



DEPARTMENT OF COMPUTER SCIENCE

CircuitFlow: A Domain Specific Language for Dataflow Programming

Riley Evans

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Engineering in the Faculty of Engineering.

Wednesday 19th May, 2021

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MEng in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Riley Evans, Wednesday 19th May, 2021

Acknowledgements

I would like to thank the following people for their help and support:

- Jamie Willis for giving me pointers on Haskell techniques to look at, and convincing me to keep going even when I thought it wasn't possible.
- My flatmates Chris & Jack for their support.
- My supervisor Meng, who noticed the value of indexed functors and encouraged me complete a project based on them.
- Finally, Sam who has provided tips and ideas that have been invaluable throughout this project, as well as taking the time to proof-read this work.

Contents

1	Introduction	1
1.1	Contributions	1
1.2	Outline	2
2	Background	3
2.1	Dataflow Programming	3
2.1.1	The Benefits	4
2.1.2	Dataflow Diagrams	4
2.1.3	Kahn Process Networks (KPNs)	5
2.2	Domain Specific Languages (DSLs)	5
2.2.1	Deep Embeddings	5
2.2.2	Shallow Embeddings	6
2.3	Higher Order Functors	6
2.3.1	Monadic Catamorphism with IFunctors	8
2.4	Data types à la carte	8
2.5	Dependently Typed Programming	10
2.5.1	DataKinds Language Extension	10
2.5.2	Singletons	10
2.5.3	Type Families	11
2.5.4	Heterogeneous Lists	11
2.5.5	Summary	13
2.6	Existential Types	13
2.7	Phantom Type Parameters	13
2.8	Monoidal Resource Theories	14
2.8.1	Preorders	14
2.8.2	Symmetric Monoidal Preorders	14
2.8.3	Wiring Diagrams	14
3	The Language	17
3.1	Language Requirements	17
3.2	Tasks	17
3.2.1	Data Stores	17
3.2.2	Task Constructor	18
3.3	Chains, A Dalliance	18
3.3.1	Trees as Chains	19
3.3.2	Evaluation	20
3.4	Solution: Circuit	21
3.4.1	Constructors	21
3.4.2	Combined Data Stores	23
3.4.3	Multi-Input Tasks	24
3.4.4	mapC operator	25
3.4.5	Completeness	25
3.4.6	Evaluation	25

4	Implementation	27
4.1	Requirements	27
4.2	Circuit AST	27
4.2.1	IFunctor	27
4.2.2	Indexed Data Types à la Carte	28
4.3	Process Network	29
4.3.1	Network Typeclass	29
4.3.2	The Basic Network Representation	30
4.4	Translation to a Network	30
4.4.1	Indexed Monadic Catamorphism	30
4.4.2	BuildNetworkAlg	31
4.4.3	The Translation	33
4.5	UUIDS	37
4.5.1	Modifications	37
4.5.2	Helper Functions	38
4.6	Failure in the Process Network	38
4.6.1	Maybe not Maybe	38
4.6.2	Except Monad	38
4.7	Evaluation	39
4.7.1	Requirements	39
4.7.2	Development Practices	40
5	Examples	43
5.1	Machine Learning (Audio Playlist Generation)	43
5.1.1	Building the pre-processing Circuit	43
5.1.2	Building the prediction Circuit	46
5.2	Build System (lhs2TeX)	47
5.2.1	Building the Circuit	47
5.2.2	Using the Circuit	48
5.3	Types saving the day	49
6	Benchmarks	53
6.1	Benchmarking Technicalities	53
6.2	Parallel vs Serial	54
6.2.1	Areas for improvement	54
6.2.2	An Aside: 1 Core Circuit vs Serial	55
6.3	CircuitFlow vs Luigi	55
6.3.1	Why is CircuitFlow so good?	56
6.3.2	Why is CircuitFlow so good? (v2)	56
7	Conclusion	57
7.1	Related Work	57
7.1.1	Existing Workflow Tools	57
7.1.2	Indexed Functors	57
7.2	Future Work	57
7.3	Summary	58

List of Figures

2.1	Luigi dependency graph [14]	3
2.2	Quartz composer [7]	4
2.3	An example dataflow and its imperative approach.	4
2.4	A sequence of node firings in a KPN	5
2.5	An example wiring diagram	14
3.1	A Pipe (a) and its corresponding dataflow diagram (b).	19
3.2	The constructors in the Circuit library alongside their graphical representation.	21
3.3	A graphical representation of a task with multiple dependencies	24
5.1	A dataflow diagram for playlist generation	43
5.2	A dataflow diagram for pre-processing the song data	44
5.3	The first layer (a) and its corresponding dataflow diagram (b).	45
5.4	The second layer (a) and its corresponding dataflow diagram (b).	45
5.5	The third layer (a) and its corresponding dataflow diagram (b).	46
5.6	An example config file for the lhs2TeX build system	49
5.7	A Broken Luigi Example	50
5.8	A Broken Circuit Example	51
6.1	Speedup provided by using CircuitFlow, relative to a serial implementation	54
6.2	A comparison of runtimes of a serial implementation vs single core CircuitFlow	55
6.3	A comparison of runtimes of CircuitFlow vs Luigi	55

Abstract

With the ever growing amount of data collected every day, new approaches are needed to build data science applications. Its desirable to have assurances that these applications will work correctly, as they can run for days. However, current dataflow tools are not good enough, too often the interactions between tasks are not made in a type-safe way, leading to undesirable runtime errors.

This thesis presents a new Haskell **e-DSL**, with a strong mathematical basis, refocusing on how data flows through an application, resulting in a more expressive solution that is faster. Monoidal resource theories are used to define a symmetric monoidal preorder on data that models dependencies in the workflow. It uses `DataKinds` and `TypeFamilies` in phantom type parameters to bake each axiom into the language constructors. Then using indexed functors and `Data` types à la carte to represent the **AST**, and makes use of principled recursion schemes to perform a type-safe translation to a Kahn Process Network (**KPN**).

Acronyms

Throughout this thesis, several acronyms will be used:

DSL Domain Specific Language

e-DSL Embedded DSL

FIFO First-In First-Out

KPN Kahn Process Network

GPL General Purpose Language

DPN Data Process Network

DAG Directed Acyclic Graph

AST Abstract Syntax Tree

PID Process Identifier

WHNF Weak Head Normal Form

Chapter 1

Introduction

It is estimated that every day over 2.5 quintillion bytes of data is generated [17] by people around the world. With such vast quantities of data, being able to process it is now becoming more important than ever before.

Applications that process this data and perform analysis, such as machine learning, tend to be a series of tasks that need computing. Tasks have a set of inputs and produce some outputs. These tasks form data workflows in a Directed Acyclic Graph (DAG). There are existing libraries that make use of the DAG structure to orchestrate tasks, such as the transformation of data or training a machine learning model: Spotify’s Luigi [28] or Apache’s Airflow [2].

These DAGs encode dependencies between each task, but these libraries have no mechanism to ensure the dependencies are valid. For example, in Luigi a task can output a file delimited by commas, but the dependent task may expect a file delimited by new lines. This bug, will cause the pipeline to crash at runtime, which is not desirable. Even worse, the bug could go unnoticed and cause havoc in later tasks. The libraries put the onus on the user to make sure these inputs and outputs of a task align and do not cause errors. With good development practices, the risk is reduced, but if the problem will never be eliminated. If only there was something that could help...

Any tool that could perform this type of static analysis could help speed up the feedback loop, and possibly avoid buggy code from being deployed to production environments. Once such tool is types! However, both of these libraries are developed using Python — the antithesis of types. This promotes the need for a new library that can safely compose tasks and make use of types to perform static analysis to ensure that dependencies are valid.

Both Luigi [28] and Airflow [2], apply techniques that describe how to compose tasks. Luigi takes an object-oriented approach, defining a class for each task and having a property defining any requirements of the task being defined. Airflow defines tasks and then composes them using an arrow like syntax.

This thesis flips the approach on its head: rather than focus on how to compose tasks together, it defines a language that describes how data flows through a workflow.. The language draws its origins from monoidal resource theory [6], which is used to reason about problems involving resources. Resources have similarities to data workflows, where the inputs and outputs of tasks are thought of as resources. It will then be possible to ask questions about these systems, such as are they valid — i.e. do they type check?

When a system is valid, it will need to be executed to perform meaningful computation. This thesis will translate this language and construct a Kahn Process Network (KPN), that executes the workflow in parallel, to provide the speed benefits of multi-core processors. The KPN will be capable of handling an exception in a task, without causing the full network to crash, allowing computation to continue after for successive inputs.

1.1 Contributions

- A new Haskell e-DSL for constructing dataflow programs that:
 - is able to exceed the performance of other competing libraries — outperforming Luigi by almost 4x on larger numbers of inputs.
 - employs state of the art DSL design techniques, using indexed data types à la carte and principled recursion, to provide interpretations for the AST.

- leverages state of the art Haskell methods, to produce a type-safe implementation: DataKinds, TypeFamilies, Singletons, HLists, and phantom types.
 - makes use of indexed functors, extended to support multiple indices, to construct a type-indexed **AST**. Then used in conjunction with an indexed monadic catamorphism to provide a type-safe translation to a **KPN**.
 - has a strong mathematical grounding in monadic resource theories, providing confidence that the language can represent all dataflow diagrams.
- Provides a range of examples that demonstrates the language’s applicability to many different problems.
 - The first known implementation of a Kahn Process Network in Haskell.

1.2 Outline

The story begins with Chapter 2, where dataflow programming is defined and the benefits that it brings to the table are described. The chapter also defines all the Haskell goodies needed to understand and implement the CircuitFlow **DSL**.

As already hinted, there are many differing approaches to the new design of a data workflow **DSL**. In Chapter 3, two possible implementations are described: Chains and Circuits. They are both evaluated against a desirable criteria for the design of the language.

Once settled on a good design for the language, it needs to be implemented: this will be described in Chapter 4. It will outline how a circuit is constructed under the hood, using a deep embedding, that allows for multiple interpretations to be given to the circuit. Then it describes a method that can be used for transforming a circuit into a process network in a type-safe way, using principled recursion schemes. With a working network in place, it will describe how a network will be modified to capture errors and propagate them through the network. Finally, the implementation is evaluated against a set of criteria.

Now with a working language, two different example applications will be explored in Chapter 5, working through their construction step-by-step and describing each step of the translation. A final example pits CircuitFlow against Luigi, demonstrating CircuitFlow’s ability to catch errors during compilation that would have caused runtime errors in Luigi.

Chapter 6, will aim to compare the performance of CircuitFlow against Luigi and serial implementations. It then investigates deeper into the possible reasons for the results that are obtained.

Chapter 2

Background

2.1 Dataflow Programming

Dataflow programming is a paradigm that models applications as a directed graph. The nodes of the graph have inputs and outputs and are pure functions, therefore have no side effects. It is possible for a node to be a: source; sink; or processing node. A source is a read-only storage: it can be used to feed inputs into processes. A sink is a write-only storage: it can be used to store the outputs of processes. Processes will read from either a source or the output of another process, and then produce a result which is either passed to another process or saved in a sink. Edges connect these nodes together, and define the flow of information.

Example - Data Pipelines A common use of dataflow programming is in pipelines that process data. This paradigm is particularly helpful as it helps the developer to focus on each specific transformation on the data as a single component. Avoiding the need for long and laborious scripts that could be hard to maintain. One example of a data pipeline tool that makes use of dataflow programming is Luigi [28]. An example dataflow graph produced by the tool is shown in Figure 2.1

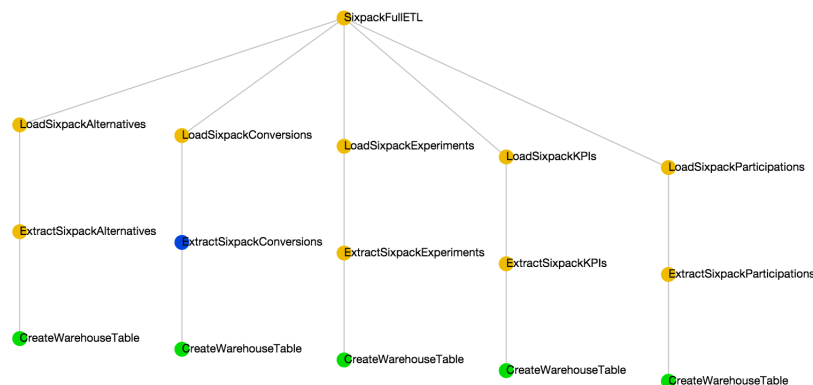


Figure 2.1: Luigi dependency graph [14]

Example - Quartz Composer Apple developed a tool included in XCode, named Quartz Composer, which is a node-based visual programming language [1]. As seen in Figure 2.2, it uses a visual approach to programming connecting nodes with edges. This allows for quick development of programs that process and render graphical data, without the user having to write a single line of code. This means that even non-programmers are able to use the tool.

Example - Spreadsheets A widely used example of dataflow programming is in spreadsheets. A cell in a spreadsheet can be thought of as a single node. It is possible to specify dependencies to other cells through the use of formulas. Whenever a cell is updated it sends its new value to those who depend on it, and so on. Work has also done to visualise spreadsheets using dataflow diagrams, to help debug ones that are complex [12].

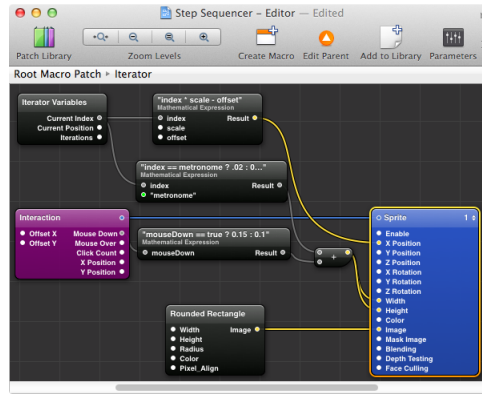


Figure 2.2: Quartz composer [7]

2.1.1 The Benefits

Visual The dataflow paradigm uses graphs, which make programming visual. It allows the end-user programmer to see how data passes through the program, much easier than in an imperative approach. In many cases, dataflow programming languages use drag and drop blocks with a graphical user interface to build programs. For example, Tableau Prep [32], that makes programming more accessible to users who do not have programming skills.

Implicit Parallelism Moore's law states that the number of transistors on a computer chip doubles every two years [24]. This meant that the chips' processing speeds also increased in alignment with Moore's law. However, in recent years this is becoming harder for chip manufacturers to achieve [4]. Therefore, chip manufactures have had to turn to other approaches to increase the speed of new chips, such as multiple cores. It is this approach the dataflow programming can effectively make use of. Since each node in a dataflow is a pure function, it is possible to parallelise implicitly. No node can interact with another node, therefore there are no data dependencies outside of those encoded in the dataflow. Thus eliminating the ability for a deadlock to occur.

2.1.2 Dataflow Diagrams

Dataflow programs are typically viewed as a graph. An example dataflow graph along with its corresponding imperative approach, can be found in Figure 2.3. The nodes 100, X , and Y are sources as they are only read from. C is a sink as it is wrote to. The remaining nodes are all processes, as they have some number of inputs and compute a result.

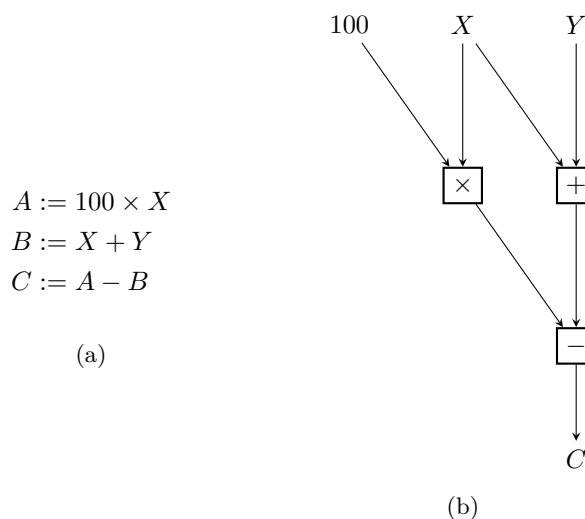


Figure 2.3: An example dataflow and its imperative approach.

In this diagram is possible to see how implicit parallelisation is possible. Both A and B can be calculated simultaneously, with C able to be evaluated after they are complete.

2.1.3 Kahn Process Networks (KPNs)

A method introduced by Gilles Kahn, Kahn Process Networks (KPNs) realised the concept of dataflow networks through the use of threads and unbounded First-In First-Out (FIFO) queues [16]. The FIFO queue is one where the items are output in the same order that they are added. A node in the dataflow becomes a thread in the process network. Each FIFO queue represents the edges connecting the nodes in a graph. The threads are then able to communicate through FIFO queues. The node can have multiple input queues and is able to read any number of values from them. It will then compute a result and add it to an output queue. Kahn imposed a restriction on a process in a KPNs that the thread is suspended if it attempts to fetch a value from an empty queue. The thread is not allowed to test for the presence of data in a queue.

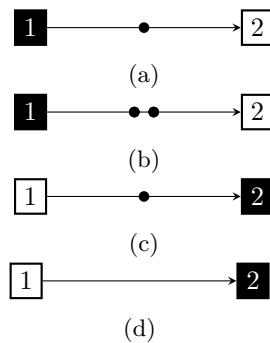


Figure 2.4: A sequence of node firings in a KPN

Data Process Networks (DPNs) Parks described a variant of KPNs, called Data Process Networks (DPNs) [21]. They recognise that if functions have no side effects then they have no values to be shared between each firing. Therefore, a pool of threads can be used with a central scheduler instead.

2.2 Domain Specific Languages (DSLs)

A DSL is a programming language that has a specialised domain or use-case. This differs from a General Purpose Language (GPL), which can be applied across a larger set of domains, and are generally turing complete. HTML is an example of a DSL: it is good for describing the appearance of websites, however, it cannot be used for more generic purposes, such as adding two numbers together.

Approaches to Implementation DSLs are typically split into two categories: standalone and embedded. Standalone DSLs require their own compiler and typically their own syntax; HTML would be an example of a standalone DSL. Embedded DSLs (e-DSLs) use an existing language as a host, therefore they use the syntax and compiler from the host. This means that they are easier to maintain and often quicker to develop than standalone DSLs. An e-DSL, can be implemented using two differing techniques: deep and shallow embeddings.

2.2.1 Deep Embeddings

A deep embedding is when the terms of the DSL will construct an Abstract Syntax Tree (AST) as a host language datatype. Semantics can then be provided later on with evaluation functions. Consider the example of a minimal non-deterministic parser combinator library [35], which will be a running example for this chapter.

```
data Parserd (a :: Type) where
  Satisfyd :: (Char → Bool) → Parserd Char
  Ord      :: Parserd a → Parserd a → Parserd a
```

This can be used to build a parser that can parse the characters 'a' or 'b'.

```
aorbd :: Parserd Char
aorbd = Satisfyd (≡ 'a') `Ord` Satisfyd (≡ 'b')
```

However, this parser does not have any semantics, therefore this needs to be provided by the evaluation function `parse`.

```
parsed :: Parserd a → String → [(a, String)]
parsed (Satisfyd p) = λcase
  []      → []
  (t : ts') → [(t, ts') | p t]
parsed (Ord px py) = λts → parsed px ts ++ parsed py ts
```

The program can then be evaluated by the `parsed` function. For example, `parsed aorbd "a"` evaluates to `[(a, "")]`, and `parsed aorbd "c"` evaluates to `[]`.

A key benefit for deep embeddings is that the structure can be inspected, and then modified to optimise the user code: Parsley [34] makes use of such techniques to create optimised parsers. Another benefit, is that you can provide multiple interpretations, by specifying different evaluation functions. However, they also have drawbacks - it can be laborious to add a new constructor to the language. Since it requires that all functions that use the deep embedding be modified to add a case for the new constructor [30].

2.2.2 Shallow Embeddings

In contrast, a shallow approach is when the terms of the **DSL** are defined as first class components of the language. For example, a function in Haskell. Components can then be composed together and evaluated to provide the semantics of the language. Again a simple parser example can be considered.

```
newtype Parsers a = Parsers { parses :: String → [(a, String)] }
ors :: Parsers a → Parsers a → Parsers a
ors (Parsers px) (Parsers py) = Parsers (λts → px ts ++ py ts)
satisfys :: (Char → Bool) → Parsers Char
satisfys p = Parsers (λcase
  []      → []
  (t : ts') → [(t, ts') | p t])
```

The same `aorbs` parser can be constructed from these functions, avoiding the need for an intermediate **AST**.

```
aorbs :: Parsers Char
aorbs = satisfys (≡ 'a') `ors` satisfys (≡ 'b')
```

Using a shallow implementation has the benefit of being able add new ‘constructors’ to a **DSL**, without having to modify any other functions. Since each ‘constructor’, produces the desired result directly. However, this causes one of the main disadvantages of a shallow embedding - the structure cannot be inspected. This means that optimisations cannot be made to the structure before evaluating it.

2.3 Higher Order Functors

It is possible to capture the shape of an abstract datatype as a **Functor**. The use of a **Functor** allows for the specification of where a datatype recurses. Consider an example on a small expression language:

```
data Expr = Add Expr Expr
          | Val Int
```

The recursion within the `Expr` datatype can be removed to form `ExprF`. The recursive steps can then be specified in the **Functor** instance.

```
data ExprF f = AddF f f
              | ValF Int
```

```
instance Functor ExprF where
  fmap f (AddF x y) = AddF (f x) (f y)
  fmap f (ValF x)   = ValF x
```

To regain a datatype that is isomorphic to the original datatype, the recursive knot need to be tied. This can be done with `Fix`, to get the fixed point of `ExprF`:

```
data Fix f = In (f (Fix f))
type Expr' = Fix ExprF
```

There is, however, one problem: a `Functor` expressing the a parser language is required to be typed. Parsers require the type of the tokens being parsed. For example, a parser reading tokens that make up an expression could have the type `Parser Expr`. A `Functor` does not retain this type information needed in a parser.

IFunctors Instead a type class called `IFunctor` [22] — also known as `HFunctor` [15] — can be used, which is able to maintain the type indicies. This makes use of \leadsto , which represents a natural transformation [20] from `f` to `g`. `IFunctor` can be thought of as a functor transformer: it is able to change the structure of a functor, whilst preserving the values inside it. Whereas a functor changes the values inside a structure.

```
type ( $\leadsto$ ) f g =  $\forall a. f a \rightarrow g a$ 
class IFunctor iF where
  imap :: (f  $\leadsto$  g)  $\rightarrow$  iF f  $\leadsto$  iF g
```

The shape of `Parser` can be seen in `ParserF` where the `f` marks the recursive spots. The type `f` represents the type of the children of that node. In most cases this will be itself.

```
data ParserF (f :: *  $\rightarrow$  *) (a :: *) where
  SatisfyF :: (Char  $\rightarrow$  Bool)  $\rightarrow$  ParserF f Char
  OrF      :: f a  $\rightarrow$  f a  $\rightarrow$  ParserF f a
```

An `IFunctor` instance can be defined, which follow the same structure as a standard `Functor` instance.

```
instance IFunctor ParserF where
  imap _ (SatisfyF s) = SatisfyF s
  imap f (OrF px py) = OrF (f px) (f py)
```

`Fix` is used to get the fixed point of a `Functor`, to get the indexed fixed point `IFix` can be used.

```
newtype IFix iF a = In (iF (IFix iF) a)
```

The fixed point of `ParserF` is `Parserfixed`.

```
type Parserfixed = IFix ParserF
```

In a deep embedding, the **AST** can be traversed and modified to make optimisations, however, it may not be the best representation when evaluating it. This means that it might be transformed to a different representation. In the case of a parser, this could be a stack machine. Now that the recursion in the datatype has been generalised, it is possible to create a mechanism to perform this transformation. An indexed *catamorphism* is one such way to do this, it is a generalised way of folding an abstract datatype. The use of a catamorphism removes the recursion from any folding of the datatype. This means that the algebra can focus on one layer at a time. This also ensures that there is no re-computation of recursive calls, as this is all handled by the catamorphism. The commutative diagram below describes how to define a catamorphism, that folds an `IFix iF a` to a `f a`.

$$\begin{array}{ccc}
\text{iF (IFix iF) a} & \xrightarrow{\text{imap (icata alg)}} & \text{iF f a} \\
\text{inop} \uparrow \downarrow \text{In} & & \downarrow \text{alg} \\
\text{IFix iF a} & \xrightarrow{\text{icata alg}} & \text{f a}
\end{array}$$

`icata` is able to fold an `IFix iF a` and produce an item of type `f a`. It uses the algebra argument as a specification of how to transform a single layer of the datatype.

```
icata :: IFunctor iF => (iF f ~> f) -> IFix iF ~> f
icata alg (Iln x) = alg (imap (icata alg) x)
```

The resulting type of `icata` is `f a`, therefore the `f` has kind `* -> *`. This could be `IFix ParserF`, which would be a transformation to the same structure, possibly applying optimisations to the [AST](#).

2.3.1 Monadic Catamorphism with IFunctors

Using an indexed catamorphism, allows for principled recursion and makes it easier to define a fold over a data type, as any recursive step is abstracted from the user. However, there may be times when there is a need for monadic computations in the algebra. To be able to do this a monadic catamorphism [10] is defined:

```
cataM :: (Traversable f, Monad m) => (forall a. f a -> m a) -> Fix f -> m a
cataM algM (Iln x) = algM <=< mapM (cataM algM) x
```

This catamorphism follows a similar pattern to a standard catamorphism, however, it upgrades the `Functor` constraint to `Traversable`, which still requires that `f` is a `Functor`, but also provides additional functions such as a monadic map — `mapM :: Monad m => (a -> m b) -> f a -> m (f b)`. This allows the monadic catamorphism to be applied recursively on the data type being folded.

This technique can also be applied to indexed catamorphisms to gain a monadic version [3], however, to do so an indexed monadic map has to be introduced. This will be included as part of the `IFunctor` type class:

```
class IFunctor iF where
  imap :: (f ~> g) -> iF f ~> iF g
  imapM :: Monad m => (forall a. f a -> m (g a)) -> iF f a -> m (iF g a)
```

`imapM` is the indexed equivalent of `mapM`, it performs a natural transformation, but is capable of also using monadic computation.

The new `IFunctor` instance for `ParserF` is defined as:

```
instance IFunctor ParserF where
  imap = ... -- previously defined in this section.
  imapM _ (SatisfyF s) = return $ SatisfyF s
  imapM f (OrF px py) = do
    px' <- f px
    py' <- f py
    return (OrF px' py')
```

The definition for `imapM` on `ParserF` is intuitively the same, however just uses `do`-notation instead. Making use of `imapM`, `icataM` is defined to be:

```
icataM :: (IFunctor iF, Monad m) => (forall a. iF f a -> m (f a)) -> IFix iF a -> m (f a)
icataM algM (Iln x) = algM <=< imapM (icataM algM) x
```

`icataM`, has a similar structure to all other catamorphisms defined, however it takes a monadic algebra, that can be used to transform the structure of the input type.

2.4 Data types à la carte

When building a [DSL](#) one problem that becomes quickly prevalent, the so called *Expression Problem* [33]. The expression problem is a trade off between a deep and shallow embedding. In a deep embedding, it is easy to add multiple interpretations to the [DSL](#) - just add a new evaluation function. However, it is not easy to add a new constructor, since all functions will need to be modified to add a new case for the constructor. The opposite is true in a shallow embedding.

One possible attempt at fixing the expression problem is *Data types à la carte* [31]. It combines constructors using the co-product of their signatures. This technique makes use of standard functors, however, an approach using higher-order functors is described in *Compositional data types* [3].

This is defined as:

```
data (iF :+: iG) f a = L (iF f a) | R (iG f a)
```

It is also the case that if both f and g are IFunctors then so is the sum $f :+: g$.

```
instance (IFunctor iF, IFunctor iG)  $\Rightarrow$  IFunctor (iF :+: iG) where  
  imap f (L x) = L (imap f x)  
  imap f (R y) = R (imap f y)
```

For each constructor it is possible to define a new data type and a **Functor** instance specifying where it recurses. This allows for the modularisation of the parser example:

```
data SatisfyF2 f a where  
  SatisfyF2 :: (Char  $\rightarrow$  Bool)  $\rightarrow$  SatisfyF2 f Char  
data OrF2 f a where  
  OrF2 :: f a  $\rightarrow$  f a  $\rightarrow$  OrF2 f a  
instance IFunctor SatisfyF2 where  
  imap f (SatisfyF2 g) = SatisfyF2 g  
instance IFunctor OrF2 where  
  imap f (OrF2 px py) = OrF2 (f px) (f py)
```

By using IFix to tie the recursive knot, the $\text{IFix (SatisfyF}_2 :+: \text{OrF}_2)$ data type would be isomorphic to the original Parser_d datatype found in Section 2.2.1.

One problem that now exists, however, is that it is now rather difficult to create expressions. Revisiting the simple example of a parser for 'a' or 'b'.

```
exampleParser :: IFix (SatisfyF2 :+: OrF2) Char  
exampleParser = lIn (R (OrF2 (lIn (L (SatisfyF2 ( $\equiv$  'a'))))) (lIn (L (SatisfyF2 ( $\equiv$  'b'))))))
```

It would be beneficial if there was a way to add these L s and R s automatically. Fortunately, there is a method using injections. The $:-:$ type class captures the notion of subtypes between IFunctors .

```
class (IFunctor iF, IFunctor iG)  $\Rightarrow$  iF  $:-:$  iG where  
  inj :: iF f a  $\rightarrow$  iG f a  
instance IFunctor iF  $\Rightarrow$  iF  $:-:$  iF where  
  inj = id  
instance (IFunctor iF, IFunctor iG)  $\Rightarrow$  iF  $:-:$  (iF :+: iG) where  
  inj = L  
instance (IFunctor iF, IFunctor iG, IFunctor iH, iF  $:-:$  iG)  $\Rightarrow$  iF  $:-:$  (iH :+: iG) where  
  inj = R  $\cdot$  inj
```

Using this type class, smart constructors are defined:

```
inject :: (iG  $:-:$  iF)  $\Rightarrow$  iG (IFix iF) a  $\rightarrow$  IFix iF a  
inject = lIn  $\cdot$  inj  
satisfy2 :: (SatisfyF2  $:-:$  iF)  $\Rightarrow$  (Char  $\rightarrow$  Bool)  $\rightarrow$  IFix iF Char  
satisfy2 f = inject (SatisfyF2 f)  
or2 :: (OrF2  $:-:$  iF)  $\Rightarrow$  IFix iF a  $\rightarrow$  IFix iF a  $\rightarrow$  IFix iF a  
or2 px py = inject (OrF2 px py)
```

Expressions can now be built using the constructors, such as $\text{satisfy}_2 (\equiv \text{'a'}) \text{`or}_2 \text{satisfy}_2 (\equiv \text{'b'})$.

A modular algebra can now be defined that provides an interpretation of this datatype.

```
newtype Size a = Size {unSize :: Int} deriving Num  
class IFunctor iF  $\Rightarrow$  SizeAlg iF where  
  sizeAlg :: iF Size a  $\rightarrow$  Size a  
instance (SizeAlg iF, SizeAlg iG)  $\Rightarrow$  SizeAlg (iF :+: iG) where  
  sizeAlg (L x) = sizeAlg x  
  sizeAlg (R y) = sizeAlg y
```

```

instance SizeAlg OrF2 where
  sizeAlg (OrF2 px py) = px + py
instance SizeAlg SatisfyF2 where
  sizeAlg (SatisfyF2 f) = 1
eval :: SizeAlg iF ⇒ IFix iF a → Size a
eval = icata sizeAlg

```

The main benefit of this approach is modularity. Each constructor is given by its interpretation in isolation and only for interpretations that make sense for it. Additionally, existing interpretations are not affected by the addition of new constructors, such as ApF_2 . This helps to solve the expression problem.

2.5 Dependently Typed Programming

Although Haskell does not officially support dependently typed programming, there are techniques available that together can be used to replicate some of the experience.

2.5.1 DataKinds Language Extension

Through the use of the DataKinds language extension [36], all data types can be promoted to also be kinds and their constructors to be type constructors. When constructors are promoted to type constructors, they are prefixed with a `'`. This allows for more interesting and restrictive types.

Consider the example of a vector that also maintains its length. Peano numbers can be used to keep track of the length, which prevents a negative length for a vector. This is where numbers are defined as zero or a number n incremented by 1.

```

data Nat = Zero
        | Succ Nat

```

A vector type can now be defined that makes use of the promoted `Nat` kind.

```

data Vec :: Type → Nat → Type where
  Nil  :: Vec a 'Zero
  Cons :: a → Vec a n → Vec a ('Succ n)

```

The use of DataKinds can enforce stronger types. For example a function can now require that a specific length of vector is given as an argument. With standard lists, this would not be possible, which could result in run-time errors when the incorrect length is used. For example, getting the head of a list. Getting the head of an empty list an error will be thrown. For a vector, a `safeHead` function can be defined that will not type check if the vector is empty.

```

safeHead :: Vec a ('Succ n) → a
safeHead (Cons x _) = x

```

2.5.2 Singletons

DataKinds are useful for adding extra information back into the types, but how can information be recovered from the types? For example, could a function that gets the length of a vector be defined?

```

vecLength :: Vec a n → Nat

```

This is enabled through the use of singletons [9]. A singleton in Haskell is a type that has just one inhabitant. That is that there is only one possible value for each type. They are written in such a way that pattern matching reveals the type parameter. For example, the corresponding singleton instance for `Nat` is `SNat`. The structure for `SNat` closely flows that of `Nat`.

```

data SNat (n :: Nat) where
  SZero :: SNat 'Zero
  SSucc  :: SNat n → SNat ('Succ n)

```


A function that fetches the length of a vector can now be definable.

```
vecLength2 :: Vec a n → SNat n
vecLength2 Nil      = SZero
vecLength2 (Cons x xs) = SSucc (vecLength2 xs)
```

Recovering an SNat Although being able to define a function that can recover the length of a vector is great, there is a more general way this can be approached. This is to define a new type class that is able to recover an `SNat` from any type level `Nat`:

```
class IsNat (n :: Nat) where
  nat :: SNat n
```

The type class has one value inside it `nat`, which can produce an `SNat` for a type level `Nat`. There are two instances from this type class: a base case and a recursive case.

```
instance IsNat 'Zero where
  nat = SZero

instance IsNat n ⇒ IsNat ('Succ n) where
  nat = SSucc nat
```

The base case matches on the type level `Nat 'Zero`, in this case `nat` is defined to be `SZero` — the singleton equivalent. The recursive step deals with the `'Succ n` case, where the singleton equivalent `SSucc` is used to define `nat`.

A new vector length function can then be defined as:

```
vecLength3 :: IsNat n ⇒ Vec a n → SNat n
vecLength3 _ = nat
```

2.5.3 Type Families

Now consider the possible scenario of appending two vectors together. How would the type signature look? This leads to the problem where two type-level `Nats` need to be added together. This is where Type Families [27] become useful, they allow for the definition of functions on types. Consider the example of appending two vectors together, this would require type-level arithmetic — adding the lengths together.

```
vecAppend :: Vec a n → Vec a m → Vec a (n :+ m)
```

This requires a `:+` type family that can add two `Nats` together.

```
type family (a :: Nat) :+(b :: Nat) where
  a :+'Zero    = a
  a :+'Succ b = 'Succ (a :+b)
```

2.5.4 Heterogeneous Lists

Heterogeneous lists [18] are a way of having multiple types in the same list. Rather than be parameterised by a single type, they instead make use of a type list, which is the list type promoted through `DataKinds` to be a kind, with its elements being types. Each element in the type list aligns with the value at that position in the list, giving its type. A heterogeneous list is defined as:

```
data HList (xs :: [Type]) where
  HNil  :: HList '[]
  HCons :: x → HList xs → HList (x ': xs)
```

This data type has two constructors:

- `HNil` represents the empty list. The type parameter is the empty type list `'[]`
- `HCons` allows a new element to be added to the list. The type parameter is the type of the item inserted consed onto the front of the types of the tail of the list.

Functions on HLists

Length Using singletons and type families, it is possible to get the length of a HList in a type-safe way. Firstly, a type family is defined that is able to return the length of a type list.

```
type family Length (l :: [k]) :: Nat where
  Length '[]      = 'Zero
  Length (e ': l) = 'Succ (Length l)
```

Length follows a similar definition to the `length :: [a] → Int` function defined in the Prelude:

```
length :: [a] → Int
length []      = 0
length (x : xs) = 1 + length xs
```

The base case of **Length** defines the length to be `'Zero`. The recursive case increments the length by 1 for each item in the list, until it reaches the base case.

Now a function is defined that returns the length of a HList:

```
lengthH :: HList xs → SNat (Length xs)
lengthH HNil      = SZero
lengthH (HCons _ xs) = SSucc (lengthH xs)
```

This follows the same structure as the **Length** type family, however instead, of working with types it uses singleton values.

Take Another function that may be helpful with HLists is **take**. This will return the first `n` items from the list. If `n` is larger than the length of the list, then the whole list will be returned. Again, to be able to do this a new type family is needed – **Take**:

```
type family Take (n :: Nat) (l :: [k]) :: [k] where
  Take 'Zero    l      = '[]
  Take ('Succ n) '[]    = '[]
  Take ('Succ n) (e ': l) = e ': Take n l
```

The type family follows the same definition as the standard `take :: Int → [a] → [a]` as defined in the Prelude. Similar to the `lengthH` function, `takeH` follows the same structure as the type family:

```
takeH :: SNat n → HList xs → HList (Take n xs)
takeH SZero    l      = HNil
takeH (SSucc n) HNil    = HNil
takeH (SSucc n) (HCons x xs) = HCons x (takeH n xs)
```

Drop The final function used in this project on Hlists is one that can drop the first `n` elements. The **Drop** type family can be defined as:

```
type family Drop (n :: Nat) (l :: [k]) :: [k] where
  Drop 'Zero    l      = l
  Drop ('Succ _) '[]    = '[]
  Drop ('Succ n) (_ ': l) = Drop n l
```

The **Drop** type family also closely follows the definition of `drop :: Int → [a] → [a]` from the Prelude, and its result is reflected in the values level just as with `lengthH` and `takeH`.

```
dropH :: SNat n → HList xs → HList (Drop n xs)
dropH SZero l = l
dropH (SSucc _) HNil = HNil
dropH (SSucc n) (HCons _ xs) = dropH n xs
```

2.5.5 Summary

Together these features allow for dependently typed programming constructs in Haskell:

- DataKinds allow for values to be promoted to types
- Singletons allow types to be demoted to values
- Type Families can be used to define functions that manipulate types.

2.6 Existential Types

Typically, when defining a data type in Haskell, every type variable that exists on the right hand side of the equals, must also be on the left hand side. For example, this is not allowed:

```
newtype Bad = Bad a
```

Existential types [23] are a way to allow this to happen, for example:

```
data Good =  $\forall a$ . Good a
```

One benefit of existential types is that the type variable no longer needs to be on the left hand side of the equals.

There is however, one problem with this approach, the variable will be a random unknown type. To avoid this problem constraints are typically added to the signature, so that there can be a set of functions that work with that variable. For example, the variable could have the `Show` constraint, so that we are able to use the `show` function with it:

```
data Showy =  $\forall a$ . Show a  $\Rightarrow$  Showy a
```

It would now be possible to build a list of items that can use the `show` function.

```
showList :: [Showy]  
showList = [Showy 123, Showy "abc"]
```

This list can now store any value with a `Show` instance defined, by wrapping it in the `Showy` constructor.

2.7 Phantom Type Parameters

Phantom type parameters [5] could be considered the opposite of existential types. This is when a type variable only appears on the left hand side of the equals. The most basic example is `Const`, it has two type arguments, but only `a` is used on the right hand side:

```
newtype Const a b = Const a
```

Phantom type parameters can be used to store information in the types, which can act as further static constraints on the types. Consider an example revolving around locking doors: it should not be possible to lock a door that is open, first it has to be closed and then it can be locked. The state of the door can be represented by a data type that is promoted to a kind with the `DataKinds` extension. A door can be represented as a type with a phantom type variable, with kind `DoorState` that record the state of the door:

```
data DoorState = Open | Closed | Locked  
data Door (state :: DoorState) where  
  Door :: Door state
```

It would then be possible to define functions that can close and lock doors:

```
closeDoor :: Door 'Open  $\rightarrow$  Door 'Closed  
lockDoor :: Door 'Closed  $\rightarrow$  Door 'Locked
```

The `closeDoor` function, enforces that only an open door can be given as input, similarly `lockDoor` prevents an open door from being locked.

2.8 Monoidal Resource Theories

Resource theories [6] are a branch of mathematics that allow for the reasoning of questions surrounding resources, for example:

- If I have some resources, can I make something?
- If I have some resources, how can I get what I want?

Resource theories provide a way to answer these questions.

2.8.1 Preorders

A preorder relation on a set X is denoted by \leq . The relation must obey two laws:

1. Reflexivity — $x \leq x$
2. Transitivity — if $x \leq y$ and $y \leq z$ then $x \leq z$.

A preorder is a pair (X, \leq) made up of a set and a preorder relation on that set. Preorders can represent many different things, such as, the less than relationship on numbers, or even dependencies between different values. For example, $A \leq B$ would mean that A is required to calculate B .

2.8.2 Symmetric Monoidal Preorders

To be a symmetric monoid then the following laws must be satisfied:

1. Monotonicity — $\forall x_1, x_2, y_1, y_2 \in X$, if $x_1 \leq y_1$ and $x_2 \leq y_2$, then $x_1 \otimes x_2 \leq y_1 \otimes y_2$
2. Unitality — $\forall x \in X$, $I \otimes x = x$ and $x \otimes I = x$
3. Associativity — $\forall x, y, z \in X$, $(x \otimes y) \otimes z = x \otimes (y \otimes z)$
4. Symmetry — $\forall x, y \in X$, $x \otimes y = y \otimes x$

A symmetric monoidal structure (X, \leq, I, \otimes) on a preorder (X, \leq) has two additional components:

1. The monoidal unit — an element $I \in X$
2. The monoidal product — a function $\otimes : X \times X \rightarrow X$

One example is where X is a collection of resources, \otimes combines resources together, and \leq defines dependencies between resources.

2.8.3 Wiring Diagrams

A graphical representation of symmetric monoidal preorders is a wiring diagram. A wiring diagram is made up of: boxes that can have multiple inputs and outputs. The boxes can be arranged in series or in parallel. Figure 2.5, shows an example wiring diagram.

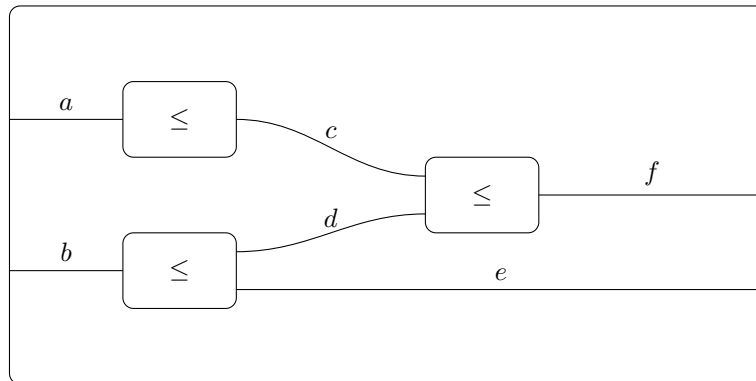
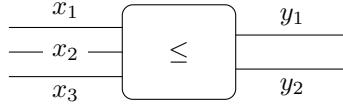


Figure 2.5: An example wiring diagram

A wiring diagram formalises a symmetric monoidal preorder, with each element $x \in X$ existing as the label on a wire. Two wires, x and y , drawn in parallel are considered to be the monoidal product $x \otimes y$. The monoidal unit is defined as a wire with the label I or no wire.

$$\frac{x}{y}$$

A box connects parallel wires on the left to parallel wires on the right. A wiring diagram is considered valid if the monoidal product of the left is less than the right.



This example wiring diagram corresponds to the inequality $x_1 \otimes x_2 \otimes x_3 \leq y_1 \otimes y_2$, which corresponds to the idea that x_1 , x_2 , and x_3 are required to get y_1 , and y_2 .

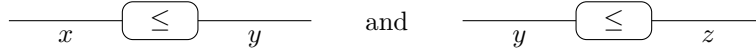
Each axiom in a symmetric monoidal preorder has a corresponding graphical form, using wiring diagrams.

Reflexivity The reflexivity law states that $x \leq x$, this states that a diagram of one wire is valid.

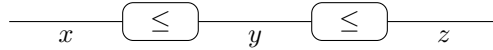
$$\frac{}{x}$$

This law corresponds to the idea that a resource is preserved.

Transitivity The transitivity law says that if $x \leq y$ and $y \leq z$ then $x \leq z$. This corresponds to connecting two diagrams together in sequence. If both of the diagrams



are valid, then they can be joined together to obtain another valid diagram.



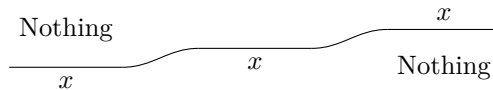
If a box is considered a task that can transform values, then this law corresponds to the idea that two tasks can be composed in sequence, with the output of one being the input to the next.

Monotonicity Monotonicity states that, if $x_1 \leq y_1$ and $x_2 \leq y_2$, then $x_1 \otimes x_2 \leq y_1 \otimes y_2$. This can be thought of as stacking two boxes on top of each other:



This law conceptualises the idea that when resources are combined, the dependencies are respected.

Unitality The unitality law states that $I \otimes x = x$ and $x \otimes I = x$, this means that a blank space can be ignored and that diagrams such as

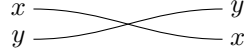


are valid.

Associativity The associativity law says that $(x \otimes y) \otimes z = x \otimes (y \otimes z)$, this states that diagrams can be built from either the top or bottom. This means that the order of grouping resources does not matter. In reality it is trivial to see how this is true with wires:

$$\frac{\frac{x}{\frac{y}{z}}}{\quad} = \frac{\frac{x}{\frac{y}{z}}}{\quad}$$

Symmetry The symmetry law states that $x \otimes y = y \otimes x$, this encodes the notion that a diagram is still valid even if the wires cross.

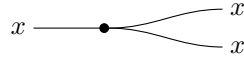


Discard Axiom There are times when there is no longer need to keep a value, it would be beneficial if it could be discarded. In a wiring diagram this is represented as:



This can be added as an additional axiom to the definition of a symmetric monoidal preorder: $\forall x \in X, x \leq I$ It corresponds to the idea that resources can be destroyed when they are no longer needed.

Copy Axiom The final axiom to add is the notion of copying a value: $\forall x \in X, x \leq x + x$. This can be represented in wiring diagram as a split wire:



This embodies the idea that it is possible to duplicate a resource.

Chapter 3

The Language

3.1 Language Requirements

For the design of the language to be considered a success, several criteria need to be met:

- **Describe any dataflow** — The combinators in the language should be able to describe any dataflow that the user needs.
- **Quickness to Learn** — The language should be quick and easy to learn. If it takes users too long to learn, they may never bother, meaning that the language will never be used.
- **Easy to write** — A user should be able to write programs easily. They should not have to spend time creating additional boilerplate code, where it is not necessary.
- **Easy to understand** — Any program written with this language should be easy to understand, so that users can review existing code and know what it will do.
- **Type-safe** — A feature missing in many dataflow tools is the lack of type checking. This causes problems later on in the development process with more debugging and testing needed. The language should be type-safe to avoid any run-time errors occurring where types do not match.

3.2 Tasks

Dataflow programming focuses on the transforming inputs to outputs. To be able to transform inputs this language will make use of tasks. They are responsible for reading from an input data source, completing some operation on the input, then finally writing to an output data sink. Tasks could take many different forms, for example they could be:

- A pure function — a function with type $a \rightarrow b$
- An external operation — interacting with some external system. For example, calling a terminal command.

A task could have a single input or multiple inputs, however, for now just a single input task will be considered. Multi-input tasks are explained further in Sub-Section [3.4.3](#)

3.2.1 Data Stores

Data stores are used to pass values between different tasks, this ensures that the input and output of tasks are closely controlled. They are used to represent different ways of storing values: one example could be a point to a CSV file. By also having just one place that defines how to read and write to data stores, it will reduce the possibility of an error occurring and make it easier to test. A data store can be defined as a type class, with two methods `fetch` and `save`:

```
class DataStore f a where
  fetch :: f a → IO a
  save  :: f a → a → IO (f a)
```

A **DataSource** has two type parameters: where **f** is the type of **DataSource** being used and **a** is the type of the value stored inside it. The aptly named methods describe their intended function: **fetch** will fetch a value from a **DataSource**, and **save** will save a value. The **fetch** method takes a **DataSource** as input and will return the value stores inside. However, the **save** method may not be as self explanatory, since it has an extra **f a** argument. This argument can be thought of as a pointer to a **DataSource**: it contains the information needed to save. For example, in the case of a file store it could be the file name.

By implementing this as a type class, there can be many different implementations of a **DataSource**. The library comes with several pre-defined **DataStores**, such as a **VariableStore**. This can be though of as an in memory storage between tasks.

```
data VariableStore a = Var a | Empty
instance DataSource VariableStore a where
  fetch (Var x) = return x
  save Empty x = return (Var x)
```

The **VariableStore** is the most basic example of a **DataSource**, a more complex example is a **FileStore**, which represents a pointer to a file:

```
newtype FileStore a = FileStore String
instance DataSource FileStore String where
  fetch (FileStore fname) = readFile fname
  save (FileStore fname) x = writeFile fname x >> return (FileStore fname)
instance DataSource FileStore [String] where
  fetch (FileStore fname) = readFile fname >> return . lines
  save (FileStore fname) x = writeFile fname (unlines x) >> return (FileStore fname)
```

The **FileStore** is only defined to store two different types: **String** and **[String]**. If a user attempts to store anything other than these two types then a compiler error will be thrown, for example:

```
ghci> save (FileStore "test.txt") (123 :: Int)
> No instance for (DataSource FileStore Int) arising from a use of ‘save’
```

Although a small set of **DataStores** are included in the library, the user is also able to add new instances of the type class with their own **DataStores**. Some example expansions, could be supporting writing to a database table, or a Hadoop file system.

3.2.2 Task Constructor

The type of a task details the inputs and outputs. A task is created via a constructor that takes two arguments: the function it represents and somewhere to store the output. This constructor makes use of GADTs syntax [26] so that constraints can be placed on the types used. It enforces that a **DataSource** must exist for the input and output types. This allows the task to make use of the **fetch** and **save** functions.

```
data Task (f :: Type → Type) (a :: Type) (g :: Type → Type) (b :: Type) where
  Task :: (DataSource f a, DataSource g b) ⇒ (f a → g b → IO (g b)) → g b → Task f a g b
```

When a **Task** is executed the stored function is executed, with the input being passed in as the first argument and the output “pointer” as the second argument. This returns an output **DataSource** that can be passed on to another **Task**.

3.3 Chains, A Dalliance

In a dataflow programming, one of the key aspects is the definition of dependencies between tasks in the flow. One possible approach to encoding this concept in the language, that was ultimately not up to scratch, is to make use of sequences of tasks — also referred to as chains. These chains compose tasks, based on their dependencies. A chain can be modelled with an abstract datatype:

```
data Chain (f :: Type → Type) (a :: Type) (g :: Type → Type) (b :: Type) where
  Chain :: Task f a g b → Chain f a g b
  Then :: Chain f a g b → Chain g b h c → Chain f a h c
```


The `Chain` constructor wraps a `Task` in the `Chain` data type. This allows for them to be easily composed with other `Tasks`. Without this there would need to be an “empty” element, however, this could lead to the construction of a chain with no tasks. The `Then` constructor combines multiple chains to form a sequence of sequential tasks. To make this easier to use an operator `>>>` is defined that represents `Then`:

```
(>>>) :: Chain f a g b → Chain g b h c → Chain f a h c
(>>>) = Then
```

This can be now be used to join sequences of tasks together, for example:

```
task1 :: Task VariableStore Int      VariableStore String
task2 :: Task VariableStore String  FileStore      [String]
task3 :: Task FileStore      [String] VariableStore Int
sequence :: Chain VariableStore Int VariableStore Int
sequence = Chain task1 >>> Chain task2 >>> Chain task3
```

`sequence` will perform the three tasks in order, starting with `task1` and finishing with `task3`.

3.3.1 Trees as Chains

Now that tasks can be performed in sequence, the next logical step will be to introduce the concept of branching out. This results in a tasks output being given to multiple tasks, rather than just 1.

To do this a new abstract datatype is required. This will be used to form a list of `Chains`, conventionally the `[]` type would be used, however this is not possible as each chain will have a different type. This means that existential types will need to be used.

```
data Pipe where
  Pipe :: ∀ f a g b. (DataStore f a, DataStore g b) ⇒ Chain f a g b → Pipe
  And :: Pipe → Pipe → Pipe
```

The `And` constructor can be used to combine multiple chains together. Figure 3.1 shows the previous sequence, with a new `task4` which also uses the input from `task2`.

```
task4 :: Task FileStore [String] VariableStore String
branchExample :: Pipe
branchExample = Pipe (Chain task1 >>> Chain task2 >>> Chain task3)
                `And`
                Pipe (Chain task2 >>> Chain task4)
```

(a)

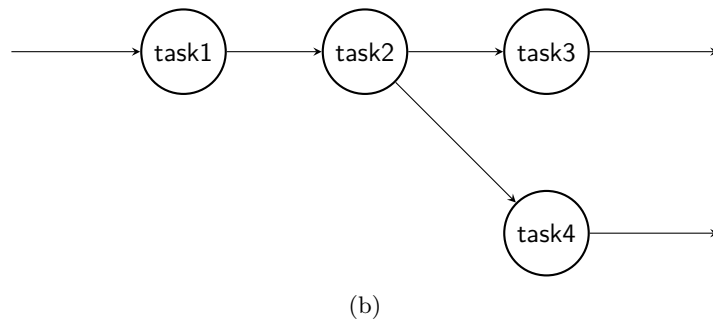


Figure 3.1: A Pipe (a) and its corresponding dataflow diagram (b).

However, there is a problem with this approach: to be able to form a network similar to that shown in Figure 3.1, the language will need to know where to join two `Chains` together. However, with the current definition of a `Task`, it is not possible to easily check the equivalence of two functions. Being able to check for equivalence will be key to defining a method to merge these chains together: it will need to match `task2` in the first chain with the `task2` in the second chain. Similarly, if a user wanted to use the same task multiple times, it would not be possible to differentiate between them. One approach to this would be to have unique identifiers for each task, such as PIDs.

Process Identifiers (PIDs) A Chain can be modified so that instead of storing a Task it instead stores a PID — a unique identifier for a task. However to do this a new PID data type is needed:

```
data PID (f :: Type → Type) (a :: Type) (g :: Type → Type) (b :: Type) where
  PID :: Int → PID f a g b
```

The PID data type has the same kind as a Task and makes use of phantom type parameters, to retain the same type information as a Task, whilst storing just an Int that can be used to identify it.

```
data Chain' (f :: Type → Type) (a :: Type) (g :: Type → Type) (b :: Type) where
  Chain' :: PID → f a g b → Chain' f a g b
  Then' :: Chain' f a g b → Chain' g b h c → Chain' f a h c
```

This new version of Chain named Chain', is almost identical in the way it behaves, however instead of storing a Task, it stores a PID.

This, however, leaves a key question, how do Tasks get mapped to PIDs. This can be done by employing the State monad. The state stores a map from PID to task and a counter for PIDs. As Tasks each have a different type a new wrapper datatype is required, making use of existential types, to close over the types, and produce values of apparently the same type. This is because a Map can only store one type. The Workflow monad is a type alias for the State monad, which stores the WorkflowState.

```
data TaskWrap = ∀f a g b.TaskWrap (Task f a g b)
data WorkflowState = WorkflowState {
  pidCounter :: Int,
  tasks :: M.Map Int TaskWrap
}
type Workflow = State WorkflowState
```

An operation that is defined for the monad is registerTask. This takes a Task and returns a Chain that stores a PID inside it. Whenever a user would like to add a new task to the workflow, they register it. They can then use this returned value to construct multiple chains, which can now be joined easily by comparing the stored PID.

```
registerTask :: Task f a g b → Workflow (Chain' f a g b)
```

One benefit to this approach is that if the user would like to use a task again in a different place, they can simply register it again and use the new PID value.

3.3.2 Evaluation

☑ **Describe any Dataflow** This method would be capable of describing any dataflow, although in its current state, it can only support trees. The algorithm that would be used to join different chains together could be developed to allow them to rejoin onto another chain. This will allow for any DAG to be defined.

☑ **Quickness to Learn** With only 4 different combinators (Chain, >>>, Pipe, And) and the Task constructor, this language should be simple to learn. There are only two main concepts the user needs to understand: how to join tasks into a chain and how to join chains together.

☑ **Easy to Write** Building chains is a very simple process, and allows the user to focus on the dependencies of one task. They do not need to be aware of the bigger picture. For example, adding a new task5 that depends on task2 in Figure 3.1. The user will just need to add one extra chain task5 >>> task2, which will have no effect on the existing chains.

☑ **Easy to Read** Although it is easy to write chains, this could lead to messy definitions with no structure. This will make it harder for a user to interpret an already defined collection of chains. Its possible that they will need to draw out some of the dependencies to further understand what is already defined.

☒ **Type-safe** Although a **Chain** can be well typed, the use of existential types to join chains together pose a problem. This causes the types to be ‘hidden’, this means that when executing these tasks, the types need to be recovered. This is possible through the use of `gcast` from the `Data.Typeable` library. However, this has to perform a reflection at run-time to compare the types. There is the possibility that the types could not matching and this would only be discovered at run-time. There is a mechanism to handle the failed match case, however, this does not fulfil the criteria of being fully type-safe.

Chains are not good enough Chains do not satisfy some of the key requirements that this library set out to solve, such as type safety. This could lead to crashes at run-time as the compiler is not able to fully verify that the system type-checks. Another approach could be to look to category theory, for ways to compose functions together in a type-safe way.

3.4 Solution: Circuit

This approach is inspired by monadic resource theory, which has a collection of mathematical operators for composing functions together. It uses parallel prefix circuits, described by Hinze [13], as a starting point for the design of the combinators. A set of constructors can be defined that are used to represent a **DAG**, which are similar to the wiring diagrams in Section 2.8.3. The constructors seen in Figure 3.2 represent the behaviour of edges in the graph.

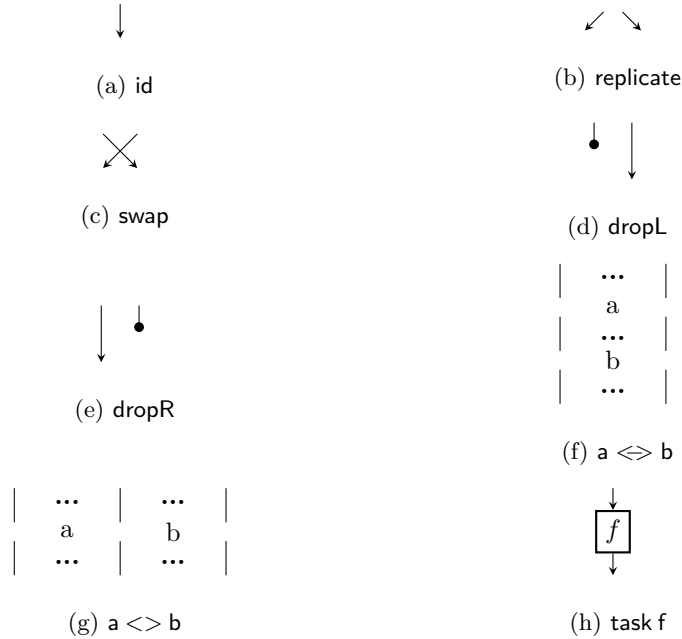


Figure 3.2: The constructors in the `Circuit` library alongside their graphical representation.

3.4.1 Constructors

Each of these constructors use strong types to ensure that they are combined correctly. A `Circuit` has 7 different type parameters:

```
Circuit (inputsStorageTypes :: [Type → Type]) (inputsTypes :: [Type]) (inputsApplied :: [Type])
        (outputsStorageTypes :: [Type → Type]) (outputsTypes :: [Type]) (outputsApplied :: [Type])
        (nInputs :: Nat)
```

A `Circuit` can be thought of as a list of inputs, which are processed and a resulting list of outputs are produced. To represent this it makes use of the `DataKinds` language extension [36], to use type-level lists and natural numbers. Each parameter represents a certain piece of information needed to construct a circuit:

- `inputsStorageTypes` is a type-list of storage types, for example `'[VariableStore, CSVStore]` — these all have kind $* \rightarrow *$.
- `inputsTypes` is a type-list of the types stored in the storage, for example `'[Int, [(String, Float)]]`.
- `inputsApplied` is a type-list of the storage types applied to the types stored, for example `'[VariableStore Int, CSVStore [(String, Float)]]`.
- `outputsStorageTypes`, `outputsTypes` and `outputsApplied` mirror the examples above, but for the outputs instead.
- `nInputs` is a type-level `Nat` that is the length of the input lists.

Although, to a human some of these types may seem irrelevant, GHC is not able to make all of the deductions itself, when type checking the code. It requires additional information, such as `inputsApplied` or `nInputs`.

In the language there are two different types of constructor, those that recurse and those that can be considered leaf nodes. The behaviour of both types of constructor is recorded within the types. For example, the `id` constructor has the type:

```
id :: DataStore' '[f] '[a] => Circuit '[f] '[a] '[f a] '[f] '[a] '[f a] N1
```

It can be seen how the type information for this constructor states that it has 1 input value (`N1 ~ Succ Zero`) of type `f a` and it returns that same value. Each type parameter in `id` is a phantom type [5], since there are no values stored in the data type that use the type parameters. The rest of the constructors that are leaf nodes are: `replicate`, `swap`, `dropL`, and `dropR`:

```
replicate :: DataStore' '[f] '[a] => Circuit '[f] '[a] '[f a] '[f, f] '[a, a] '[f a, f a] N1
swap      :: DataStore' '[f, g] '[a, b] => Circuit '[f, g] '[a, b] '[f a, g b] '[g, f] '[b, a] '[g b, f a] N2
dropL     :: DataStore' '[f, g] '[a, b] => Circuit '[f, g] '[a, b] '[f a, g b] '[g] '[b] '[g b] N2
dropR     :: DataStore' '[f, g] '[a, b] => Circuit '[f, g] '[a, b] '[f a, g b] '[f] '[a] '[f a] N2
```

The `replicate` constructor states that a single input value of type `f a` should be input, and that value should then be duplicated and output. The `swap` constructor takes two values as input: `f a` and `g b`. It will then swap these values over, such that the output will now be: `g b` and `f a`. `dropL` will take two inputs: `f a` and `g b`. It will then drop the left argument and return just a `g b`. The `dropR` has the same behaviour as `dropL`, it just drops the right argument instead.

To be able to make use of the leaf nodes, they need to be combined in some way. To do this two new recursive constructors named ‘beside’ and ‘then’ will be used. However, before defining these constructors there are some tools that are required. This is due to the types no longer being concrete. For example, the input type list is no longer known: it can only be referred to as `fs` and `as`. This means it is much harder to specify the new type of the `Circuit`.

Apply Type Family It would not be possible to use a new type variable `xs` for the `inputsApplied` parameter. This is because it needs to be constrained so that it is equivalent to `fs` applied to `as`. To solve this a new closed type family [8] is created that is able to apply the two type lists together. This type family pairwise applies a list of types with kind $* \rightarrow *$ to a list of types with kind $*$ to form a new list containing types of kind $*$. For example, `Apply '[f, g, h] '[a, b, c] ~ '[f a, g b, h c]`.

```
type family Apply (fs :: [Type -> Type]) (as :: [Type]) where
  Apply '[] '[] = '[]
  Apply (f ': fs) (a ': as) = f a ': Apply fs as
```

Append Type Family There will also be the need to append two type level lists together. Lists would need to be appended in this way, when combining inputs and outputs to form a larger circuit. For example, `'[a, b, c] :++' [d, e, f] ~ '[a, b, c, d, e, f]`. To do this an append type family [18] can be used:

```
type family (:++) (l1 :: [k]) (l2 :: [k]) :: [k] where
  (:++) '[] l = l
  (:++) (e ': l) l' = e ': (l :++ l')
```

The `:++` type family is defined in the same way as the standard `++` function on value lists, however, it appends type lists together instead. This type family makes use of the language extension `PolyKinds` [36] to allow for the append to be polymorphic on the kind stored in the type list. This will avoid defining multiple versions to append `fs` with `gs`, and `as` with `bs`.

The ‘Then’ Constructor This constructor — denoted by `<>` — is used to stack two circuits on top of each other. It is used to encapsulate the idea of dependencies, between different circuits. Through types it enforces that the output of the top circuit is the same as the input to the bottom circuit.

```
(<>) :: (DataStore' fs as, DataStore' gs bs, DataStore' hs cs)
  => Circuit fs as (Apply fs as) gs bs (Apply gs bs) nfs
  => Circuit gs bs (Apply gs bs) hs cs (Apply hs cs) ngs
  => Circuit fs as (Apply fs as) hs cs (Apply hs cs) nfs
```

It employs a similar logic to function composition $(\cdot) :: (a \rightarrow b) \rightarrow (b \rightarrow c) \rightarrow (a \rightarrow c)$. The resulting type from this constructor uses the input types from the first argument `fs as (Apply fs as)`, and the output types from the second argument `hs cs (Apply hs cs)`. It then forces the constrain that the output type of the first argument and the input type of the second are the same — `gs bs (Apply gs bs)`.

The ‘Beside’ Constructor Denoted by `<>`, the beside constructor is used to place two circuits side-by-side. The resulting `Circuit` has the types of left and right circuits appended together.

```
(<>) :: (DataStore' fs as, DataStore' gs bs, DataStore' hs cs, DataStore' is ds)
  => Circuit fs as (Apply fs as) gs bs (Apply gs bs) nfs
  => Circuit hs cs (Apply hs cs) is ds (Apply is ds) nhs
  => Circuit (fs :++ hs) (as :++ cs) (Apply fs as :++ Apply hs cs)
            (gs :++ is) (bs :++ ds) (Apply gs bs :++ Apply is ds)
            (nfs :+ nhs)
```

This constructor works by making use of the `:++` type family to append the input and output type list of the left constructor to those of the right constructor. It also makes use of the `:+` type family — defined in Section 2.5.3 — to add the number of inputs from the left and right together.

3.4.2 Combined Data Stores

A keen eyed reader may notice that all of these constructors have not been using the original `DataStore` type class. Instead they have all used the `DataStore'` type class. This is a special case of a `DataStore`, allowing for constructors to also be defined over type lists, not just a single type. Combined data stores make it easier for tasks to fetch from multiple inputs. Users will just have to call a single `fetch'` function, rather than multiple. However, since tasks can only have one output, there is no need for a `save'` function, that would be able to save to multiple data stores.

To be able to define `DataStore'`, heterogeneous lists [18] are needed — specifically two different forms. `HList'` stores values of type `f a` and is parameterised by two type lists `fs` and `as`. `IOList` stores items of type `IO a` and is parameterised by a type list `as`. Using an `IOList` makes it easier to define a function that produces a list of IO computations. Their definitions are:

```
data HList' (fs :: [Type → Type]) (as :: [Type]) where
  HCons' :: f a → HList' fs as → HList' (f ': fs) (a ': as)
  HNil'  :: HList' '[] '[]

data IOList (xs :: [Type]) where
  IOCons :: IO x → IOList xs → IOList (x ': xs)
  IONil  :: IOList '[]
```

Now that there is a mechanism to represent a list of different types, it is possible to define `DataStore'`:

```
class DataStore' (fs :: [Type → Type]) (as :: [Type]) where
  fetch' :: HList' fs as → IOList as
```

To save the user of the cumbersome task of having to define an instance of `DataStore'` for every possible combination of data stores, the instance is derived from the previous `DataStore` type class. This means

that a user does not need to create any instances of `DataStore'`. They can instead focus on each single case, with the knowledge that they will automatically be able to combine them with other data stores.

```
instance {-# OVERLAPPING #-} (DataStore f a) => DataStore' '[f]' '[a] where
    fetch' (HCons' x HNil') = IOCons (fetch x) IONil
instance (DataStore f a, DataStore' fs as) => DataStore' (f ': fs) (a ': as) where
    fetch' (HCons' x xs)    = IOCons (fetch x) (fetch' xs)
```

One of these instances makes use of the `{-# OVERLAPPING #-}` pragma. In most cases the base case instance that would be defined is `DataStore' '[]' []`. However, it does not make sense to have an empty data store. Therefore, the base case is selected to be a list with one element `DataStore' '[f]' '[a]`. This leads to a problem: GHC is unable to decide which instance to use. It could use either the `DataStore' '[f]' '[a]` or the `DataStore' (f ': fs) (a ': as)` instance. The overlapping pragma tells GHC, that if it encounters this scenario, it should choose the one with the pragma.

3.4.3 Multi-Input Tasks

With a `Circuit`, it is possible to represent a **DAG**. This means that a node in the graph can now have multiple dependencies, as seen in Figure 3.3.

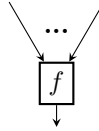


Figure 3.3: A graphical representation of a task with multiple dependencies

To support this, a modification is made to the `task` constructor: rather than have an input value type of `f a`, it can now have an input value type of `HList' fs as`. The function executed in the task can now use `fetch'` to fetch all inputs with one function call.

```
task :: (DataStore' fs as, DataStore g b)
      => (HList' fs as -> g b -> IO (g b))
      -> g b
      -> Circuit fs as (Apply fs as) '[g]' '[b]' '[g b]' (Length fs)
```

Now that the length of the inputs is unknown, in order to specify the `nInputs` type parameter, the `Length` type family defined in Section 2.5.4 must be used. This will return a type-level `Nat`, which is the length of the input array `fs`.

Smart Constructors There could be many times that the flexibility provided by defining your own tasks from scratch could cause a large amount of boiler plate code. For example, there may be times that a user already has pre-defined function and would like to convert it to a task. Therefore there are also two smart constructors that they are able to use:

```
multiInputTask :: (DataStore' fs as, DataStore g b)
                 => (HList as -> b)
                 -> g b
                 -> Circuit fs as (Apply fs as) '[g]' '[b]' '[g b]' (Length fs)
functionTask :: (DataStore f a, DataStore g b)
               => (a -> b)
               -> g b
               -> Circuit '[f]' '[a]' '[f a]' '[g]' '[b]' '[g b]' N1
```

The first allows for a simple function with multiple inputs to be defined. With the fetching and saving handled by the smart constructor. The second allows for a simple `a -> b` function to be turned into a `Task`.

3.4.4 mapC operator

Currently a circuit has a static design — once it has been created it cannot change. There are times when this could be a flaw in the language. For example, when there is a dynamic number of inputs. This could be combated with more smart constructors to generate more complex circuits, with the pre-existing constructors. Another approach would be to add new constructors that allow for more dynamic circuits, such as **mapC**. This new constructor is used to map a circuit on a single input containing a list of items. The input is fed into the inner circuit, accumulated back into a list, and then output.

```
mapC :: (DataStore '[f] '[a]), DataStore g [b])
  => Circuit '[VariableStore] '[a] '[VariableStore a] '[VariableStore] '[b] '[VariableStore b] N1
  -> g [b]
  -> Circuit '[f]                '[[a]] '[f [a]]                '[g]                '[[b]] '[g [b]]                N1
```

This example can be thought of a production line ($[a] \rightarrow [b]$). The circuit given as an argument describes how to produce one item ($a \rightarrow b$). The **mapC** can then be provided with a pallet (or a list) of resources to build multiple items ($[a]$), it will then return a pallet of made items ($[b]$).

3.4.5 Completeness

The constructors in this library make up a symmetric monoidal preorder. For simplicity only the **inputsApplied** and **outputsApplied** type parameters will be used to formalise a **Circuit** — all other type parameters are only required to aid GHC in compilation.

A preorder is defined over tasks and **DataStores**. The preorder relation \leq , can be used to describe the dependencies in the **DataStores**, with a task being able to transform **DataStores** into new **DataStores**. The relation is defined over the set X , which describes the set of all possible **DataStores**.

The monoidal product \otimes can be thought of as the concatenation of multiple **DataStores** into type-lists. For example the monoidal product of $(f\ a) \otimes (g\ b) = '[f\ a] :++ '[g\ b] \sim '[f\ a, g\ b]$. The monoidal unit, is tricky to define as it has no real meaning within a **Circuit**, however it could be considered the empty **DataStore**: $'[]$.

The axioms are then satisfied as follows:

1. Reflexivity — this is the `id :: Circuit '[f a] '[f a]` constructor, it represents a straight line with the same input and output.
2. Transitivity — this is the `<=> :: Circuit x y -> Circuit y z -> Circuit x z` constructor, it allows for circuits to be placed in sequence.
3. Monotonicity — this is the `<> :: Circuit x1 y1 -> Circuit x2 y2 -> Circuit (x1 :++ x2) (y1 :++ y2)` constructor. This can place circuits next to each other.
4. Unitality — given the monoidal unit `'[]` and a **DataStore** `xs`, then the rules hold true: `'[] :++ xs ~ xs` and `xs :++ '[] ~ xs`.
5. Associativity — given three **DataStores**: `xs`, `ys`, `zs`. Since concatenation of lists is associative then this rule holds: `(xs :++ ys) :++ zs ~ xs :++ (ys :++ zs)`.
6. Symmetry — this is the `swap :: Circuit '[f a, g b] '[g b, f a]` constructor, it allows for values to swap over.
7. Delete Axiom — this is satisfied by the `dropL :: Circuit '[f a, g b] '[g b]` and `dropR :: Circuit '[f a, g b] '[f a]`. Although this does not directly fit with the axiom, it also has to ensure the constraint on a circuit that there must always be 1 output value.
8. Copy Axiom — this is the `replicate :: Circuit '[f a] '[f a, f a]` constructor. It allows for a **DataStore** to be duplicated.

By satisfying all the axioms a **Circuit** is a symmetric monoidal preorder.

3.4.6 Evaluation

☑ **Describe any Dataflow** It is possible to represent a **DAG** with the constructors in this library. Its completeness has been formalised as a symmetric monoidal preorder.

☑ **Quickness to Learn** The language uses combinators that also have a visual representation, this makes it easy to quickly understand how they all work. A user can then also benefit from the familiarity of using the host language Haskell.

☑ **Easy to Write** Although a circuit may appear hard to construct initially, the skills of the domain expert also need to be considered. To be able to define a circuit the user needs to have an understanding of the shape of the dataflow diagram: a skill the domain expert is expected to have. Once the user has a sketch for the dataflow they would like to create, translating to a circuit is a relatively simple job. This creates more upfront work for the user, however, it is offset by the additional benefits that a circuit brings.

☑ **Easy to Read** Due to the graphical nature of the constructors, it is relatively simple to build up a picture of how an existing circuit works. The user is able to infer the shape of the dataflow diagram easily by working their way down a circuit from top to bottom, and visualising how different tasks are connected. It could also be possible to pretty print a circuit to recreate the dataflow diagram — although this has not been implemented.

☑ **Type-safe** One key benefit that a `Circuit` brings is that constructing them uses strong types. Each constructor encodes its behaviour within the types. This allows the GHC type checker to validate a `Circuit` at compile-time, to ensure that each task is receiving the correct values. This avoids the possibility of crashes at run-time, where types do not match correctly. There is, however, a consequence of this type-safety: the user now needs to add some explicit types on a `Circuit` to help the type checker.

Chapter 4

Implementation

4.1 Requirements

The implementation of the network itself has several requirements that are separate from the language design:

- **Type-safe** — This is a continuation of the previous requirement for the language. It is also important that once the user has built a well-typed `Circuit`, that the code also continues to be executed in a well-typed environment, to ensure that all inputs and outputs are correctly typed.
- **Parallel** — One of the key benefits that comes from dataflow programming is implicit parallelisation. With this `DSL` being tailored towards data pipelines, which could be computationally expensive, it should be able to benefit from parallel execution.
- **Competitive Speed** — This library should be able to execute dataflows in a competitive time, with other libraries that already exist.
- **Failure Tolerance** — It is important that if one invocation of a task crashes, it does not crash the whole program. This implementation should be able to gracefully handle errors and propagate them through the circuit.
- **Usable** — The implementation of the library should not break any of the usability of the language design.
- **Maintainable** — It should be easy to maintain the library and add new constructors in the future.

4.2 Circuit AST

The constructors for the language are actually *smart constructors*. They provide a more elegant way to build an `AST`, which represents the circuit. They give the ability to gain the benefits of extensibility and modularity, usually found in a shallow embedding, while still having a fixed core `AST` that can be used for interpretation.

4.2.1 IFunctor

To build the fixed core `AST` for a `Circuit`, indexed functors — also known as `IFunctor` — are used. `IFunctors` are, however, only defined to take a single type index — a `Circuit` needs 7. To do this `IFunctor7` can be defined:

```
class IFunctor7 iF where
  imap7 :: (∀ a b c d e f g. f' a b c d e f g → g' a b c d e f g) → iF f' a b c d e f g → iF g' a b c d e f g
```

`IFunctor7` follows a similar structure to a standard `IFunctor`, it just has 7 type indices instead. This indexed functor can be used to mark the recursive points of the data types used to construct the `Circuit AST`.

However, any data type that is converted to use an `IFunctor7`, will need a way to tie the recursive knot. A new type called `IFix7` can be used, it will follow a similar pattern to `IFix`, but with 7 type indices.

```
newtype IFix7 iF a b c d e f g = IIn7 (iF (IFix7 iF) a b c d e f g)
```

4.2.2 Indexed Data Types à la Carte

To build the **AST**, the data types à la carte [31] approach is taken. This allows for a modular approach, making the library more extendable later on. To be able to use this approach, it needs to be modified to support the 7 type indices — a through to g:

```
data (iF :: iG) (f' :: i → j → k → l → m → n → o → Type)
    (a :: i) (b :: j) (c :: k) (d :: l) (e :: m) (f :: n) (g :: o) where
    L :: iF f' a b c d e f g → (iF :: iG) f' a b c d e f g
    R :: iG f' a b c d e f g → (iF :: iG) f' a b c d e f g
infixr ::+:
```

As mentioned in Section 2.4, using the `::+` operator comes with problem of many L's and R's, when trying to create the **AST** structure. To avoid this, the `::<` operator can also be extended to work with `IFunctor7`. The definition follows closely to the original definition, with some minor modifications to the types and constraints.

```
class (IFunctor7 iF, IFunctor7 iG) ⇒ iF ::< iG where
    inj :: iF f' a b c d e f g → iG f' a b c d e f g
instance IFunctor7 iF ⇒ iF ::< iF where
    inj = id
instance (IFunctor7 iF, IFunctor7 iG) ⇒ iF ::< (iF ::+ iG) where
    inj = L
instance (IFunctor7 iF, IFunctor7 iG, IFunctor7 iH, iF ::< iG) ⇒ iF ::< (iH ::+ iG) where
    inj = R · inj
```

Defining a constructor Data types for each constructor can be defined individually. The `Then` constructor is used as an example, however, the process can be applied to all constructors in the language.

```
data Then (iF :: [Type → Type] → [Type] → [Type]
    → [Type → Type] → [Type] → [Type] → Nat → Type)
    (inputsS :: [Type → Type]) (inputsT :: [Type]) (inputsA :: [Type])
    (outputsS :: [Type → Type]) (outputsT :: [Type]) (outputsA :: [Type])
    (nininputs :: Nat) where
    Then :: (DataStore' fs as, DataStore' gs bs, DataStore' hs cs)
        ⇒ iF fs as (Apply fs as) gs bs (Apply gs bs) nfs
        → iF gs bs (Apply gs bs) hs cs (Apply hs cs) ngs
        → Then iF fs as (Apply fs as) hs cs (Apply hs cs) nfs
```

Each `iF` denotes the recursive points in the data type, with the subsequent type arguments mirroring those seen in Section 3.4.1. A corresponding `IFunctor7` instance formalises the points of recursion, by showing how to transform the structure inside it.

```
instance IFunctor7 Then where
    imap7 f (Then x y) = Then (f x) (f y)
```

The smart constructor, that injects the L's and R's automatically can be defined for `Then` as:

```
(<=>) :: (Then :-<: iF, DataStore' fs as, DataStore' gs bs, DataStore' hs cs)
  => IFix7 iF fs as (Apply fs as) gs bs (Apply gs bs) nfs
  → IFix7 iF gs bs (Apply gs bs) hs cs (Apply hs cs) nfs
  → IFix7 iF fs as (Apply fs as) hs cs (Apply hs cs) nfs
(<=>) l r = lln7 (inj (Then l r))
infixr 4 <=>
```

The constructor adds one extra constraint, to the constructor defined in Section 3.4.1 — `Then :-<: iF`. This allows the smart constructor to produce an node in the **AST** for any sum of data types, that includes the `Then` data type.

Representing a Circuit Once each constructor has been defined then they can be combined together to form the `CircuitF` type, which can be used to represent a circuit.

```
type CircuitF = Id :+: Replicate :+: Then :+: Beside :+: Swap :+: DropL :+: DropR :+: Task :+: Map
```

The fixed-point of the `CircuitF` datatype can be defined with `IFix7`:

```
type Circuit = IFix7 CircuitF
```

4.3 Process Network

Now that it is possible to build a `Circuit`, which can be considered a specification for how to execute a set of tasks, there needs to be a mechanism in place to execute the specification. The standard implementation of a process network will use a Kahn Process Network (**KPN**). This means that each task in a circuit will run on its own separate thread, with inputs being passed between them on unbounded channels.

4.3.1 Network Typeclass

To allow for different process networks, a typeclass will be used to specify all the functions that every network should have. The `Network` typeclass is defined as:

```
class Network n where
  startNetwork :: Circuit inputsS inputsT inputsA outputsS outputsT outputsA nInputs
    → IO (n inputsS inputsT inputsA outputsS outputsT outputsA)
  stopNetwork :: n inputsS inputsT inputsA outputsS outputsT outputsA
    → IO ()
  write :: HList' inputsS inputsT
    → n inputsS inputsT inputsA outputsS outputsT outputsA
    → IO ()
  read :: n inputsS inputsT inputsA outputsS outputsT outputsA
    → IO (HList' outputsS outputsT)
```

This type class requires that a network has 4 different functions:

- `startNetwork` is responsible for converting the circuit into the underlying representation for a process network: it will be discussed in more detail in Section 4.4.
- `stopNetwork` is for cleaning up the network after it is no longer needed. For example, this could be stopping the threads running. This could be particularly important if embedding a circuit into a larger program, where unused threads could be left hanging.
- `write` should take some input values and add them into the network, so that they can be processed.
- `read` should retrieve some output values from the network.

Examples of how to use a network are included in Chapter 5.

4.3.2 The Basic Network Representation

An implementation of the `Network` typeclass is a `BasicNetwork`. This implementation makes use of a special case of heterogeneous list. A `PipeList` is used to represent a heterogeneous list of channels. This allows the `BasicNetwork` to store multiple channels in the same list without the need for existential types — one of the problems previously encountered with chains in Section 3.3.

```
data PipeList (fs :: [Type → Type]) (as :: [Type]) (xs :: [Type]) where
  PipeCons :: Chan (f a) → PipeList fs as xs → PipeList (f ' : fs) (a ' : as) (f a ' : xs)
  PipeNil  :: PipeList '[] '[] '[]
```

Making use of `PipeLists`, the `BasicNetwork` data type can be defined. This definition makes use of record syntax, this allows for named fields, with accessors automatically generated.

```
data BasicNetwork (inputsS :: [Type → Type]) (inputsT :: [Type]) (inputsA :: [Type])
  (outputsS :: [Type → Type]) (outputsT :: [Type]) (outputsA :: [Type]) where
  BasicNetwork :: {
    threads :: [ThreadId],
    inputs  :: PipeList inputsS inputsT inputsA,
    outputs :: PipeList outputsS outputsT outputsA }
    → BasicNetwork inputsS inputsT inputsA outputsS outputsT outputsA
```

The `BasicNetwork` has three fields:

- `threads` is a list of `ThreadId`s, this allows for the threads to be managed after their creation.
- `inputs` is a `PipeList` containing the channels that the initial input into the network.
- `outputs` is a `PipeList` that stores channels, which the output of the network can be read from.

The `Network` type instance for a `BasicNetwork` is relatively trivial to implement: if given a function to transform a `Circuit` to it.

```
instance Network BasicNetwork where
  startNetwork = buildBasicNetwork -- Definition to come...
  stopNetwork n = forM_ (threads n) killThread
  write uuid xs n = writePipes xs (inputs n)
  read n          = readPipes (outputs n)
```

The `writePipes` function will input a list of values into each of the respective pipes. The `readPipes` function will make a blocking call to each channel to read an output from it. This function will block till an output is read from every output channel.

4.4 Translation to a Network

There is now a representation for a `Circuit` that the user will build, and a representation used to execute the `Circuit`. However, there is no mechanism to convert between them. This can be achieved by folding the circuit data type into a network. This fold, however, will need to create threads and channels, both of which IO actions. The current definition for the fold `icata7` is not able perform monadic computation inside the algebra. To solve this `unsafePerformIO` could be used, however, for this to be safe the IO computation needs to have no side-effects. This fold will violate this rule, therefore, the only other way to support this is to modify the catamorphism to support monadic computation.

4.4.1 Indexed Monadic Catamorphism

An indexed monadic catamorphism, found in Section 2.3.1, can be used to perform this fold: it will allow for monadic computation within the algebra. However, `icataM` needs to be modified to support the 7 type indices needed. The first step is to define a monadic `imap` that supports the needed number of type indices. This will be added by extending the `IFunctor7` instance to also include a function named `imapM7`.

```

class IFunctor7 iF where
  imap7 :: (∀ a b c d e f g. f' a b c d e f g → g' a b c d e f g) → iF f' a b c d e f g → iF g' a b c d e f g
  imapM7 :: Monad m ⇒ (∀ a b c d e f g. f' a b c d e f g → m (g' a b c d e f g))
    → iF f' a b c d e f g
    → m (iF g' a b c d e f g)

```

The definition for this new function `imapM7` for each instance closely follows the non-monadic version, however, now has a monadic function to map on the input. Here is the definition of the new `IFunctor7` instance for `Beside`:

```

instance IFunctor7 Beside where
  imap7 f (Beside l r) = Beside (f l) (f r)
  imapM7 f (Beside l r) = do
    l' ← f l
    r' ← f r
    return (Beside l' r')

```

The definition is intuitively the same, just using do-notation instead.

Now that there is a indexed monadic map, it is possible to define the a monadic catamorphism for an `IFunctor7`:

```

icataM7 :: (IFunctor7 iF, Monad m)
  ⇒ (∀ a b c d e f g. iF f' a b c d e f g → m (f' a b c d e f g))
  → IFix7 iF a b c d e f g
  → m (f' a b c d e f g)
icataM7 algM (lIn x) = algM ≡≡ imapM7 (icataM7 algM) x

```

`icataM7` is almost identical to `icataM`, however, it makes use of `imapM7` instead of `imapM`.

4.4.2 BuildNetworkAlg

To use `icataM7` to fold a `Circuit` into a `BasicNetwork`, an algebra is required. However, a standard algebra will not be able to complete this transformation. Consider this example `Circuit` with two tasks executed in sequence.

```

example = task1
        <=>
        task2

```

In a standard algebra both sides of the `Then` constructor would be evaluated independently. In this case it would produce two disjoint networks, both with their own input and output channels. The algebra for `Then`, would then need to join the output channels of `task1` with the input channels of `task2`. However, it is not possible to join channels together. Instead, the output channels from `task1` need to be accessible when creating `task2`. This is referred to as a *context-sensitive* or *accumulating* fold. An accumulating fold forms series of nested functions, that collapse to give a final value once the base case has been applied. A simple example of an accumulating fold could be, implementing `foldl` in terms of `foldr`.

```

foldl :: (b → a → b) → b → [a] → b
foldl f b as = foldr (λ a g x → g (f x a)) id as b

```

A simple example of `foldl` can be considered.

```

foldl (+) 0 [1, 2]
≡
(λ x → (λ x → id (x + 2)) (x + 1)) 0

```

To be able to have an accumulating fold inside an indexed catamorphism a carrier data type is required to wrap up this function. This carrier, which shall be named `AccuN`, contains a function that when given a network that has been accumulated up to that point, then it is able to produce a network including the next layer in a circuit. This can be likened to the lambda function given to `foldr`, when defining `foldl`. The type of the layer being folded will be `Circuit a b c d e f g`.

```

newtype AccuN n asS asT asA a b c d e f g = AccuN
  { unAccuN :: n asS asT asA a b c → IO (n asS asT asA d e f) }

```

This newtype has 3 additional type parameters at the beginning, namely: `asS`, `asT`, `asA`. They represent the input types to the initial circuit. Since the accumulating fold will work layer by layer from the top downwards, these types will remain constant and never change throughout the fold.

Classy Algebra To ensure that the approach remains modular, the algebra takes the form of a type class: the interpretation of a new constructor is just a new type class instance.

```

class (Network n, IFunctor7 iF) ⇒ BuildNetworkAlg n iF where
  buildNetworkAlg :: iF ( AccuN n asS asT asA ) bsS bsT bsA csS csT csA nbs
    → IO ((AccuN n asS asT asA) bsS bsT bsA csS csT csA nbs)

```

This algebra type class takes two parameters: `n` and `iF`. The `n` is constrained to have a `Network` instance, this allows the same algebra to be used for defining folds for multiple network types. The `iF` is the `IFunctor7` that this instance is being defined for, an example is `Then` or `Id`. This algebra uses the `AccuN` data type to perform an accumulating fold. The input to the algebra is an `IFunctor7` with the inner elements containing values of type `AccuN`. The function can be retrieved from inside `AccuN` to perform steps that are dependent on the previous, for example, in the `Then` constructor.

For an algebra to handle sums of `IFunctor7s`, `:+:` has an algebra instance to direct it through the nest of `Ls` and `Rs`. It knows what to do in each case, due to the constraint that each side of the sum must also have an instance of the algebra.

```

instance (BuildNetworkAlg n iF, BuildNetworkAlg n iG) ⇒ BuildNetworkAlg n (iF :+: iG) where
  buildNetworkAlg (L x) = buildNetworkAlg x
  buildNetworkAlg (R y) = buildNetworkAlg y

```

The Initial Network Before being able to define the actual translation, there is one more base to cover. This is an accumulating fold that depends on the previous layer to be able to define the current one. However, what happens on the first layer? There is no previous `Network` to use. The fold needs an initial network that has matching input and output types, this means that the input channels should be the same as the output channels.

The following new type class will generate a `PipeList` of channels that will be stored in the initial network:

```

class InitialPipes (inputsS :: [Type → Type]) (inputsT :: [Type]) (inputsA :: [Type]) where
  initialPipes :: IO (PipeList inputsS inputsT inputsA)

```

This type class constructs an `initialPipes` based on the type required in the initial network. The following two instances will allow for the construction of the `initialPipes` value:

```

instance InitialPipes '[]' '[]' '[]' where
  initialPipes = return PipeNil
instance InitialPipes fs as xs ⇒ InitialPipes (f ': fs) (a ': as) (f a ': xs) where
  initialPipes = do
    c ← newChan :: IO (Chan (f a))
    PipeCons c <$> (initialPipes :: IO (PipeList fs as xs))

```

The first instance deals with the base case: when the type lists are empty, an empty `PipeList` is created. The latter more interesting case deals with a cons in the type lists. Here a new channel is created with the same type as the type removed from the front of the type list. The channel is then consed to the front of a `PipeList` with the remaining list generated by a recursive call.

Now that there is a method for creating a `PipeList` that matches a type-list, the initial network is defined:

```
initialNetwork
  :: ∀inputsS inputsT inputsA
  . (InitialPipes inputsS inputsT inputsA)
  ⇒ IO (BasicNetwork inputsS inputsT inputsA inputsS inputsT inputsA)
initialNetwork = do
  ps ← initialPipes :: IO (PipeList inputsS inputsT inputsA)
  return $ BasicNetwork [] ps ps
```

This creates a network that has matching input and output types. To do so `initialNetwork` creates a `PipeList` of the initial channels, which is then used as both the inputs and outputs of the network.

4.4.3 The Translation

Now that the algebra type class, and the initial input to the accumulating fold is defined, each instance of the type class are defined.

Basic Constructors There are several constructors that just manipulate the output `PipeList`, these constructors are `Id`, `Replicate`, `Swap`, `DropL`, and `DropR`. The `Swap` constructor takes two inputs and then swaps them over:

```
instance BuildNetworkAlg BasicNetwork Swap where
  buildNetworkAlg Swap = return $ AccuN $
    (λn → do
      output ← swapOutput (outputs n)
      return $ BasicNetwork (threads n) (inputs n) output
    )
  where
    swapOutput :: PipeList '[f, g] '[a, b] '[f a, g b]
      → IO (PipeList '[g, f] '[b, a] '[g b, f a])
    swapOutput (PipeCons c1 (PipeCons c2 PipeNil)) =
      return $ PipeCons c2 (PipeCons c1 PipeNil)
```

The instance for `Swap`, defines a function wrapped by `AccuN`, that takes the current accumulated network, up to this point. It then is able to transform the outputs and build a new `BasicNetwork`. To transform the networks it makes use of a function named `swapOutput`, which unpacks the `PipeList` and swaps the two channels `c1` and `c2` over.

Another basic constructor is `Replicate`: the purpose of this constructor is to duplicate the input to produce two outputs. The instance for `Replicate` is defined as:

```
instance BuildNetworkAlg BasicNetwork Replicate where
  buildNetworkAlg Replicate = return $ AccuN
    (λn → do
      output ← dupOutput (outputs n)
      return $ BasicNetwork (threads n) (inputs n) output
    )
  where
    dupOutput :: PipeList '[f] '[a] '[f a]
      → IO (PipeList '[f, f] '[a, a] '[f a, f a])
    dupOutput (PipeCons c PipeNil) = do
      c' ← dupChan c
      return $ PipeCons c (PipeCons c' PipeNil)
```

This instance follows a similar pattern to the `Swap` instance — defining a function which retrieves the accumulated network, then manipulates the outputs. However, the `dupOutput` function also has to make use of an operation on the channel — `dupChan`. This will create a new channel that will mirror the inputs of the original channel.

All other basic constructors will follow this pattern:

- `Id`, will return the same outputs as the accumulated network.
- `DropL`, will drop the *first* item in the output `PipeList` of the accumulated network.
- `DropR`, will drop the *last* item in the output `PipeList` of the accumulated network.

Task In a `BasicNetwork` a task will run as a separate thread, to do this `forkIO :: IO () → IO ThreadId` will be used. Using this function requires some `IO ()` computation to run, this will be defined by `taskExecutor`:

```
taskExecutor
  :: Task iF inputsS inputsT inputsA outputS outputT outputsA ninputs
  → PipeList inputsS inputsT inputsA
  → PipeList outputS outputT outputA
  → IO ()
taskExecutor (Task f outStore) inPipes outPipes = forever
  (do
    taskInput ← readPipes inPipes
    r ← f taskInputs outStore
    writePipes (HCons' r HNil') outPipes
  )
```

The `taskExecutor` has three arguments:

- The `Task` to be executed on the thread.
- A `PipeList` which has channels containing the input values.
- A `PipeList` to output the results of the `Task`.

A `taskExecutor` will, read a value from each of input channels, execute the task with those inputs, and then write the output to the output channels. This computation is then repeated forever, using the aptly named function `forever`.

Making use of the `taskExecutor`, the algebra instance for `Task` is defined as:

```
instance BuildNetworkAlg BasicNetwork Task where
  buildNetworkAlg (Task t out) = return $ AccuN
    (λn → do
      c ← newChan
      let output = PipeCons c PipeNil
      threadId ← forkIO (taskExecutor (Task t out) (outputs n) output)
      return $ BasicNetwork (threadId : threads n) (inputs n) output
    )
```

This instance first creates a new output channel, this will be given to the task to send its outputs on. It then forks a new thread with the computation generated by `taskExecutor`. The executor is given the output values of the accumulated network and the output channel, just created. The resulting network has the same inputs, but now adds a new thread id to the list and the outputs set to be the output channels from the task.

Then The `Then` constructor is responsible for connecting circuits in sequence. When converting this to a network, this will involve making use of the accumulated network value to generate the next layer. The instance is defined as:

```
instance BuildNetworkAlg BasicNetwork Then where
  buildNetworkAlg (Then (AccuN fx) (AccuN fy)) = return $ AccuN (fx >=> fy)
```

This instance has an interesting definition: firstly it takes the accumulated network `n` as input. It then uses the function `fx`, with the input `n` to generate a network for the top half of the `Then` constructor. Finally, it takes the returned network `nx`, from the top half of the constructor, and generates a network using the function `fy` representing the bottom half of the constructor.

Beside The **Beside** constructor places two circuits side by side. This is the most difficult algebra to define as the accumulated network needs to be split in half to pass to the two recursive sides of **Beside**. An instance of the algebra is defined as:

```
instance BuildNetworkAlg BasicNetwork Beside where
  buildNetworkAlg = beside
```

This requires a **beside** function, however to define this function some extra tools are required. The first is **takeP**, which will take the first n elements from a **PipeList**:

```
takeP :: SNat n → PipeList fs as xs → PipeList (Take n fs) (Take n as) (Take n xs)
takeP SZero _ = PipeNil
takeP (SSucc _) PipeNil = PipeNil
takeP (SSucc n) (PipeCons x xs) = PipeCons x (takeP n xs)
```

This makes use of the **Take** type family to take n elements from each of the type lists: **fs**, **as**, and **xs**. It follows the same structure as the **take** :: **Int** → **[a]** → **[a]** defined in the **Prelude**.

The next function is **dropP**, it drops n elements from a **PipeList**:

```
dropP :: SNat n → PipeList fs as xs → PipeList (Drop n fs) (Drop n as) (Drop n xs)
dropP SZero _ = _
dropP (SSucc _) PipeNil = PipeNil
dropP (SSucc n) (PipeCons _ xs) = dropP n xs
```

This function again follows the same structure as **drop** :: **Int** → **[a]** → **[a]** defined in the **Prelude**. Both **takeP** and **dropP** are used to split the outputs of a network after n elements. This requires the knowledge of what n is at the value level, however n is only stored at the type level as the argument **ninputs**. To be able to recover this value the **IsNat** type class, as defined in Section 2.5.2 is used. The **recoverNInputs** function is able to direct the **IsNat** type class to the correct type argument, and produces an **SNat** with the same value as that stored in the type.

```
recoverNInputs :: (Length bsS ~ Length bsT, Length bsT ~ Length bsA, Length bsA ~ Length bsS,
  ninputs ~ Length bsS, IsNat ninputs, Network n)
  ⇒ (N n asS asT asA) bsS bsT bsA csS csT csA (ninputs :: Nat)
  → SNat (Length bsS)
circuitInputs _ = nat
```

After splitting a network and generating two new networks, the outputs will need to be joined together again: this will require the appending of two **PipeLists**. To do this an **AppendP** type class is defined:

```
class AppendP fs as xs gs bs ys where
  appendP :: PipeList fs as xs → PipeList gs bs ys → PipeList (fs :++ gs) (as :++ bs) (xs :++ ys)
```

This type class has one function **appendP**, it is able to append two **PipeLists** together. It makes use of the **:++** type family to append the type lists together. The instances for this type class are made up of two cases: the base case and a recursive case.

```
instance AppendP '[] '[] '[] gs bs ys where
  appendP PipeNil _ = ys
instance (AppendP fs as xs gs bs ys) ⇒ AppendP (f ' : fs) (a ' : as) (f a ' : xs) gs bs ys where
  appendP (PipeCons x xs) ys = PipeCons x (appendP xs ys)
```

The base case corresponds to having an empty list on the left, with some other list on the right. Here the list on the right is returned. The recursive case, simply takes 1 element from the left hand side and conses it onto the front of a recursive call, with the rest of the left hand side.

It is now possible to define the `beside` function. The result is calculated in 4 steps, with helper functions for each step:

1. Get the number of inputs (`ninputs`) on the left hand side of the `Beside` constructor. This will give the information needed to split the inputted accumulated network `n`.
2. Split the network into a left and right hand side. This will retain the same input type to the network, as there is no information on how to split that. Only the output `PipeList` will be split into two parts.
3. Translate the both the left and right network. This will perform the recursive step and generate two new networks with the networks from the left and right added to the accumulated network `n`.
4. Join the networks back together. Now that the left and right hand side of this layer has been added to the accumulated network, the two sides need to be joined back together to get a single network that can be returned.

```

beside :: ∀asS asT asA bsS bsT bsA csS csT csA (nbs :: Nat)
  .Beside (AccuN BasicNetwork asS asT asA) bsS bsT bsA csS csT csA nbs
  → IO ((AccuN BasicNetwork asS asT asA) bsS bsT bsA csS csT csA nbs)
beside (Beside l r) = return $ AccuN
  (λn → do
    let ninputs = circuitInputs l
    (nL, nR) ← splitNetwork ninputs n
    (newL, newR) ← translate ninputs (nL, nR) (l, r)
    joinNetwork (newL, newR)
  )
where
  splitNetwork :: SNat nbsL
    → BasicNetwork asS asT asA bsS bsT bsA
    → IO (BasicNetwork asS asT asA (Take nbsL bsS) (Take nbsL bsT) (Take nbsL bsA),
          BasicNetwork asS asT asA (Drop nbsL bsS) (Drop nbsL bsT) (Drop nbsL bsA))
  splitNetwork nbs n = return
    (BasicNetwork (threads n) (inputs n) (takeP nbs (outputs n)),
     BasicNetwork (threads n) (inputs n) (dropP nbs (outputs n)))
  translate :: SNat nbsL
    → (BasicNetwork asS asT asA (Take nbsL bsS) (Take nbsL bsT) (Take nbsL bsA),
        BasicNetwork asS asT asA (Drop nbsL bsS) (Drop nbsL bsT) (Drop nbsL bsA))
    → ((AccuN BasicNetwork asS asT asA)
        (Take nbsL bsS) (Take nbsL bsT) (Take nbsL bsA) csLS csLT csLA nbsL,
        (AccuN BasicNetwork asS asT asA)
        (Drop nbsL bsS) (Drop nbsL bsT) (Drop nbsL bsA) csRS csRT csRA nbsR)
    → IO (BasicNetwork asS asT asA csLS csLT csLA,
          BasicNetwork asS asT asA csRS csRT csRA)
  translate _ (nL, nR) (N cL, N cR) = do
    nL' ← cL nL
    nR' ← cR nR
    return (nL', nR')
  joinNetwork :: (AppendP csLS csLT csLA csRS csRT csRA)
    ⇒ (BasicNetwork asS asT asA csLS csLT csLA, BasicNetwork asS asT asA csRS csRT csRA)
    → IO (BasicNetwork asS asT asA (csLS :++ csRS) (csLT :++ csRT) (csLA :++ csRA))
  joinNetwork (nL, nR) = return
    $ BasicNetwork (nub (threads nL ++ threads nR)) (inputs nL) (outputs nL `appendP` outputs nR)

```

The `splitNetwork` function creates two new `BasicNetwork`s. To split the output values, `takeP`, and `dropP` are used. `translate` performs the recursive step in the accumulating fold, which produces two new networks that include this layer. `joinNetwork` takes the two new networks and appends the outputs with `appendP`. It also has to append the thread ids from both sides, however, this will now include duplicates as threads were not split in `splitNetwork`. To combat this `nub` is used, which returns a list containing all the unique values in the original.

Using the algebra in conjunction with `icataM7`, `buildBasicNetwork` is now defined:

```
buildBasicNetwork :: InitialPipes a b c ⇒ Circuit a b c d e f g → IO (BasicNetwork a b c d e f)
buildBasicNetwork x = do
  n ← icataM7 buildNetworkAlg x
  n' ← initialNetwork
  unAccuN n n'
```

`icataM7` builds a value of type `AccuN`, which is unpacked and applied with the `initialNetwork`. The application step, causes the nest of accumulator functions to collapse and generate a new network, creating the channels and threads.

4.5 UUIDS

When inputting multiple values into a `Network` problems can occur. For example, a task's output pointer is statically defined. If this were a file, it would result in files being overwritten before they have been read. This is eliminated through the use of UUIDs, they act as a unique identifier for each input into the network.

4.5.1 Modifications

To be able to support a UUID, several small modifications need to be made.

Data Store The value is accessible to the data store when saving by making a small modification:

```
class DataStore f a where
  fetch :: UUID → f a → IO a
  save :: UUID → f a → a → IO (f a)
```

This can then be made use of when reading or writing to a data store. For example, a filename could be prepended with the unique identifier, or it could be used as a primary key when saving to a database table.

PipeList To transfer the value around the network, the `PipeList` data type is modified to store channels of type `(UUID, f a)`, instead of just `f a`:

```
data PipeList (fs :: [Type → Type]) (as :: [Type]) (xs :: [Type]) where
  PipeCons :: Chan (UUID, f a) → PipeList fs as xs → PipeList (f ': fs) (a ': as) (f a ': xs)
  PipeNil :: PipeList '[] '[] (Apply '[] '[])
```

Task Executor The task executor needs to be modified, so that it retrieves the UUID from the input channels, gives it to the task, and passes it on down the output channels.

```
taskExecutor :: Task iF inputsS inputsT inputsA outputS outputT outputsA ninputs
  → PipeList inputsS inputsT inputsA
  → PipeList outputS outputT outputA
  → IO ()
taskExecutor (Task f outStore) inPipes outPipes = forever
  (do
    (uuid, taskInput) ← readPipes inPipes
    r ← f uuid taskInputs outStore
    writePipes uuid (HCons' r HNil') outPipes
  )
```

Network: Read & Write The read and write methods defined in the `Network` type class are modified to also take a `UUID`:

```
class Network n where
  ...
  read :: n inputsS inputsT inputsA outputsS outputsT outputsA → IO (UUID, HList' outputsS outputsT)
  write :: UUID → HList' inputsS inputsT → n inputsS inputsT inputsA outputsS outputsT outputsA → IO ()
```

4.5.2 Helper Functions

There are several helper functions for reading and writing into a network:

```
input :: Network n ⇒ HList' inputsS inputsT
  → n inputsS inputsT inputsA outputsS outputsT outputsA
  → IO UUID
input_ :: Network n ⇒ HList' inputsS inputsT
  → n inputsS inputsT inputsA outputsS outputsT outputsA
  → IO ()
output_ :: Network n ⇒ n inputsS inputsT inputsA outputsS outputsT outputsA
  → IO (HList' outputsS outputsT)
```

Both of the input functions generate a random `UUID`, meaning the user does not have to specify one. `input` will return this generated values, whereas `input_` will not. `output_` fetches values from a network, but does not return the `UUID` with the outputs.

4.6 Failure in the Process Network

Currently, when an exception occurs in a task the whole network crashes — this isn't desired behaviour. There are many ways to model failure in Haskell, one such example could be the `Maybe` monad.

4.6.1 Maybe not Maybe

`Maybe` can capture failure, with `Just x` being the success case, and `Nothing` being the failed case. When `Nothing` is produced the rest of the computation automatically fails due to the definition of `>>=`.

```
instance Monad Maybe where
  return x = Just x
  (>>=) Nothing _ = Nothing
  (>>=) (Just x) f = f x
```

This would work well in a network as an error in one task can be propagated to all it dependents, causing them to fail gracefully, and the error propagating further through the network. However, there is one problem with `Maybe`, it does not retain any information about the error that occurred. This information would be very useful to a user, as it helps them debug the issue.

4.6.2 Except Monad

The `Except` monad is based on `Either`, with the `Left` constructor representing failure, and the `Right` constructor indicating success. This means that an error message can now be stored, when a failure may occur. Since tasks already execute in the `IO` monad, a monad transformer `ExceptT` is required, so that failure can be implemented into the network. This allows for computation in both `IO`, and `Except`, however, any `IO` computation will need to be lifted into the `ExceptT` monad.

The following modifications are required to add modelling of failure in a network.

PipeList A `PipeList` will now need to also transfer information about whether the previous task failed to execute. To do this it will carry an `Either TaskError (f a)`, with `TaskError` being a custom data type storing the error message text.

```
data PipeList (fs :: [Type → Type]) (as :: [Type]) (xs :: [Type]) where
  PipeCons :: Chan (UUID, Either TaskError (f a))
    → PipeList fs as xs
    → PipeList (f ' : fs) (a ' : as) (f a ' : xs)
  PipeNil :: PipeList '[] '[] (Apply '[] '[])
```

Task Executor The task executor will need to be modified so that it executes the tasks in the `ExceptT` monad.

```
taskExecutor :: Task iF inputsS inputsT inputsA outputS outputT outputsA ninputs
  → PipeList inputsS inputsT inputsA
  → PipeList outputS outputT outputA
  → IO ()
taskExecutor (Task f outStore) inPipes outPipes = forever
  (do
    (uuid, taskInputs) ← readPipes inPipes
    r ← runExceptT $
      (do
        input ← (ExceptT · return) taskInputs
        r ← catchE (intercept (f uuid input outStore))
          (throwE · TaskError · ExceptionMessage · displayException)
        return (HCons' (r `deepseq` r) HNil')
      )
    writePipes uuid r outPipes
  )
```

This version of the executor, first reads the values from the input channels. It then runs some computation in the `ExceptT` monad to get a return value `r :: Either TaskError (HList' outputsS outputsT)`. The return value is then sent along the output channels.

The `catchE` function can be used to catch an exception thrown in a task, the exception is then converted to a `TaskError`, and re-thrown. This will mean that only a `Either TaskError (HList' outputsS outputsT)` is returned, rather than another type of error.

There is, however, an error that is caused by Haskell's laziness: if a value is not evaluated inside the `runExceptT` block then it will not be caught, and the program will continue to crash. To solve this the `deepseq` function is used, which fully evaluates the left argument, before returning the right argument. This forces any error that could occur to happen inside the `runExceptT` block.

4.7 Evaluation

4.7.1 Requirements

☑ **Type-safe** A `Network` uses strong types to verify the correctness of the implementation and the translation to it. The implementation does not make use of any unsafe casts or coerces. The strength of types during the translation caught numerous bugs: the implementation and translation worked first time after it fully compiled.

☑ **Parallel** The network achieves parallel computation through its use of threads and channels. This allows it to perform multiple tasks simultaneously, increasing the run-time. The `BasicNetwork`, however, does have a limitation that it has to create a thread for every task: this may not be idea for large networks.

☑ **Competitive Speed** The network is, indeed, able to compute results in a time that is competitive with other libraries. More detail on this can be found in Chapter 6.

☑ **Failure Tolerance** A network is able to catch an error occurring in a task, and then propagate it through the network. Due to this an exception in a task cannot crash the program, instead the error is return as the output of the network. This means that if multiple inputs are inputted into the network, and one of them crashes, all results will be returned apart from the one that crashed.

☑ **Usable** The implementation of constructors does not add any extra constraints that would impact on the original language design. Smart constructors have been used to hide the **AST**, that is built under the hood to represent a **Circuit**.

☑ **Maintainable** The implementation has been designed with maintainability in mind. It is possible to add now constructors, without impacting on those currently defined. The translation uses individual algebra instances for each constructor, which means they are only required to be defined where it makes sense.

The network system has also been defined in a way that is easily extendable in the future. New instances could be defined that allow for other types of network, such as a profiling network, or a distributed network.

4.7.2 Development Practices

Unit Tests The implementation of the language has been verified by a set of unit tests. A test for each constructor has been defined that ensures it has the expected behaviour. For example, the test for replicate is:

```
replicateCircuit :: Circuit
  '[VariableStore]
  '[Int]
  '[VariableStore Int]
  '[VariableStore, VariableStore]
  '[Int, Int]
  '[VariableStore Int, VariableStore Int]
  N1
replicateCircuit = replicate
replicateTests :: TestTree
replicateTests = testGroup
  "replicate should"
  [testCase "return a duplicated input value" $ do
    let i = HCons' (Var 0) HNil'
    n ← startNetwork circuit :: IO (BasicNetwork a b c d e f)
    input_ i n
    o ← output_ n
    stopNetwork n
    o @? = Right (HCons' (Var 0) (HCons' (Var 0) HNil'))
  ]
```

This test defines a circuit that uses just the replicate operator. It then starts and network and inputs one value into it. The result is then read from the network and compared to the expected output.

The tests for Task, also ensure that if an exception occurs it is caught correctly and the error message is returned.

Source Code Management & Continuous Integration The source code for this implementation is stored in a GitHub repository: <https://github.com/RileyEv/CircuitFlow>. The repository for this project makes use of continuous integration (GitHub Actions), to ensure that no breaking changes are introduced into the code. Each new feature has been developed and merged using pull requests that enforce several checks before any new feature can be merged:

- The code must build
- All tests must pass

- The HLint tool must report that there are no linting issues.

This ensures that the library is well maintained and that no changes accidentally introduce breaking faults.

Documentation Unlike most Haskell libraries, this implementation is fully documented with 100% coverage. The generation of the documentation is integrated into the continuous integration workflow, with it being deployed to a publicly hosted system: <https://circuitflow.rly.rocks>.

Chapter 5

Examples

5.1 Machine Learning (Audio Playlist Generation)

A use case for CircuitFlow is when building data pipelines. Here there are many tasks that can only be executed when other data files have been produced. A data pipeline will also benefit from being parallel to improve run-times. CircuitFlow can help to build a parallel data pipeline, without the user having to worry about how to combine all their data manipulations together, in a way that will not run into concurrency problems.

Consider the example where an audio streaming service would like to create a playlist full of new songs to listen to. This could require a machine learning model that can predict your songs based on the top 10 artists and songs that you have listened to over the last 3 months. However, each of the months data is stored in different files that need aggregating together, before they can be input into the model.

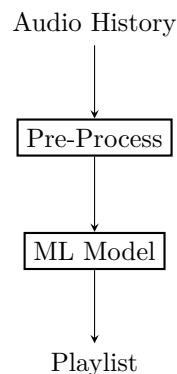


Figure 5.1: A dataflow diagram for playlist generation

5.1.1 Building the pre-processing Circuit

This circuit will need to have 3 different inputs — each months listening history. It will also have 2 outputs: the top 10 songs and artists. To begin with a dataflow diagram can be constructed — seen in Figure 5.2. This will model all the dependencies between each of the pre-processing tasks.

To write this in CircuitFlow, the first step is to create all the tasks that will be used. Both the `AggSongs` and `AggArtists` tasks will require three inputs, which will be CSVs. This means that it is possible to make use of the pre-defined data store: `NamedCSVStore`. The type instances that are required to parse the CSV will be omitted, as they are not relevant to the discussion.

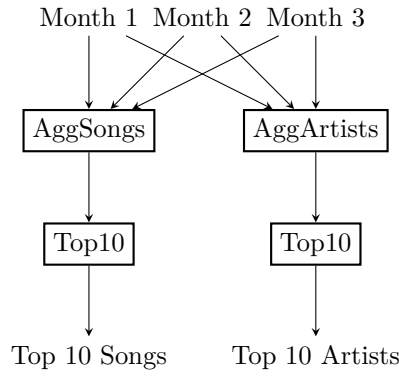


Figure 5.2: A dataflow diagram for pre-processing the song data

```

aggSongsTask :: Circuit
  '[NamedCSVStore,      NamedCSVStore,      NamedCSVStore]
  '[      [Listen],      [Listen],      [Listen]]
  '[NamedCSVStore [Listen], NamedCSVStore [Listen], NamedCSVStore [Listen]]
  '[VariableStore]
  '[      [TrackCount]]
  '[VariableStore [TrackCount]]
  N3
aggSongsTask = multiInputTask f Empty
where
  f :: HList '[[Listen], [Listen], [Listen]] → [TrackCount]
  f (HCons month1 (HCons month2 (HCons month3 HNil))) =
    (map (uncurry TrackCount) · reverse · count_ · map track) (month1 ++ month2 ++ month3)

```

This task applies a composition of functions:

1. Firstly, the track is extracted from the `Listen` data type using the function `track :: Listen → Track`.
2. Next, a list of unique tracks are extracted, along with the number of appearances they made in the original list. This makes use of a special variant `count_ :: [a] → [(a, Int)]`, which will also sort the list in ascending order.
3. This list is then flipped to descending order with `reverse :: [a] → [a]`.
4. Finally, a `TrackCount` value is made from the previous tuple `uncurry TrackCount :: (Track, Int) → TrackCount`.

```

aggArtistsTask :: Circuit
  '[NamedCSVStore,      NamedCSVStore,      NamedCSVStore]
  '[      [Listen],      [Listen],      [Listen]]
  '[NamedCSVStore [Listen], NamedCSVStore [Listen], NamedCSVStore [Listen]]
  '[VariableStore]
  '[      [ArtistCount]]
  '[VariableStore [ArtistCount]]
  N3
aggArtistsTask = multiInputTask f Empty
where
  f :: HList '[[Listen], [Listen], [Listen]] → [ArtistCount]
  f (HCons month1 (HCons month2 (HCons month3 HNil))) =
    (map (uncurry ArtistCount) · reverse · count_ · map (artist · track)) (day1 ++ day2 ++ day3)

```

The `aggArtistsTask` is constructed in a similar way to the `aggSongsTask`, however, it extracts the artist from the track before aggregating the data.

The final task to define is `Top10`, however, this task is used multiple times. Therefore, it would be beneficial if the type was polymorphic so that it can receive and input of both `[ArtistCount]` and `[TrackCount]`. This is possible to do:

```
top10Task :: (ToNamedRecord a, FromNamedRecord a, DefaultOrdered a)
  => FilePath
  -> Circuit '[VariableStore] '[a] '[VariableStore [a]]
            '[NamedCSVStore] '[a] '[NamedCSVStore [a]]
            N1
top10Task filename = functionTask (take 10) (NamedCSVStore filename)
```

The `top10Task` takes an input list and then returns the first 10 from the list.

Now that all of the tasks have been defined, they need to be combined into a circuit. The dataflow diagram seen in Figure 5.2, will prove to be helpful. The first obvious problem is how to create a circuit that can achieve the top part of the diagram, transforming the inputs so that they can be passed into two tasks. This can be dealt with in layers: using the diagrams associated with each constructor, it is possible to convert a layer into a composition of Circuits. The first layer, seen in Figure 5.3 will be responsible for duplicating each of the three inputs.

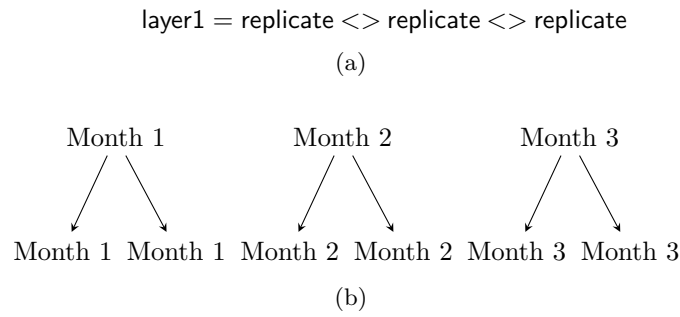


Figure 5.3: The first layer (a) and its corresponding dataflow diagram (b).

Scanning down the dataflow diagram the next layer to deal with is the two swaps that occur: month 1 and 2, and month 2 and 3. This can be seen in Figure 5.4

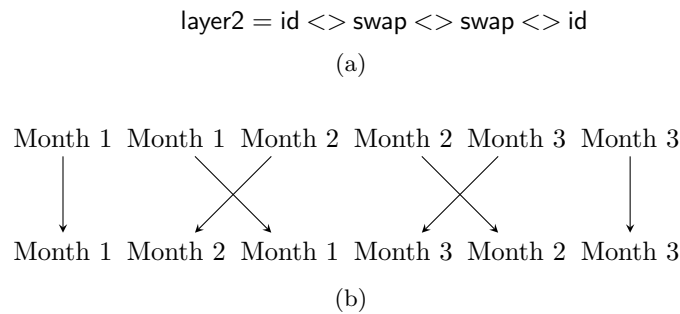


Figure 5.4: The second layer (a) and its corresponding dataflow diagram (b).

The final layer, seen in Figure 5.5, will swap month 1 with month 3.

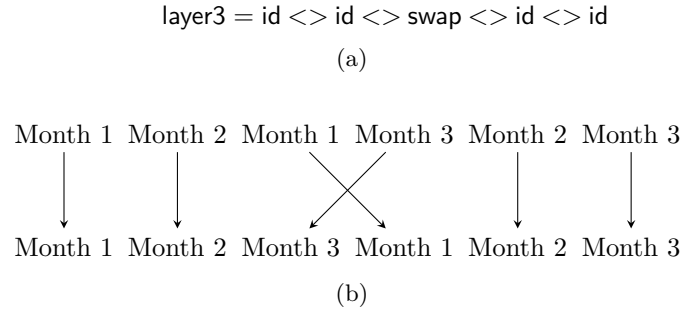


Figure 5.5: The third layer (a) and its corresponding dataflow diagram (b).

These layers can be stacked on top of each other to provide a full transformation:

```
organiseInputs = layer1
                  <>
                  layer2
                  <>
                  layer3
```

The tasks can now be combined with this transformation to provide the full circuit:

```
preProcPipeline :: Circuit
  '[NamedCSVStore,      NamedCSVStore,      NamedCSVStore]
  '[      [Listen],      [Listen],      [Listen]]
  '[NamedCSVStore [Listen], NamedCSVStore [Listen], NamedCSVStore [Listen]]
  '[NamedCSVStore, NamedCSVStore]
  '[      [ArtistCount],      [TrackCount]]
  '[NamedCSVStore [ArtistCount], NamedCSVStore [TrackCount]]
  N3
preProcPipeline = organiseInputs
                  <>
                  aggSongsTask      <> aggArtistsTask
                  <>
                  top10Task "top10Artists.csv" <> top10Task "top10Artists.csv"
```

Again it can be seen how this structure of tasks directly correlates with the dataflow diagram previously seen in Figure 5.2. This helps to make it easier when designing circuits as it can be constructed visually level by level.

5.1.2 Building the prediction Circuit

Now that there is a pipeline to pre-process the data for the model, a new playlist can be created. Again it is possible to build a circuit that combines with the pre-processing circuit.

The audio company train models for each user and store them in a cloud storage service. This would be a good use-case for creating a new `DataStore` — called a `ModelStore`. Say that there are two types defined `ModelStore` and `NewSongPlaylist`: a new `DataStore` instance can be defined for each different model. Here is the `NewSongsPlaylist` as an example:

```
instance DataStore ModelStore NewSongsPlaylist where
  -- Fetch from cloud storage and load into the program
  fetch (ModelStore NewSongsPlaylist) = ...
  save (ModelStore NewSongsPlaylist) _ = ...
```

`fetch` is used to download a trained model from the cloud store. `save` will be used when training a new model, it will allow the model to be saved for use in the future.

With this `DataStore` a new task can be defined that will, take a users top 10 songs and artists, and a model as input. It will then output a list of songs to create a playlist.

```

predictTask :: Circuit
  '[NamedCSVStore,           NamedCSVStore, ,           ModelStore]
  '[           [ArtistCount],           [TrackCount],           NewSongsPlayList]
  '[NamedCSVStore [ArtistCount], NamedCSVStore [TrackCount], ModelStore NewSongsPlayList]
  '[NamedCSVStore]
  '[           [Track]]
  '[NamedCSVStore [Track]]
  N3
predictTask = multiInputTask f (NamesCSVStore "newSongsPlaylist.csv")
  where
    f :: HList '[ [ArtistCount], [TrackCount], NewSongsPlayList ] → [Track]
    f (HCons topArtists (HCons topSongs (HCons model HNil))) =
      predict model topArtists topSongs

```

The `predictTask` can now be combined with the pre-processing circuit to create a circuit that is able to create a new playlist from listening history.

```

createPlaylist = preProcPipeline <> id
                <>
                predictTask

```

The `createPlaylist`, circuit can now be converted into a `Network` and used on user data to generate new playlists, based on their top songs. `preProcPipeline` will be revisited in Chapter 6, to act as a benchmark to compare the performance of the `CircuitFlow` library.

5.2 Build System (lhs2TeX)

Another use case for a task based dependency system is a build system. For example, a Makefile is a way of specifying the target files from some source files, with a command that can be used to generate the target file. `CircuitFlow`: could also be used to model such a system.

Consider this dissertation, which is made using `LATEX`. This project is made up of multiple sub-files, each written in a literate Haskell format. Each of these files needs to be pre-processed by the `lhs2TeX` command to produce the `.tex` source file. Once each of these files has been generated, then the `LATEX`project can be built into a PDF file.

5.2.1 Building the Circuit

The Circuit defined here makes use of the `mapC` operator. To do so a `Circuit` is defined that is able to build a single `.tex` file from a `.lhs`. This has to make use of the standard `task` constructor:

```

buildLhsTask :: Circuit '[VariableStore] '[String] '[VariableStore String]
               '[VariableStore] '[String] '[VariableStore String]
               N1
buildLhsTask = task f Empty
  where
    f :: UUID
      → HList '[VariableStore] '[String]
      → VariableStore String
      → ExceptT SomeException IO (VariableStore String)
    f _ (HCons' (Var flnName) HNil') _ = do
      let fOutName = flnName <.> ".tex"
      lift (callCommand ("lhs2tex -o " ++ fOutName ++ " " ++ flnName))
      return (Var fOutName)

```

This Circuit makes use of the `callCommand` function from the `System.Process` library. This allows the task to execute external commands, that may not necessarily be defined in Haskell. A similar Circuit can

be defined that will compile the `.tex` files and produce a PDF.

```
buildTexTask :: String → Circuit '[VariableStore] '[[String]] '[VariableStore [String]]
                                     '[VariableStore] '[String] '[VariableStore String]
                                     N1
buildTexTask name = task f Empty
where
  f :: UUID
    → HList '[VariableStore] '[String]
    → VariableStore String
    → ExceptT SomeException IO (VariableStore String)
  f mainFileName (HCons' (Var _) HNil') _ = do
    lift
      (callCommand
        ("texfot -no-stderr latexmk -interaction=nonstopmode -pdf "
          ++ "-no-shell-escape -bibtex -jobname="
          ++ name
          ++ " "
          ++ mainFileName
        )
      )
    return (Var "dissertation.pdf")
```

The `buildTexTask` also demonstrates how it is possible to pass a global parameter into a task. This can allow tasks to be made in a more reusable way. Saving a user from defining multiple variations of the same task.

These two tasks can now be combined into a `Circuit`. This `Circuit` can be interpreted as inputting a list of `.lhs` files, which are sequentially compiled to `.tex` files by the `mapC` operator. These files are *then* built by the `buildTexTask` to produce a PDF file as output.

```
buildDiss :: String → Circuit '[VariableStore] '[[String]] '[VariableStore [String]]
                                     '[VariableStore] '[String] '[VariableStore String]
                                     N1
buildDiss name = mapC buildLhsTask Empty
                ⇔
                buildTexTask name
```

`mapC` takes two arguments, the inner circuit to execute and a pointer to the location it should store its results. In this case the output data store is a `VariableStore`, therefore, the pointer to this location is just an `Empty` variable.

5.2.2 Using the Circuit

To use this `Circuit`, a `Config` data type is used to store the information needed within the system to build the project:

```
data Config = Config
  { mainFile    :: FilePath
  , outputName :: String
  , lhsFiles    :: [FilePath]
  }
deriving (Generic, FromJSON, Show)
```

This data type uses record syntax to have name fields:

- `mainFile` is the name of the root file that should be used for compilation.
- `outputName` is the desired name for the output PDF file.
- `lhsFiles` are all the literate haskell files required to build the \LaTeX document.

The `Config` data type also derives the `Generic` and `FromJSON` instance. This allows it to be used in conjunction with a YAML file to specify these parameters. The config can be loaded with:

```
loadConfig :: IO Config
loadConfig = loadYamlSettings ["dissertation.tex-build"] [] ignoreEnv
```

An example config file can be seen in Figure 5.6.

```
mainFile: dissertation.lhs
outputName: dissertation
lhsFiles: [dissertation.lhs
           , chapters/introduction.lhs
           , chapters/background.lhs
           , chapters/the-language.lhs
           , chapters/implementation.lhs
           , chapters/evaluation.lhs
           , chapters/examples.lhs]
```

Figure 5.6: An example config file for the `lhs2TeX` build system

To be able to use the system a `main` function is defined, which will serve as the entry point to the executable:

```
main :: IO ()
main = do
  config <- loadConfig
  n <- startNetwork (buildDiss (outputName config)) :: IO
    (BasicNetwork '[VariableStore] '[String] '[VariableStore String]
      '[VariableStore] '[String] '[VariableStore String])
  write (mainFile config -<.> "tex") (HCons' (Var (lhsFiles config)) HNil) n
  _ <- read n
  stopNetwork n
```

The `main` function in the build system has 5 steps:

1. The config file is loaded.
2. A network is started based on the `buildDiss` circuit. The type of network to be started has to be annotated, so that the type system knows which `Network` instance to use.
3. The build job is input into the network, with the input values being a list of `.lhs` files to compile.
4. A call is made to the blocking function `read`, although the outputs are not needed, the call to `read` is. This prevents the program ending before the network has complete processing values.
5. Finally, the network is destroyed.

This dissertation makes use of this build system to be able include literate Haskell files.

5.3 Types saving the day

Consider an example shown in the docs for Luigi [29], that is made up of two tasks. The first generates a list of words and saves it to a file and second counts the number of letters in each of those words. The counting letters task is dependent on the words being generated.

Figure 5.7, shows an implementation of such a system, in the Python library called Luigi. However, this implementation has a very subtle bug! `GenerateWords` writes the words to a file separated by new lines, but `CountLetters` reads that same file as a comma-separated list. This shows a key flaw in this system, it is up to the programmer to ensure that they write the outputs correctly, and then that they read that same file in the same way. This error, would not even cause a run-time error, instead, it will just produce the incorrect result. For a developer this is extremely unhelpful, it means more of time is used writing tests — something that no one enjoys.

```

import luigi

class GenerateWords(luigi.Task):
    def output(self):
        return luigi.LocalTarget('words.txt')

    def run(self):
        # write a dummy list of words to output file
        words = ['apple', 'banana', 'grapefruit']

        with self.output().open('w') as f:
            for word in words:
                f.write('{word}\n'.format(word=word))

class CountLetters(luigi.Task):
    def requires(self):
        return GenerateWords()

    def output(self):
        return luigi.LocalTarget('letter_counts.txt')

    def run(self):
        # read in file as list
        with self.input().open('r') as infile:
            words = infile.read().split(',')

        # write each word to output file with its corresponding letter count
        with self.output().open('w') as outfile:
            for word in words:
                outfile.write('{word}:{letter_count}\n'.format(
                    word=word,
                    letter_count=len(word)
                ))

```

Figure 5.7: A Broken Luigi Example

The Fix Why not eliminate the need for all of this with `DataStores` and types. As previously mentioned in Section 3.2.1, a `DataStore` can be used to abstract the reading and writing of many different sources. This will help to ensure correctness of this step, by eliminating any possible duplicated code. Instead, just having the `fetch` and `save` methods to test.

The second greater benefit, is to use `DataStores` in combination with the type system. Each constructor for a `Circuit` will, enforce that the types of a `DataStore` align correctly. It would not be possible to feed the output of one task, with the type `FileStore[String]` into a task that expects a `CommaSepFile[String]`. The same example as before can be seen in Figure 5.8. In this example it will fail to compile, giving the error:

```
> Couldn't match type 'CommaSepFile' with 'FileStore'
```

This will benefit the user as it reduces the feedback loop of knowing if the program will succeed. Previously the whole data pipeline had to be run, whereas now this information can be informed to the user at compile-time.


```
generateWords :: Circuit '[VariableStore] '[(())] '[VariableStore ()]
                  '[FileStore] '[[String]] '[FileStore [String]]
                  N1
generateWords = functionTask (const ["apple", "banana", "grapefruit"]) (FileStore "fruit.txt")

countLetters :: Circuit '[CommaSepFile] '[[String]] '[CommaSepFile [String]]
                  '[FileStore] '[[String]] '[FileStore [String]]
                  N1
countLetters = functionTask (map f) (FileStore "count.txt")
  where
    f word = (concat [word, ":", show (length word)])

circuit :: Circuit '[VariableStore] '[(())] '[VariableStore ()]
                  '[FileStore] '[[String]] '[FileStore [String]]
                  N1
circuit = generateWords <=> countLetters
```

Figure 5.8: A Broken Circuit Example

Chapter 6

Benchmarks

To perform benchmarks in this Chapter, the data pre-processing pipeline that aggregated audio history will be used from Section 5.1.1. Each benchmark made will be tested on 5 different numbers of inputs: 1, 10, 100, 1000, 2000. Three months of my own audio history will be used, to ensure that the data closely aligns with the real world. This allows for the evaluation of how each implementation scales with more inputs. All benchmarks will take place on a dual-core Intel i5-6276U CPU with hyper-threading, giving 4 virtual cores.

6.1 Benchmarking Technicalities

When benchmarking Haskell, there are several important features that should be considered to ensure correct and meaningful timings are obtained.

Lazy Evaluation Haskell makes use of lazy evaluation, in many cases this is a very useful feature. However, when timing code, care needs to be taken to ensure that values are actually evaluated, within the timed block. Consider this example of timed code:

```
newtype Value = Value Int deriving Show
add :: Int → Int → IO (Value)
add x y = return (Value (x + y))
main :: IO ()
main = do
    startTime ← getTime clock
    r          ← add 1 2
    endTime   ← getTime clock
    print (diffTimeSpec endTime startTime)
    print r
```

To an untrained eye, this would appear to time how long it takes to add the two numbers together, but this isn't the case. The laziness of Haskell means that, the evaluation of `r` will not take place until `print` needs it.

Luckily, Haskell provides infrastructure to avoid these problems. `deepseq :: a → b → b` forces the full evaluation of `a`, before it returns `b`: making use of this function the evaluation of `r` is forced in the timed code. The `BangPatterns` language extension is used to force evaluation to Weak Head Normal Form (**WHNF**), when an expression is evaluated the the outermost data constructor or lambda expression. `"he" ++ "llo"` is an example of an expression not in **WHNF**, this example in **WHNF** is `'h':("e" ++ "llo")`, with the outermost constructor being `:`.

```
main :: IO ()
main = do
    startTime ← getTime clock
    r          ← add 1 2
    let !r' = r `deepseq` r
    endTime   ← getTime clock
```

```
print (diffTimeSpec endTime startTime)
print r'
```

This examples firstly uses `deepseq` to force the full evaluation of the result, however, Haskell will lazily evaluate this application of `deepseq`. This is because the result of `r `deepseq` r` is not needed until the final `print`, therefore, a bang pattern is used to force the evaluation the application of `deepseq`.

Multi-Core Haskell By default the Haskell runtime does not enable multi-core processing. Considering the aim of this project partly involves making `CircuitFlow` run in parallel, multi-core processing is crucial. To enable this the `-threaded` flag is set when building the binary. Then, using the runtime options, the number of threads can be set by adding `+RTS -N` flags when running the binary. The `-N` allows the runtime to select the optimal number of threads for the program.

6.2 Parallel vs Serial

The first test will ensure that `CircuitFlow`'s parallelisation has a positive effect on run-times. To ensure that the test is fair, the serial implementation will make use of the same tasks in the pre-processing pipeline. The inputs and outputs will just be manually fed into each task, in a sequential way. The results from this test are found in figure 6.1

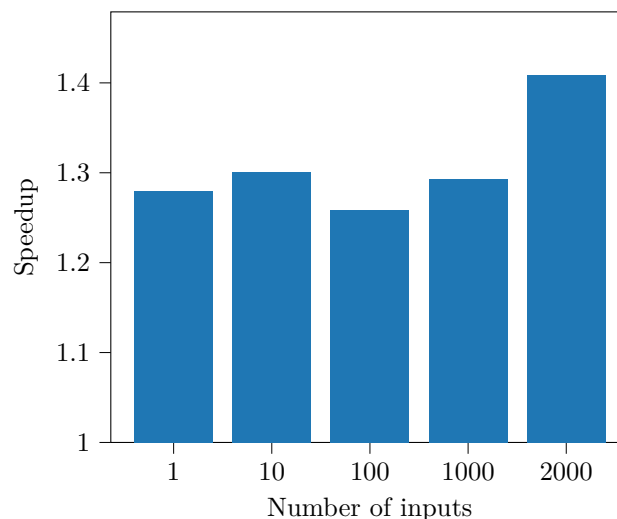


Figure 6.1: Speedup provided by using `CircuitFlow`, relative to a serial implementation

This shows that `CircuitFlow` does indeed provide a performance gain, with on average speedup of 1.3x. However, with 4 threads available a higher speedup may have been expected, as there is now 4x as much processing power, especially on larger numbers of inputs, where the channels should buffer and allow for all 4 threads to be running simultaneously.

6.2.1 Areas for improvement

Profiling the circuit shows that a significant proportion of time is spend reading CSV files. Improving the speed of parsing CSV files is outside the scope of this project, however there could be ways to optimise how often a CSV is read. For example, replicating a data store leads to the same value being read from the data store twice, the use of caching could allow for the value to only be read once. In the pipeline example, this would mean that each input CSV is only read once.

Another area for improvement is that there is an expectation on the user to know where is best to split up the workflow into tasks. With a user who is familiar with the domain this should not be too hard, but it would also be beneficial if a circuit could automatically fuse tasks together, then it would have a positive effect on the runtime.

6.2.2 An Aside: 1 Core Circuit vs Serial

Another interesting scenario to test is checking if the network structure adds additional overhead, in a situation where there is only 1 core. To test this, the multi-core support of the Haskell runtime will not be enabled: this will then simulate multiple cores with context switching. Figure 6.2, shows the results of this benchmark.

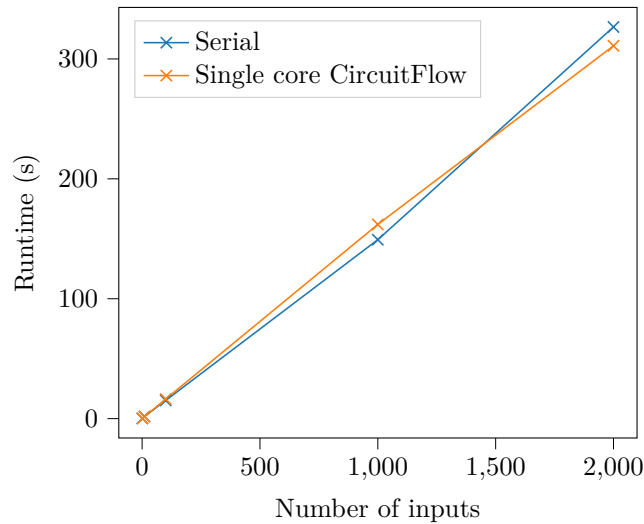


Figure 6.2: A comparison of runtimes of a serial implementation vs single core CircuitFlow

This shows that both the linear and single core implementation scale together in a linear fashion. Most importantly, neither implementation is faster than the other, therefore, it is concluded that **CircuitFlow** does not add additional overheads. This will be particularly helpful for a user which needs to run code on multiple types of devices. There is no need for them to create a different implementation for devices where parallelisation may not be possible.

6.3 CircuitFlow vs Luigi

The final benchmark on CircuitFlow is comparing it to widely used library: Luigi by Spotify [28]. Since Luigi uses a **DPN**, it can use any number of threads: in this test it is set to 4 — the same as CircuitFlow. Figure 6.3, shows the results of the benchmark.

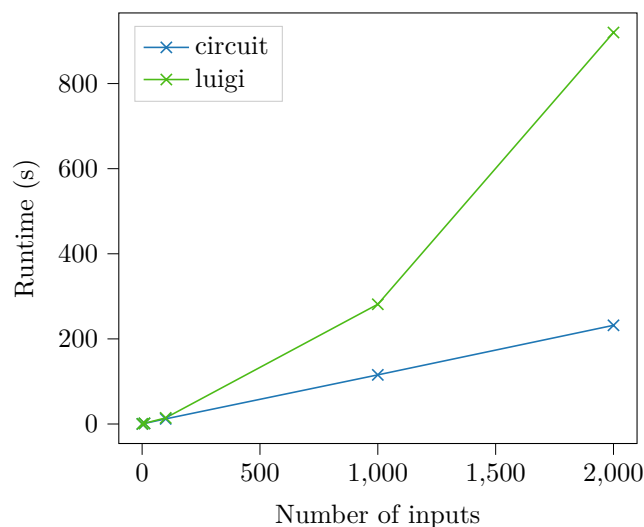


Figure 6.3: A comparison of runtimes of CircuitFlow vs Luigi

This shows clearly that CircuitFlow out performs Luigi on larger numbers of inputs. CircuitFlow

scales linearly with the number of inputs, whereas Luigi's runtime appears to grow at a quicker rate than linear. CircuitFlow always out performs Luigi on any number of inputs, this can lead to the conclusion that CircuitFlow is able to out perform another dataflow DSL — exceeding the requirements.

6.3.1 Why is CircuitFlow so good?

There are differences between CircuitFlow and Luigi which could explain these significant differences in runtime — especially for larger numbers of inputs.

Computation Models CircuitFlow uses a Kahn Process Network (KPN) to compute the results of the system: this makes use of threads and channels, with the computation on each thread set statically at compile-time. Luigi uses a variant of a KPN, known as Data Process Network (DPN): making use of a central scheduler and threads that can perform any task. One possible reason for slowdown in Luigi is this central scheduler: this scheduler checks tasks dependencies at runtime, therefore it will need to resolve the dependencies in tasks to work out which ones to execute. This is the opposite to CircuitFlow, where dependencies between tasks are resolved at compile time.

Use of Threads CircuitFlow aims to be as lightweight on threads as possible, it only starts 1 thread for each task in a circuit: this will reduce the overhead of starting new threads. Luigi, does not make use of threads, instead it creates new processes, however it does so in an inefficient way: it creates a new process for each invocation of a task, which is destroyed after. This means that in the example benchmarked with 2000 inputs, it creates 8000 new processes: this adds significant overhead that CircuitFlow does not have.

6.3.2 Why is CircuitFlow so good? (v2)

Luigi and Circuit flow have their differences, which will likely explain why there is a difference in run times, especially with larger numbers of inputs.

Computation Models The two libraries use variants of the same computation model: CircuitFlow uses a KPN and Luigi uses a DPN. This difference is the main reason why CircuitFlow scales linearly when it needs to process more input values. CircuitFlow makes use of buffered channels to keep a queue of all inputs that need to be processed. However, Luigi does not rely on this design, instead it has a pool of workers with a scheduler controlling what is executed on each worker. It is this scheduler that causes Luigi to scale non-linearly. As the number of inputs grow, the scheduler will have to schedule more and more tasks: this process is not $\mathcal{O}(n)$.

Multi-processing in Python Circuit Flow makes use of a static number of threads * defined by the number of tasks in a circuit. Luigi on the other hand can support any number of workers, however, Luigi suffers from a downfall of Python: threads cannot run in parallel due to the Global Interpreter Lock. To avoid this Luigi uses processes not threads, which adds extra overhead. Luigi also creates a new process for each invocation of a task, which CircuitFlow does not do. This means that Luigi will start 8000 processes vs CircuitFlow's 4 threads for the 2000 inputs benchmark. Luigi's static number of threads could also be considered a downside due to the lack of flexibility depending on run-time values. To combat this more combinators can be introduced that allow for branching or other similar operations, in fact, mapC is a combinator of this type.

More Lightweight There are other features that Luigi has, which Circuit flow does not, that cause CircuitFlow to gain an unfair speed advantage over Luigi — one such feature is back filling. This allows Luigi to avoid running tasks that have already been run. This feature means that before executing a task the Luigi scheduler has to check if a task has already been executed. This adds additional overhead to the scheduler that Circuit Flow does not have. Although this feature does have its benefits, after the first run of Luigi all run times after are very quick as no tasks will need to be executed. If CircuitFlow were to implement this feature any overhead it adds will be partially mitigated by the checks being distributed across multiple threads, instead of in one central scheduler.

Chapter 7

Conclusion

7.1 Related Work

7.1.1 Existing Workflow Tools

Composing Effects into Tasks and Workflows [25] Another approach taken that solves the untyped task composition problem is the Funflow library based on the techniques in the paper *Composing Effects into Tasks and Workflows*. This work notices that tasks in a workflow are similar to effects in the functional community. It draws from existing work on combining and analysing effects, with categories and arrows, and applies this to constructing workflows. It differs from CircuitFlow by using arrows to compose tasks together, whereas CircuitFlow makes use of circuits based on symmetric monoidal preorders.

SciPipe: A workflow library for agile development of complex and dynamic bioinformatics pipelines [19] Whilst CircuitFlow and many other libraries place emphasis on defining tasks as functions within the embedding language, SciPipe takes a different approach. It instead focuses on the orchestration of external jobs that are called with Bash commands, which allows for simple integration with pre-existing binaries. It also allows for the library to be able to orchestrate tasks written in the best language for its requirements.

Luigi Luigi is used to orchestrate tasks in a data workflow, it is a library that falls into the trap of untyped task dependencies. It makes use of a central scheduler and workers, allowing work to be distributed across multiple machines. It also comes with built in support for many different output formats, such as files in a Hadoop file system.

Pipes [11] The CircuitFlow examples described in this thesis focus on processing a group of jobs and waiting until they are complete before moving on, another approach is to support streaming of data. This is the route that the Haskell library pipes takes, it allows the user to build streaming pipelines, that can be composed together. This approach could also be adopted by CircuitFlow without any modifications, allowing for inputs and outputs to be streamed in, with the channels acting as buffers on the stream.

7.1.2 Indexed Functors

Staged Selective Parser Combinators [34] Indexed functors [22], are a new technique for building typed **DSLs**. This paper makes use of this new tool to have a type index representing the type of a parser. This allows it to make optimisations and translations while ensuring that the value parsed never changes.

7.2 Future Work

Performance Improvements Whilst CircuitFlow beats the competition, there are some possible improvements that could help to improve the runtime even further. One optimisation is to fuse tasks together, when the fusing would have a greater increase than the parallelisation from being left un-fused. This would require analysis of the computation inside a task, to rate the relative computational intensity of each task. Unfortunately, it is not possible to inspect the **AST** of another Haskell function, so another approach must be taken. This could involve using defunctionalisation and building a new **DSL** that can construct tasks. This would generate an **AST**, that can be traversed to measure its computational

intensity. However, this could restrict the computation a task is capable of doing. Another approach could be, introducing a new **Network** instance that is capable of profiling a network with a few example inputs. The information can then be used to perform optimisations to the original network.

Network Visualiser A network could be constructed of long running tasks, making it desirable to see progress of the network in a quick and simple way. A visualiser could be build that shows all the tasks currently being executed and the number of pending values on all the channels in the system. This could help a user find bottlenecks in the system and where backlogs of pending values are occurring.

More Networks The network system has been designed in a way that allows for multiple instances. One future task could be introducing more variants. One example has already been discussed — a profiling network. Another example could be a distributed network, which would become extremely helpful when data workflows become too big for a single machine. To do this it is likely that data stores would need to be serialisable, so that it can be transferred between different nodes in the network.

Further Benchmarks Currently the only benchmark is a pre-processing pipeline for song data aggregation, however, there are many other uses for **CircuitFlow**. More benchmarks could be defined across a wider array of uses, this will allow for further comparison to other similar systems. However, for the scope of this project it was enough to beat a leading tool by Spotify at its own game — aggregating song data.

More Circuit Smart Constructors **CircuitFlow** focuses on the transformations made to data as it passes through the system, however, there are many transformations a developer may find themselves repeating. For example, what if a user wanted to replicate an input to produce three outputs instead of two: they'd find themselves repeating `replicate <> id <> replicate` multiple times. What if they wanted 4 or more outputs? Smart constructors can be used to combat this problem, with new ones being defined for common patterns that occur. For this specific example a new smart constructor `replicateN :: SNat n → Circuit '[f] '[a] '[fa] (Repeat n f) (Repeat n a) (Repeat n (fa)) N1`, where `Repeat` is a type family that produces a list of the argument repeated `n` times.

Better Support for External Tasks The `lhs2TeX` build system made use of external tasks, however, it did so at a very raw level. More smart constructors for tasks can be introduced that allows for easier definition of external tasks. There could also be a more universal method for how inputs and outputs are handled with those external tasks, possibly through `stdin` and `stdout`, or as file.

Visual Construction All the constructors for **CircuitFlow** have graphical meanings, so why not have a graphical way to compose them? A new GUI could be introduced that allows users to drag and drop building blocks to create circuits. This would enable the library to be used by developers who may have less experience or prefer to view visual representations instead.

7.3 Summary

This thesis has introduced a new DSL to construct data workflows. It set out to be type-safe and run in a competitive time, but has exceeded expectations, proving to be almost 4x faster than **Luigi**.

The battle for type-safety was a tough one, needing to employ new and advanced Haskell techniques for creating a strongly typed program. **CircuitFlow** employs phantom type parameters consisting of data types promoted to `Kinds` by `DataKinds`, to add constraints for how each constructor can be combined. This allows for a language that is strongly typed and avoids the problems demonstrated with **Luigi** in Chapter 5. The types proving pivotal to ensuring the language's correctness, with `GHC` refusing to compile for weeks on end, until finally it compiled and worked first time.

The design of the language draws its origins from a strong mathematical background, with each constructor directly representing an axiom in a symmetric monoidal preorder. This demonstrates the language's completeness at being able to represent any **DAG**, that a data workflow may need.

The work on **CircuitFlow** is far from over though! As the future work suggested there are many additional features that would allow **CircuitFlow** to become even easier to use and more available to more use cases. **CircuitFlow** is already released as an open-source project, with the intention of developing it further.

Bibliography

- [1] Quartz composer user guide, Jul 2007.
- [2] Apache. Airflow: <https://airflow.apache.org>.
- [3] Patrick Bahr and Tom Hvitved. Compositional data types. In *Proceedings of the Seventh ACM SIGPLAN Workshop on Generic Programming*, WGP '11, page 83–94, New York, NY, USA, 2011. Association for Computing Machinery.
- [4] Dr Peter Bentley. The end of moore’s law: what happens next?, Apr 2020.
- [5] James Cheney and Ralf Hinze. First-class phantom types. 01 2003.
- [6] Bob Coecke, Tobias Fritz, and Robert W. Spekkens. A mathematical theory of resources. *Information and Computation*, 250:59–86, Oct 2016.
- [7] Edward Costello. *Figure 15. Connecting a Math patch to the Sprite Width and Height inlets*. Creating Graphical User Interfaces for Csound using Quartz Composer. Csound Journal, Nov 2012.
- [8] Richard A. Eisenberg, Dimitrios Vytiniotis, Simon Peyton Jones, and Stephanie Weirich. Closed type families with overlapping equations. In *Proceedings of the 41st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '14, page 671–683, New York, NY, USA, 2014. Association for Computing Machinery.
- [9] Richard A. Eisenberg and Stephanie Weirich. Dependently typed programming with singletons. In *Proceedings of the 2012 Haskell Symposium*, Haskell '12, page 117–130, New York, NY, USA, 2012. Association for Computing Machinery.
- [10] M.M. Fokkinga. Monadic maps and folds for arbitrary datatypes. *Memoranda informatica*, (94-28):–, June 1994. Imported from EWI/DB PMS [db-utwente:tech:0000003538].
- [11] Gabriel Gonzalez. pipes: <https://hackage.haskell.org/package/pipes>.
- [12] Felienne Hermans, Martin Pinzger, and Arie van Deursen. Breviz: Visualizing spreadsheets using dataflow diagrams, 2011.
- [13] Ralf Hinze. An algebra of scans. In Dexter Kozen, editor, *Mathematics of Program Construction*, pages 186–210, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.
- [14] Tiffany Hu. Building out the seatgeek data pipeline, Jan 2015.
- [15] Patricia Johann and Neil Ghani. Foundations for structured programming with gadts. In *Proceedings of the 35th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '08, page 297–308, New York, NY, USA, 2008. Association for Computing Machinery.
- [16] Gilles Kahn. The semantics of a simple language for parallel programming. In Jack L. Rosenfeld, editor, *Information Processing, Proceedings of the 6th IFIP Congress 1974, Stockholm, Sweden, August 5-10, 1974*, pages 471–475. North-Holland, 1974.
- [17] Dhruba Karki. Can you guess how much data is generated every day?, Nov 2020.
- [18] Oleg Kiselyov, Ralf Lämmel, and Kean Schupke. Strongly typed heterogeneous collections. In *Proceedings of the 2004 ACM SIGPLAN Workshop on Haskell*, Haskell '04, page 96–107, New York, NY, USA, 2004. Association for Computing Machinery.

- [19] Samuel Lampa, Martin Dahlö, Jonathan Alvarsson, and Ola Spjuth. SciPipe: A workflow library for agile development of complex and dynamic bioinformatics pipelines. *GigaScience*, 8(5), 04 2019. giz044.
- [20] S.M. Lane, S.J. Axler, Springer-Verlag (Nowy Jork)., F.W. Gehring, and P.R. Halmos. *Categories for the Working Mathematician*. Graduate Texts in Mathematics. Springer, 1998.
- [21] E. A. Lee and T. M. Parks. Dataflow process networks. *Proceedings of the IEEE*, 83(5):773–801, 1995.
- [22] Conor McBride. Functional pearl: Kleisli arrows of outrageous fortune. *Journal of Functional Programming (accepted for publication)*, 2011.
- [23] John C. Mitchell and Gordon D. Plotkin. Abstract types have existential type. *ACM Trans. Program. Lang. Syst.*, 10(3):470–502, July 1988.
- [24] G. E. Moore. Cramming more components onto integrated circuits, reprinted from electronics, volume 38, number 8, april 19, 1965, pp.114 ff. *IEEE Solid-State Circuits Society Newsletter*, 11(3):33–35, 2006.
- [25] Yves Parès, Jean-Philippe Bernardy, and Richard A. Eisenberg. Composing effects into tasks and workflows. In *Proceedings of the 13th ACM SIGPLAN International Symposium on Haskell*, Haskell 2020, page 80–94, New York, NY, USA, 2020. Association for Computing Machinery.
- [26] Simon Peyton Jones, Dimitrios Vytiniotis, Stephanie Weirich, and Geoffrey Washburn. Simple unification-based type inference for gadts. *SIGPLAN Not.*, 41(9):50–61, September 2006.
- [27] Tom Schrijvers, Simon Peyton Jones, Manuel Chakravarty, and Martin Sulzmann. Type checking with open type functions. In *Proceedings of the 13th ACM SIGPLAN International Conference on Functional Programming*, ICFP ’08, page 51–62, New York, NY, USA, 2008. Association for Computing Machinery.
- [28] Spotify. Luigi: <https://github.com/spotify/luigi>.
- [29] Spotify. Tasks, Apr 2020.
- [30] Josef Svenningsson and Emil Axelsson. Combining deep and shallow embedding of domain-specific languages. *Computer Languages, Systems & Structures*, 44:143–165, 2015. SI: TFP 2011/12.
- [31] WOUTER SWIERSTRA. Data types à la carte. *Journal of Functional Programming*, 18(4):423–436, 2008.
- [32] Tableau. Tableau prep builder & prep conductor: A self-service data preparation solution, 2021.
- [33] Phillip Wadler. The expression problem, Nov 1998.
- [34] Jamie Willis, Nicolas Wu, and Matthew Pickering. Staged selective parser combinators. *Proc. ACM Program. Lang.*, 4(ICFP), August 2020.
- [35] Nicolas Wu. Yoda: A simple combinator library, 2018.
- [36] Brent A. Yorgey, Stephanie Weirich, Julien Cretin, Simon Peyton Jones, Dimitrios Vytiniotis, and José Pedro Magalhães. Giving haskell a promotion. In *Proceedings of the 8th ACM SIGPLAN Workshop on Types in Language Design and Implementation*, TLDI ’12, page 53–66, New York, NY, USA, 2012. Association for Computing Machinery.