- [29] T. Altenkirch and C. McBride, “Towards observational type theory,” in *In preparation*, 2006.
- [30] F. Genovese, A. Gryzlov, J. Herold, A. Knispel, M. Perone, E. Post, and A. Videla, “idris-ct: A library to do category theory in idris,” *Electronic Proceedings in Theoretical Computer Science*, vol. 323, p. 246–254, Sep 2020. arXiv: 1912.06191.
- [31] R. Atkey, “Parameterised notions of computation,” *Journal of Functional Programming*, vol. 19, p. 335–376, Jul 2009.
- [32] E. Brady, “State machines all the way down,” ML 2017, Sep 2017.
- [33] K. Matsuda and M. Wang, “Sparcl: a language for partially-invertible computation,” *Proceedings of the ACM on Programming Languages*, vol. 4, p. 1–31, Aug 2020.
- [34] M. Gaboardi, A. Haeberlen, J. Hsu, A. Narayan, and B. C. Pierce, “Linear dependent types for differential privacy,” *ACM SIGPLAN Notices*, vol. 48, p. 357–370, Jan 2013.
- [35] P. Wadler, “Propositions as sessions,” 2012.
- [36] P. Wadler, “There is no substitute for linear logic,” 2003.
- [37] S. Fowler, S. Lindley, and P. Wadler, “Mixing metaphors: Actors as channels and channels as actors,” 2017.

- [38] K. Kokke, J. Morris, and P. Wadler, “Towards races in linear logic,” in *Coordination Models and Languages* (H. Riis Nielson and E. Tuosto, eds.), vol. 11533 of *Lecture Notes in Computer Science*, pp. 37–53, Springer, Cham, Aug. 2019. 21st IFIP WG 6.1 International Conference, COORDINATION 2019, International Conference on Coordination Models and Languages : Held as Part of the 14th International Federated Conference on Distributed Computing Techniques, DisCoTec 2019, Kongens Lyngby, Denmark, June 17â 21, 2019, COORDINATION 2019 ; Conference date: 17-06-2019 Through 21-06-2019.
- [39] F. Genovese, “The essence of petri net gluings,” *arXiv:1909.03518 [cs, math]*, Sep 2019. arXiv: 1909.03518.
- [40] P. Sobocinski, P. W. Wilson, and F. Zanasi, “Cartographer: A tool for string diagrammatic reasoning (tool paper),” p. 7 pages, 2019.
- [41] J. Chapman, P.- . Dagand, C. McBride, and P. Morris, “The gentle art of levitation,” Sep 2010.
- [42] M. Abbott, T. Altenkirch, and N. Ghani, *Categories of Containers*, vol. 2620 of *Lecture Notes in Computer Science*, p. 23–38. Springer Berlin Heidelberg, 2003.
- [43] M. Abbott, T. Altenkirch, C. McBride, and N. Ghani, “ for data: Differentiating data structures,” *IOS Press*, vol. 65, pp. 1–28, 2005.
- [44] T. Altenkirch, N. Ghani, P. Hancock, C. McBride, and P. Morris, “Indexed containers,” *Journal of Functional Programming*, vol. 25, p. e5, 2015.
- [45] Statebox, *Typedefs*.
- [46] N. D. Matsakis and F. S. Klock, “The rust language,” in *Proceedings of the 2014 ACM SIGAda annual conference on High integrity language technology - HILT ’14*, p. 103–104, ACM Press, 2014.
- [47] A. S. Al-Sibahi, *The Practical Guide to Levitation*. PhD thesis, IT University Copenhagen, Sep 2014.
- [48] K. Matsuda, “Modular inference of linear types for multiplicity-annotated arrows,” *arXiv:1911.00268 [cs]*, Feb 2020. arXiv: 1911.00268.
- [49] W. Partain, *The nofib Benchmark Suite of Haskell Programs*, p. 195–202. Workshops in Computing, Springer London, 1993.

## 10 Appendix

### Sat solver benchmark input

```
c FILE: aim-50-1_6-yes1-4.cnf
c
c SOURCE: Kazuo Iwama, Eiji Miyano (miyano@cscu.kyushu-u.ac.jp),
c         and Yuichi Asahiro
c
c DESCRIPTION: Artifical instances from generator by source.  Generators
c               and more information in sat/contributed/iwama.
c
c NOTE: Satisfiable
c
p cnf 50 80
16 17 30 0
-17 22 30 0
-17 -22 30 0
16 -30 47 0
16 -30 -47 0
-16 -21 31 0
-16 -21 -31 0
-16 21 -28 0
-13 21 28 0
13 -16 18 0
13 -18 -38 0
13 -18 -31 0
31 38 44 0
-8 31 -44 0
8 -12 -44 0
8 12 -27 0
12 27 40 0
-4 27 -40 0
12 23 -40 0
-3 4 -23 0
3 -23 -49 0
3 -13 -49 0
-23 -26 49 0
12 -34 49 0
```

-12 26 -34 0  
 19 34 36 0  
 -19 26 36 0  
 -30 34 -36 0  
 24 34 -36 0  
 -24 -36 43 0  
 6 42 -43 0  
 -24 42 -43 0  
 -5 -24 -42 0  
 5 20 -42 0  
 5 -7 -20 0  
 4 7 10 0  
 -4 10 -20 0  
 7 -10 -41 0  
 -10 41 46 0  
 -33 41 -46 0  
 33 -37 -46 0  
 32 33 37 0  
 6 -32 37 0  
 -6 25 -32 0  
 -6 -25 -48 0  
 -9 28 48 0  
 -9 -25 -28 0  
 19 -25 48 0  
 2 9 -19 0  
 -2 -19 35 0  
 -2 22 -35 0  
 -22 -35 50 0  
 -17 -35 -50 0  
 -29 -35 -50 0  
 -1 29 -50 0  
 1 11 29 0  
 -11 17 -45 0  
 -11 39 45 0  
 -26 39 45 0  
 -3 -26 45 0  
 -11 15 -39 0  
 14 -15 -39 0

```

14 -15 -45 0
14 -15 -27 0
-14 -15 47 0
17 17 40 0
1 -29 -31 0
-7 32 38 0
-14 -33 -47 0
-1 2 -8 0
35 43 44 0
21 21 24 0
20 29 -48 0
23 35 -37 0
2 18 -33 0
15 25 -45 0
9 14 -38 0
-5 11 50 0
-3 -13 46 0
-13 -41 43 0

```

## Sat solver benchmark output

```

,,,,,,,,,average,variance,,,,,,,,,
normal,5m 51s 878ms,5m 52s 27ms,6m 29s 173ms,6m 35s 147ms,6m 34s 151ms,6m 35s 888ms,6m 36s 815ms,6m 35s 96
mutat,6m 27s 894ms,6m 21s 214ms,6m 10s 131ms,6m 27s 206ms,6m 4s 71ms,6m 25s 495ms,6m 27s 899ms,6m 13s 771m
normal milisec,351878,352027,389173,395147,394151,395888,396815,395960,393967,394964,364124,384008.54
mutating milisec,387894,381214,370131,387206,364071,385495,387899,373771,351541,347685,348300,371382.
normal milisec * 10^5,3.51878,3.52027,3.89173,3.95147,3.94151,3.95888,3.96815,3.9596,3.93967,3.94964,3
mutating milisec * 10^5,3.87894,3.81214,3.70131,3.87206,3.64071,3.85495,3.87899,3.73771,3.51541,3.47685

```