

# CONTEXT-AWARE PROGRAMMING LANGUAGES

TOMAS PETRICEK

Computer Laboratory  
University of Cambridge

2014



## ABSTRACT

---

The development of programming languages needs to reflect important changes in the way programs execute. In recent years, this has included e.g. the development of parallel programming models (in reaction to the multi-core revolution) or improvements in data access technologies. This thesis is a response to another such revolution – the diversification of devices and systems where programs run.

The key point made by this thesis is the realization that an execution environment or a *context* is fundamental for writing modern applications and that programming languages should provide abstractions for programming with context and verifying how it is accessed.

We identify a number of program properties that were not connected before, but model some notion of context. Our examples include tracking different execution platforms (and their versions) in cross-platform development, resources available in different execution environments (e.g. GPS sensor on a phone and database on the server), but also more traditional notions such as variable usage (e.g. in liveness analysis and linear logics) or past values in stream-based data-flow programming.

Our first contribution is the discovery of the connection between the above examples and their novel presentation in the form of calculi (*coeffect systems*). The presented type systems and formal semantics highlight the relationship between different notions of context. Our second and third contributions are two unified coeffect calculi that capture commonalities in the presented examples. In particular, our *flat coeffect calculus* models languages with contextual properties of the execution environment and our *structural coeffect calculus* models languages where the contextual properties are attached to the variable usage. We close the thesis with a *unified coeffect calculus* that captures the two notions using a single parameterized system.

Although the focus of this thesis is on the syntactic properties of the presented systems, we also discuss their category-theoretical motivation. We introduce the notion of an *indexed comonad* (based on dualisation of the well-known monad structure) and show how they provide semantics of the two coeffect calculi.



## CONTENTS

1	WHY CONTEXT-AWARE PROGRAMMING MATTERS	1
1.1	Why context-aware programming matters . . . . .	2
1.1.1	Context awareness #1: Platform versioning . . . . .	3
1.1.2	Context awareness #2: System capabilities . . . . .	4
1.1.3	Context awareness #3: Confidentiality and provenance . . . . .	5
1.1.4	Context-awareness #4: Checking array access patterns . . . . .	5
1.2	Towards context-aware languages . . . . .	6
1.2.1	Context-aware languages in action . . . . .	7
1.2.2	Understanding context with types . . . . .	7
1.3	Theory of context dependence . . . . .	9
1.4	Thesis outline . . . . .	11
2	PATHWAYS TO COEFFECTS	13
2.1	Through type and effect systems . . . . .	13
2.1.1	Simple effect system. . . . .	13
2.1.2	Simple coeffect system. . . . .	14
2.2	Through language semantics . . . . .	15
2.2.1	Effectful languages and meta-languages . . . . .	15
2.2.2	Marriage of effects and monads . . . . .	17
2.2.3	Context-dependent languages and meta-languages . . . . .	17
2.3	Through substructural and bunched logics . . . . .	21
2.3.1	Substructural type systems. . . . .	21
2.3.2	Bunched type systems. . . . .	22
2.4	Context oriented programming . . . . .	23
2.5	Summary . . . . .	23
3	CONTEXT-AWARE SYSTEMS	25
3.1	Structure of coeffect systems . . . . .	25
3.1.1	Effectful lambda abstraction . . . . .	26
3.1.2	Notions of context . . . . .	26
3.1.3	Scalars and vectors . . . . .	28
3.2	Flat coeffect systems . . . . .	29
3.2.1	Implicit parameters and type classes . . . . .	29
3.2.2	Distributed computing . . . . .	34
3.2.3	Liveness analysis . . . . .	37
3.2.4	Data-flow languages . . . . .	41
3.2.5	Permissions and safe locking . . . . .	46
3.3	Structural coeffect systems . . . . .	47
3.3.1	Liveness analysis revisited . . . . .	47
3.3.2	Bounded variable use . . . . .	52
3.3.3	Data-flow languages revisited . . . . .	55
3.3.4	Security, tainting and provenance . . . . .	58
3.4	Beyond passive contexts . . . . .	59
3.5	Summary . . . . .	61
4	THE FLAT COEFFECT CALCULUS	63
4.1	Introduction . . . . .	63
4.1.1	Related work . . . . .	64
4.2	Flat coeffect calculus . . . . .	64
4.2.1	Reconciling lambda abstraction . . . . .	64
4.2.2	Understanding flat coeffects . . . . .	65
4.2.3	Flat coeffect algebra . . . . .	65

4.2.4	Flat coeffect types . . . . .	67
4.2.5	Examples of flat coeffects . . . . .	68
4.2.6	Typing of let binding . . . . .	69
4.3	Categorical motivation . . . . .	70
4.3.1	Categorical semantics . . . . .	70
4.3.2	Introducing comonads . . . . .	71
4.3.3	Generalising to indexed comonads . . . . .	72
4.3.4	Flat indexed comonads . . . . .	73
4.3.5	Properties and related notions . . . . .	76
4.3.6	Semantics of flat calculus . . . . .	77
4.4	Syntactic equational theory . . . . .	79
4.4.1	Syntactic properties . . . . .	79
4.4.2	Call-by-value evaluation . . . . .	81
4.4.3	Call-by-name evaluation . . . . .	82
4.4.4	Internalized sequencing . . . . .	85
4.5	Syntactic properties and extensions . . . . .	87
4.5.1	Subtyping for coeffects . . . . .	88
4.5.2	Properties of lambda abstraction . . . . .	88
4.5.3	Language with pairs and unit . . . . .	90
4.6	Related work . . . . .	91
4.6.1	When is coeffect not a monad . . . . .	91
4.6.2	When is coeffect a monad . . . . .	91
4.7	Conclusions . . . . .	93
5	THE STRUCTURAL COEFFECT CALCULUS . . . . .	95
5.1	Introduction . . . . .	95
5.1.1	Related work . . . . .	96
5.2	Structural coeffect calculus . . . . .	96
5.2.1	Structural coeffect algebra . . . . .	97
5.2.2	Structural coeffect types . . . . .	98
5.2.3	Understanding structural coeffects . . . . .	99
5.2.4	Examples of structural coeffects . . . . .	100
5.3	Categorical motivation . . . . .	101
5.3.1	Semantics of vectors . . . . .	101
5.3.2	Indexed comonads, revisited . . . . .	102
5.3.3	Structural indexed comonads . . . . .	103
5.3.4	Semantics of structural caluculus . . . . .	104
5.3.5	Examples of structural indexed comonads . . . . .	106
5.4	Syntactic equational theory . . . . .	109
5.4.1	From flat coeffects to structural coeffects . . . . .	109
5.4.2	Holes and substitution lemma . . . . .	110
5.4.3	Reduction and expansion . . . . .	112
5.5	Conclusions . . . . .	113
6	UNIFIED COEFFECT LANGUAGES . . . . .	115
6.1	The unified coeffect calculus . . . . .	115
6.1.1	Shapes and containers . . . . .	116
6.1.2	Structure of coeffects . . . . .	116
6.1.3	Unified coeffect type system . . . . .	119
6.1.4	Structural coeffects . . . . .	121
6.1.5	Flat coeffects . . . . .	122
6.1.6	Semanitics of unified calculus . . . . .	124
6.2	Towards practical coeffects . . . . .	125
6.2.1	Embedding contextual computations . . . . .	126
6.2.2	Coeffect annotations as types . . . . .	126

6.2.3	Alternative formulation using coeffect lattice . . . . .	127
6.3	Coeffect meta-language . . . . .	129
6.3.1	Coeffects and contextual modal type theory . . . . .	129
6.3.2	Coeffect meta-language . . . . .	130
6.3.3	Embedding flat coeffect calculus . . . . .	131
6.4	Conclusions . . . . .	132
7	CONCLUSIONS . . . . .	133
7.1	Unified presentation. . . . .	133
7.2	Flat coeffect calculus. . . . .	133
7.3	Structural coeffect calculus. . . . .	134
7.4	Summary . . . . .	134
	BIBLIOGRAPHY . . . . .	135
A	APPENDIX A . . . . .	143
A.1	Coeffect typing for implicit parameters . . . . .	143
A.2	Coeffect typing for liveness . . . . .	143
A.3	Coeffect typing for data-flow . . . . .	144
B	APPENDIX B . . . . .	149
B.1	Substitution for flat coeffects . . . . .	149
B.2	Substitution for structural coeffects . . . . .	151





## WHY CONTEXT-AWARE PROGRAMMING MATTERS

---

Many advances in programming language design are driven by practical motivations. Sometimes, these practical motivations are easy to see – for example, when they come from an external change such as the rise of multi-core processors. Sometimes, discovering the practical motivations is a difficult task – perhaps because we are so used to a certain way of doing things that we do not even *see* the flaws of our approach.

Before exploring the motivations for to this thesis, we briefly consider two recent practical concerns that have led to the development of new programming languages. This helps to explain why context-aware programming is important. The examples are by no means representative, but they illustrate various kinds of motivations well.

**PARALLEL PROGRAMMING.** The rise of multi-core CPUs is a clear example of an external development influencing programming language research. As multi-core and multi-processor systems became ubiquitous, languages had to provide better abstractions for parallel programming. This led to the industrial popularity of *immutable* data structures (and functional programming in general), software transactional memory [40], data-parallelism and also asynchronous computing [85].

In this case, the motivation is easy to see – writing multi-core programs using earlier abstractions, such as threads and locks, is difficult and error-prone. At the same time, multi-core CPUs became a standard very quickly and so the lack of good language abstractions was apparent.

**DATA ACCESS.** Accessing “big data” sources is an example of a more subtle challenge. Initiatives like open government data<sup>1</sup> certainly make more data available. However, to access the data, one has to parse CSV and Excel files, issue SQL or SPARQL queries (to query database and the semantic web, respectively).

Technologies like LINQ [54] make querying data significantly easier. But perhaps because accessing data became important more gradually, it was not easy to see that SQL queries, embedded as parameterized strings<sup>2</sup>, are a poor solution *before* better approaches were developed.

This is even more the case for *type providers* – a recent feature in F# that integrates external data sources directly into the type system of the language and thus makes data explorable directly from the source code editor (through features such as auto-completion on object members). It is not easy to see the limitations of standard techniques (using HTTP requests to query REST services or parsing CSV files and using string-based lookup) until one sees just how much type providers change the data-scientist’s workflow<sup>3</sup>.

---

<sup>1</sup> In the UK, the open government data portal is available at: <http://data.gov.uk/>

<sup>2</sup> The dominant approach is demonstrated, for example, by a review of SQL injection prevention techniques by Clarke [19]

<sup>3</sup> This is difficult to explain in writing and so the reader is encouraged to watch a video showing type providers for the WorldBank and CSV data sources [69].

CONTEXT-AWARE PROGRAMMING. In this thesis, we argue that the next important practical challenge for programming language designers is designing languages that are better at working with (and understanding) the *context in which programs are executed*.

This challenge is of the kind that is not easy to see, perhaps because we are so used to doing things in certain ways that we cannot see their flaws. In this chapter, we aim to expose such flaws. We look at a number of basic programs that rely on contextual information, we explain why the currently dominant solutions are inappropriate and then briefly outline how this thesis solves the problems.

Putting deeper philosophical questions about the nature of scientific progress aside, the goal of programming language research is generally to design languages that provide more *appropriate abstractions* for capturing common problems, are *simple* and more *unified*. These are exactly the aims that we follow in this thesis. In this chapter, we explain what the common problems in context-dependent programming are. In Chapter 4 and Chapter 5, we develop two simple calculi to understand and capture the structure of those problems and, finally, Chapter 6 unifies the two abstractions.

### 1.1 WHY CONTEXT-AWARE PROGRAMMING MATTERS

The phrase *context in which programs are executed* sounds rather abstract and generic. What notions of *context* can be identified in modern software systems? Different environments provide different resources (e. g. a database or GPS sensors), environments are increasingly diverse (e. g. different mobile platforms with multiple partially incompatible versions). Web applications are split between client, server and mobile components; mobile applications must be aware of the physical environment while the “internet of things” makes the environment even more heterogeneous. At the same time, applications access rich data sources and need to be aware of provenance information and respect the security policies from the environment.

Writing such context-aware (or environment-aware) applications is a fundamental problem of modern software engineering. The state of the art relies on ad-hoc approaches – using hand-written conditions or pre-processors for conditional compilation. Common problems that developers face include:

- **SYSTEM CAPABILITIES.** Libraries such as LINQ [54] let developers write code in a host language like C# and then cross-compile it to multiple targets (including SQL, OpenCL or JavaScript [52]). Part of the compilation (e. g. generating the SQL query) occurs at runtime and developers have no guarantee that it will succeed until the program is executed, because only subset of the host language is supported.
- **PLATFORM VERSIONS.** When developing cross-platform applications, different platforms (and different versions of the same platform) provide different API functions. Writing a cross-platform code usually relies on (fragile) conditional compilation or (equally fragile) dynamic loading.
- **SECURITY AND PROVENANCE.** When working with data (be it sensitive database or social network data), we may have permission to access only some of the data and we may want to track *provenance* information. However, this is not checked – if a program attempts to access unavailable data, the access will be refused at run-time.

```

for header, value in header do
    match header with
    | "accept" → req.Accept ← value
#if FX_NO_WEBREQUEST_USERAGENT
    | "user-agent" → req.UserAgent ← value
#else
    | "user-agent" → req.Headers.[HttpHeader.UserAgent] ← value
#endif
#if FX_NO_WEBREQUEST_REFERER
    | "referer" → req.Referer ← value
#else
    | "referer" → req.Headers.[HttpHeader.Referer] ← value
#endif
    | other → req.Headers.[other] ← value

```

Figure 1: Conditional compilation in the HTTP module of the F# Data library

- **RESOURCES & DATA AVAILABILITY.** When creating a mobile application, the program may (or may not) be granted access to device capabilities such as GPS sensor, social updates or battery status. We would like to know which of the capabilities are required and which are optional (i. e. enhance the user experience, but there is a fallback strategy). Equally, on the server-side, we might have access to different database tables, depending on the role of the user.

Most developers do not perceive the above as programming language flaws. They are simply common programming problems (at most somewhat annoying and tedious) that have to be solved. However, this is because it is not apparent that a suitable language extension could make the above problems significantly easier to solve. As the number of distinct contexts and their diversity increases, these problems will become even more commonplace.

The following sub-sections explore 4 examples in more detail. The examples are chosen to demonstrate two distinct forms of contexts that are studied in this thesis – first two are related to the program environment and the latter two are associated with individual variables of the program.

#### 1.1.1 Context awareness #1: Platform versioning

The diversity across devices means that developers need to target an increasing number of platforms and possibly also multiple versions of each platform. For Android, there is a number called API level [36] which “uniquely identifies the framework API revision offered by a version of the Android platform”. Most changes in the libraries (but not all) are additive.

Equally, in the .NET ecosystem, there are multiple versions of the .NET runtime, mobile and portable versions of the framework etc. The differences may be subtle – for example, some instance methods and properties are omitted to make the mobile version of the library smaller, some functionality is not available at all, but naming can also vary between versions.

For example, the Figure 1 shows an excerpt from the Http module in the F# Data library<sup>4</sup>. The example uses conditional compilation to target

<sup>4</sup> The file version shown here is available at: <https://github.com/fsharp/FSharp.Data/blob/b4c58f4015a63bb9f8bb4449ab93853b90f93790/src/Net/Http.fs>

multiple versions of the .NET framework. Such code is difficult to write – to see whether a change is correct, it had to be recompiled for all combinations of pre-processor flags – and maintaining the code is equally hard. The above example could be refactored and the .NET API could be cleaner, but the fundamental issue remains. If the language does not understand the context (here, the different platforms and platform versions), it cannot provide any static guarantees about the code.

As an alternative to conditional compilation, developers can use dynamic loading. For example, on Android, programs can access API from higher level platform dynamically using techniques like reflection and writing wrappers. This is even more error prone. As noted in an article<sup>5</sup> introducing the technique “Remember the mantra: if you haven’t tried it, it doesn’t work.” Again, it would be reasonable to expect that statically-typed languages can provide a better solution.

### 1.1.2 Context awareness #2: System capabilities

Another example related to the previous one is when libraries use meta-programming techniques, such as LINQ [54, 18] or F# quotations [82], to translate code written in a subset of a host language to some other target language, such as SQL, OpenCL or JavaScript. For database access, this is a recently developed technique replacing embedded SQL discussed in the introduction, but it is a more broadly applicable technique for programming in heterogeneous environments. It lets developers targets multiple runtimes that have limited execution capabilities.

For example, the following LINQ query written in C# queries a database and selects those product names where the first upper case letter is “C”:

```
var db = new NorthwindDataContext();
from p in db.Products
where p.ProductName.First(λc → Char.IsUpper(c)) == "C"
select p.ProductName;
```

This appears as a perfectly valid code and the C# compiler accepts it. However, when the program is executed, it fails with the following error:

```
Unhandled Exception: System.NotSupportedException: Sequence
operators not supported for type System.String.
```

The problem is that LINQ can only translate a *subset* of normal C# code. The above snippet uses the `First` method to iterate over characters of a string, which is not supported. This is not a technical limitation of LINQ, but a fundamental problem of the approach.

When cross-compiling to a limited environment, we cannot always support the full source language. The example with LINQ and SQL demonstrates the importance of this problem. As of March 2014, Google search returns 11,800 results for the message above and even more results (44,100) for a LINQ error message “*Method X has no supported translation to SQL*” caused by a similar limitation.

<sup>5</sup> Retrieved from: <http://android-developers.blogspot.com/2009/04/backward-compatibility-for-android.html>

### 1.1.3 Context awareness #3: Confidentiality and provenance

The previous two examples were related to the non-existence of some library functions in a different execution environment. Another common factor was that they were related to the execution context of the whole program or a function scope. However, contextual properties can also be associated with specific variables.

For example, consider the following code sample that accesses a database by building a SQL query using string concatenation. For the purpose of the demonstration, this example does not use LINQ, but an older approach with a parameterized SQL query written as a string:

```
let query = sprintf "SELECT * FROM Products WHERE Name='%s'" name
let cmd = new SqlCommand(query)
let reader = cmd.ExecuteReader()
```

The code compiles without error, but it contains a major security flaw called *SQL injection* [19]. An attacker could enter `''; DROP TABLE Products --` as their name and delete the database table “Products”! For this reason, most libraries discourage building SQL commands by string concatenation, but there are still many systems that do so.

Again, this example demonstrates a more general property. Sometimes, it is desirable to track additional meta-data about variables that are in some ways special. Such meta-data can determine how the variables can be used. Here, `name` comes from the user input. This information about the value should be propagated to `query`. The `SqlCommand` object should then require arguments that can not directly contain user input (in an unchecked form).

Similarly, if we had `password` or `creditCard` variables in a client/server web application, these should be annotated as sensitive and it should not be possible to send their values over an unsecured network connection.

In the security context, such marking of values (but at run-time) is called *tainting* [39], but the technique is a special case of more general *provenance* tracking [17]. This can be useful when working with data in other contexts. For example, data journalists might want to propagate meta-data about the quality and the information source – is the source trustworthy? Is the data up-to-date? Such meta-data could propagate to the result and tell us important information about the calculated results.

### 1.1.4 Context-awareness #4: Checking array access patterns

The final example leaves the topic of cross-platform and distributed computing. We focus on checking how arrays are accessed. This is a simpler version of the data-flow programming examples used later in the thesis.

Consider a simple programming language with arrays where  $n^{th}$  element of an array `arr` is accessed using `arr[n]`. We focus on writing stencil computations (such as image blurring, Conway’s game of life or convolution) where all arrays are of the same size and the system provides a *cursor* pointing to a current location in the stencil. We assume that the keyword `cursor` returns the current location in the stencil.

The following example implements a simple one-dimensional cellular automaton, reading from the input array and writing to output:

```
let sum = input[cursor - 1] + input[cursor] + input[cursor + 1]
if sum = 2 || (sum = 1 && input[cursor - 1] = 0)
then output[cursor] ← 1 else output[cursor] ← 0
```

In this example, we use the term *context* to refer to the values in the array around the current location provided by `cursor`. The interesting question is, how much of the context (i. e. how far in the array) does the program access.

This is contextual information attached to individual (array) variables. In the above example, we want to track that input is accessed in the range  $\langle -1, 1 \rangle$  while output is accessed in the range  $\langle 0, 0 \rangle$ . When calculating the ranges, we need to be able to compose ranges  $\langle -1, -1 \rangle$ ,  $\langle 0, 0 \rangle$  and  $\langle 1, 1 \rangle$  (based on the three accesses on the first line).

The information about access patterns can be used to efficiently compile the computation by preallocating the necessary space (as we know which sub-range of the array might be accessed). It also allows better handling of boundaries [64]. For example, to simplify wrap-around behaviour we could pad the input with a known number of elements from the other side of the array.

## 1.2 TOWARDS CONTEXT-AWARE LANGUAGES

The four examples presented in the previous section cover different kinds of *context*. The context includes notions such as execution environment, capabilities provided by the environment or input and meta-data about the input and variables through which it is accessed.

The different applications can be broadly classified into two categories – those that speak about the environment and those that speak about individual inputs (variables). In this thesis, we refer to them as *flat coeffects* and *structural coeffects*, respectively:

- **FLAT COEFFECTS** – represent additional data, resources and meta-data that are available in the execution environment (regardless of how they are accessed in a program). Examples include resources such as GPS sensors and battery status (on a phone), databases (on the server), or software framework (or library) version.
- **STRUCTURAL COEFFECTS** – capture additional meta-data related to inputs. This can include provenance (source of the input value), usage information (how often is the value accessed and in what ways) or security information (whether it contain sensitive data or not).

This thesis follows the tradition of statically typed programming languages. As such, we attempt to capture such contextual information in the type system of context-aware programming languages. The type system should provide both safety guarantees (as in the first three examples) and also static analysis useful for optimization (as in the last example).

Although the main focus of this thesis is on the underlying theory of *coeffects* and on their structure, the following section briefly demonstrates the features that a practical context-aware language, based on the theory of coeffects, can provide.

```

let fetchNews(loc) =
  let cmd = sprintf "SELECT * FROM News WHERE Location='%s'" loc
  query(cmd, password)

let fetchLocalNews() =
  let loc = gpsLocation()
  remote fetchNews(loc)

let iPhoneMain() =
  createiPhoneListing(fetchLocalNews)

let windowsMain() =
  createWindowsListing(fetchLocalNews)

```

---

Figure 2: Client/server news reader app implemented in a context-aware language

### 1.2.1 Context-aware languages in action

As an example, consider a news reader app consisting of a server-side component (which stores the news in an SQL database) and a number of clients applications for popular platforms (Android, Windows Phone, etc.). A simplified code excerpt that might appear somewhere in the implementation is shown in Figure 2.

We assume that the language supports cross-compilation and splits the single program into three components: one for the server-side and two for the client-side, for iPhone and Windows platforms, respectively. The cross-compilation could be done in a way similar to Links [21], but we do not require explicit annotations specifying the target platform.

If we were writing the code using current mainstream technologies, we would have to create three completely separate components. The server-side would include the `fetchNews` function, which queries the database. The iPhone version would include `fetchLocalNews`, which gets the current GPS location and performs a call to the remote server and `iPhoneMain`, which constructs the user-interface. For Windows, we would also need `fetchLocalNews`, but this time with `windowsMain`. When using a language that can be compiled for all of the platforms, we would need a number of `#if` blocks to delimit the platform-specific parts.

To support cross-compilation, the language needs to be context-aware. Each of the function has a number of context requirements. The `fetchNews` function needs to have access to a database; `fetchLocalNews` needs access to a GPS sensor and to a network (to perform the remote call). However, it does not need a specific platform – it can work on both iPhone and Windows. The last two platform-specific functions inherit the requirements of `fetchLocalNews` and additionally also require a specific platform.

### 1.2.2 Understanding context with types

The approach advocated in this thesis is to track information about context requirements using the type system. To make this practical, the system should also provide at least a partial support for automatic type inference, as the information about context requirements makes the types more complex. An inspiring example might be the F# support for units of measure [46] –



the user has to explicitly annotate constants, but the rest of the information is inferred automatically.

Furthermore, integrating contextual information into the type system can provide information for modern developer tools. For example, many editors for F# display inferred types when placing mouse pointer over an identifier. For `fetchLocalNews`, the tip could appear as follows:

#### **fetchLocalNews**

```
unit @ { gps, rpc } → (news list) async
```

Here, we use the notation  $\tau_1 @ c \rightarrow \tau_2$  to denote a function that takes an input of type  $\tau_1$ , produces a result of type  $\tau_2$  and has additional context requirements specified by  $c$ . In the above example, the annotation  $c$  is simply a set of required resources or capabilities. However, a more complex structure could be used as well, for example, including the Android API level as an integer.

The following summary shows the types of the functions from the code sample in Figure 2. These guide code generation by specifying which function should be compiled for which of the platforms, but they also provide documentation for the developers. In addition to function annotations, we also show the annotation attached to the password variable:

```
password      : string @ sensitive
fetchNews     : location @ { database } → news list

gpsLocation   : unit @ { gps } → location
fetchLocalNews : location @ { gps, rpc } → news list

iPhoneMain    : unit @ { ios, gps, rpc } → unit
windowsMain   : unit @ { windows, gps, rpc } → unit
```

The example combines two separate notions of context. The variable `password` is annotated with a single (per-variable) annotation specifying tainting while functions are annotated with a set of resource requirements.

The concrete syntax used here is just for illustration. Furthermore, some information could even be mapped to other visual representations – for example, differently coloured backgrounds for platform-specific functions. The key point is that the type provides a number of useful information:

- The `password` variable is available in the context (we assume it has been declared earlier), but is marked as `sensitive`, which restricts how it can be used. In particular, we cannot return it as a result of a function that is called via a remote call (e.g. `fetchNews`) as that would leak sensitive data over an unsecured connection.
- The `fetchNews` function requires database access and so it can only run on the server-side (or on a thick client with a local copy of the database, such as a desktop computer with an offline mode).
- The `gpsLocation` function accesses the GPS sensor and since we call it in from `fetchLocalNews`, this function also requires GPS (the requirement is propagated automatically).
- We can compile the program for two client-side platforms - the entry points are `iPhoneMain` and `windowsMain` and require iOS and Windows user-interface libraries, together with GPS and the ability to perform remote calls over the network.



The details of how the cross-compilation would work are out of the scope of this thesis. However, one can imagine that the compiler would take multiple sets of references (representing the different platforms), expose the *union* of the functions, but annotate each with the required platform. Then, it would produce multiple different binaries – here, one for the server-side (containing `fetchNews`), one for iPhone and one for Windows.

In this scenario, the main benefit of using an integrated context-aware language would be the ability to design appropriate abstractions using standard mechanisms of the language. For cross-compilation, we can structure code using functions, rather than relying on `#if` directives. Similarly, the splitting between client-side, server-side and shared code can be done using ordinary functions and modules (with shared functions reused) – rather than having to split the application into separate independent libraries or projects.

The purpose of this section was to show that many modern programs rely on the context in which they execute in non-trivial ways. Thus designing context-aware languages is an important practical problem for language designers. The sample serves more as a motivation than as a technical background for this thesis. We explore more concrete examples of properties that can be tracked using the systems developed in this thesis in Chapter 3.

### 1.3 THEORY OF CONTEXT DEPENDENCE

The previous section introduced the idea of context-aware languages from the practical perspective. As already discussed, we approach the problem from the perspective of statically typed programming languages. This section outlines how can contextual information be integrated into the standard framework of static typing. This section is intended only as an informal overview and complete review of related work is available in Chapter 2.

**TYPE SYSTEMS.** A type system is a form of static analysis that is usually specified by *typing judgements* such as  $\Gamma \vdash e : \tau$ . The judgement specifies that, given some variables described by the context  $\Gamma$ , the expression  $e$  has a type  $\tau$ . The variable context  $\Gamma$  is necessary to determine the type of expressions. Consider an expression  $x + y$ . In many languages, including Java, C# and F#, the type could be `int`, `float`, or even `string`, depending on the types of the variables. For example, the following is a valid typing judgement in F# [88]:

$$x:\text{int}, y:\text{int} \vdash x + y : \text{int}$$

This judgement assumes that the type of both  $x$  and  $y$  is `int` and so the result must also be `int`. In F#, the expression would also be typeable in a context  $x:\text{string}, y:\text{string}$ , but not, for example, in a context where  $x$  has a type `int` and  $y$  has a type `string`.

**TRACKING EVALUATION EFFECTS.** Type systems can be extended in numerous ways. The types can be more precise, for example, by specifying the range of an integer. However, it is also possible to track what program *does* when executed. In ML-like languages, the following is a valid judgement:

$$x:\text{int} \vdash \text{print } x : \text{unit}$$

The judgement states that the expression `print x` has a type `unit`. This is correct, but it ignores the important fact that the expression has a *side-effect* and prints a number to the console. In purely functional languages, this would

not be possible. For example, in Haskell, the type would be `IO unit` meaning that the result is a *computation* that performs I/O effects and then returns unit value. Here, we look at another option for tracking effects, which is to extend the judgement with additional information about the effects. The judgement in a language with effect system would look as follows:

$$x:\text{int} \vdash \text{print } x : \text{unit} \ \& \ \{\text{console}\}$$

Effect systems add *effect annotation* as another component of the typing judgement. In the above example, the return type is unit, but the effect annotation informs us that the expression also accesses console as part of the evaluation. To track such information, the compiler needs to understand the effects of primitive built-in functions – such as `print`.

The crucial part of type systems is dealing with different forms of composition. Assume we have a function `read` that reads from the console and a function `send` that sends data over the network. The type system should correctly infer that the effects of an expression `send(read())` are `{console, network}`.

Effect systems are an established idea, but they are suitable only for tracking properties of a certain kind. They can be used for properties that describe how programs *affect* the environment. For context-aware languages, we instead need to track what programs *require* from the environment.

**TRACKING CONTEXT REQUIREMENTS.** The systems for tracking of context requirements developed in this thesis are inspired by the idea of effect systems. To demonstrate our approach, consider the following call from the sample program shown earlier – first using standard ML-like type system:

$$\text{password}:\text{string}, \text{cmd}:\text{string} \vdash \text{query}(\text{cmd}, \text{password}) : \text{news list}$$

The expression queries a database and gets back a list of news values as the result. Recall from the earlier discussion that there are two contextual information that are desirable to track for this expression. First, the call to the `query` primitive requires *database access*. Second, the `password` argument needs to be marked as *sensitive value* to avoid sending it over an unsecure network connection. The *coeffect systems* developed in this thesis capture this information in the following way (we slightly refine the notation later):

$$\begin{aligned} &(\text{password}:\text{string} @ \text{ sensitive}, \text{cmd}:\text{string}) @ \{\text{database}\} \\ &\vdash \text{query}(\text{cmd}, \text{password}) : \text{news list} \end{aligned}$$

Rather than attaching the annotation to the *resulting type*, we attach them to the variable context  $\Gamma$ . In other words, coeffect systems do not keep track just of the variables available in the context – they also capture detailed information about the execution environment. In the above example, the system tracks meta-data about the variables and annotates `password` as sensitive. Furthermore, it tracks requirements about the execution environment, for example, that the execution requires an access to database.

The example demonstrates the two kinds of coeffect systems outlined earlier. The tracking of *whole-context* information (such as environment requirements) is captured by the *flat coeffect calculus* developed in Chapter 4, while the tracking of *per-variable* information is captured by the *structural coeffect calculus* developed in Chapter 5.

It is well-known fact that *effects* correspond to *monads* and languages such as Haskell use monads to provide a limited form of effect system. An interesting observation made in this thesis is that *coeffects*, or systems for tracking contextual information, correspond to the category theoretical dual of monads called *comonads*. The details are explained throughout the thesis.

## 1.4 THESIS OUTLINE

The key claim of this thesis is that programming languages need to provide better ways of capturing how programs rely on the context, or execution environment, in which they execute. This chapter shows why this is an important problem. We looked at a number of properties related to context that are currently handled in ad-hoc and error-prone ways. Next, we considered the properties in a simplified, but realistic example of a client/server application for displaying local news.

Tracking of contextual properties may not be initially perceived as a major problem – perhaps because we are so used to write code in certain ways that prevent us from seeing the flaws. The purpose of this chapter was to expose the flaws and convince the reader that there should be a better solution. Finding the foundations of such better solution is the goal of this thesis:

- In Chapter 2, we give an overview of related work. Most importantly, we show that the idea of context-aware computations can be naturally approached from a number of directions developed recently in the theory of programming languages (including type and effect systems, categorical semantics and sub-structural logics).
- In Chapter 3, we present the first contribution of the thesis – the discovery of the connection between a number of existing programming language features that are related to context. The chapter presents type systems and semantics for a number of systems and analyses (including data-flow, liveness analysis, distributed programming and Haskell's type classes). Our novel presentation reveals their similarity.
- In Chapter 4 and Chapter 5, we present the key contributions of this thesis. We develop the *flat* and *structural* calculi, show how they capture important contextual properties and develop their categorical semantics using a notion based on comonads. Chapter 6 links the two systems using a single formalism that is capable of capturing both flat and structural properties and also discusses an alternative presentation that is more suitable for automatic type inference of coeffects.
- Related work is presented in Chapter 2 and throughout the thesis. One important direction deserves further exploration, and so Chapter 6 also includes a brief discussion of a different approach to tracking contextual information that arises from modal logics. This is followed by an outline of the future work that is needed in order to implement the presented theory in main-stream programming languages.

If there is a one thing that the reader should remember from this thesis, it is the fact that there is a unified notion of *context*, capturing many common scenarios in programming, and that programming language designers need to provide ways for working with this context (using *coeffects* or not). This greatly reduces the number of distinct concepts that software developers need to keep in mind of when building applications for the rich and diverse execution environments of the future.



There are many different directions from which the concept of *coeffects* can be approached and, indeed, discovered. In the previous chapter, we motivated it by practical applications, but coeffects also naturally arise as an extension to a number of programming language theories. Thanks to the Curry-Howard-Lambek correspondence, we can approach coeffects from the perspective of type theory, logic and also category theory. This chapter gives an overview of the most important directions.

We start by looking at coeffects as the dual of effect systems (Section 2.1). Next, we extend the duality to category theory, looking at the categorical dual of monads known as *comonads* (Section 2.2). Finally we look at logically inspired type systems that are closely related to our structural coeffects (Section 2.3).

This chapter serves two purposes. Firstly, it provides a high-level overview of the related work, although technical details are often postponed until later. Secondly it recasts existing ideas in a way that naturally leads to the coeffect systems developed later in the thesis. For this reason, we are not always faithful to the referenced work – sometimes we focus on aspects that the authors consider unimportant or present the work differently than originally intended. The reason is to fulfil the second goal of the chapter. When we do so, this is explicitly said in the text.

## 2.1 THROUGH TYPE AND EFFECT SYSTEMS

Introduced by Gifford and Lucassen [33, 53], type and effect systems have been designed to track effectful operations performed by computations. Examples include tracking of reading and writing from and to memory locations [86], communication in message-passing systems [44] and atomicity in concurrent applications [30].

Type and effect systems are usually specified as judgements of the form  $\Gamma \vdash e : \tau, \mathbf{r}$ , meaning that the expression  $e$  has a type  $\tau$  in a (free-variable) context  $\Gamma$  and additionally may have effects described by  $\mathbf{r}$ . Effect systems are typically added to a language that already supports effectful operations as a way of increasing the safety – the type and effect system provides stronger guarantees than a plain type system. Filinsky [28] refers to this approach as *descriptive*<sup>1</sup>.

### 2.1.1 Simple effect system.

The structure of a simple effect system<sup>2</sup> is demonstrated in Figure 3. The example shows typing rules for a simply typed lambda calculus with an additional (effectful) operation  $l \leftarrow e$  that writes the value of  $e$  to a mutable location  $l$ . The type of locations ( $\text{ref}_\rho \tau$ ) is annotated with a *memory region*  $\rho$  of the location  $l$ . The effects tracked by the type and effect system over-

<sup>1</sup> In contrast to *prescriptive* effect systems that implement computational effects in a pure language – such as monads in Haskell.

<sup>2</sup> Most work on effect systems uses  $\sigma$  for effect annotations. We use letters  $\mathbf{r}, \mathbf{s}, \mathbf{t}$  and also distinguish effect or coeffect annotations by colour.

$$\begin{array}{c}
\text{(var)} \quad \frac{x:\tau \in \Gamma}{\Gamma \vdash x : \tau, \emptyset} \\
\\
\text{(write)} \quad \frac{\Gamma \vdash e : \tau, \mathbf{r} \quad l : \text{ref}_\rho \quad \tau \in \Gamma}{\Gamma \vdash l \leftarrow e : \text{unit}, \mathbf{r} \cup \{\text{write}(\rho)\}} \\
\\
\text{(abs)} \quad \frac{\Gamma, x:\tau_1 \vdash e : \tau_2, \mathbf{r}}{\Gamma \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{r}} \tau_2, \emptyset} \\
\\
\text{(app)} \quad \frac{\Gamma \vdash e_1 : \tau_1 \xrightarrow{\mathbf{r}} \tau_2, \mathbf{s} \quad \Gamma \vdash e_2 : \tau_1, \mathbf{t}}{\Gamma \vdash e_1 e_2 : \tau_2, \mathbf{r} \cup \mathbf{s} \cup \mathbf{t}}
\end{array}$$

Figure 3: Simple effect system

approximate the actual effects and memory regions provide a convenient way to build such over-approximation. The effects are represented as a set of effectful actions that an expression may perform and the effectful action (*write*) adds a primitive effect  $\text{write}(\rho)$ .

The remaining rules are shared by a majority of effect systems. Variable access (*var*) has no effects, application (*app*) combines the effects of both expressions, together with the latent effects of the function to be applied. Finally, lambda abstraction (*abs*) is a pure computation that turns the *actual* effects of the body into *latent* effects of the created function.

### 2.1.2 Simple coeffect system.

When writing the judgements of coeffect systems, we want to emphasize the fact that coeffect systems talk about *context* rather than *results*. For this reason, we write the judgements in the form  $\Gamma @ \mathbf{r} \vdash e : \tau$ , associating the additional information with the context (left-hand side) of the judgement rather than with the result (right-hand side) as in  $\Gamma \vdash e : \tau, \mathbf{r}$ . This change alone would not be very interesting – we simply used different syntax to write a predicate with four arguments. The more interesting difference is how the lambda abstraction rule looks.

The language in Figure 4 extends simple lambda calculus with resources and with a construct **access**  $e$  that obtains the resource specified by the expression  $e$ . Most of the typing rules correspond to those of effect systems. Variable access (*var*) has no context requirements, application (*app*) combines context requirements of the two sub-expressions and latent context-requirements of the function.

However, the (*abs*) rule is different than the corresponding rule for effect systems – the resource requirements of the body  $\mathbf{r} \cup \mathbf{s}$  are split between the *immediate context-requirements* associated with the current context  $\Gamma @ \mathbf{r}$  and the *latent context-requirements* of the function.

As will be discussed in more detail in Section 3.2.1, this means that a resource can be captured when a function is declared (e.g. when it is constructed on the server-side where database access is available), or when a function is called (e.g. when a function created on server-side requires access to current time-zone, it can use the resource available on the client-side). In other words, resources in this example support both static (lexical) and dynamic scoping.

$$\begin{array}{l}
\text{(var)} \quad \frac{x:\tau \in \Gamma}{\Gamma @ \emptyset \vdash x:\tau} \\
\text{(access)} \quad \frac{\Gamma @ r \vdash e : \text{res}_\rho \tau}{\Gamma @ r \cup \{\text{access}(\rho)\} \vdash \text{access } e : \tau} \\
\text{(abs)} \quad \frac{(\Gamma, x:\tau_1) @ r \cup s \vdash e : \tau_2}{\Gamma @ r \vdash \lambda x. e : \tau_1 \xrightarrow{s} \tau_2} \\
\text{(app)} \quad \frac{\begin{array}{c} \Gamma \vdash e_1 : \tau_1 \xrightarrow{r} \tau_2, s \\ \Gamma \vdash e_2 : \tau_1, t \end{array}}{\Gamma \vdash e_1 e_2 : \tau_2, r \cup s \cup t}
\end{array}$$

Figure 4: Simple coeffect system

## 2.2 THROUGH LANGUAGE SEMANTICS

Another pathway to coeffects leads through the semantics of effectful and context-dependent computations. In a pioneering work, Moggi [55] showed that effects (including partiality, exceptions, non-determinism and I/O) can be modelled using the category theoretic notion of *monad*.

When using monads, we distinguish effect-free values  $\tau$  from programs, or computations  $M\tau$ . The *monad*  $M$  abstracts the *notion of computation* and provides a way of constructing and composing effectful computations:

**Definition 1.** A monad over a category  $\mathcal{C}$  is a triple  $(M, \text{unit}, \text{bind})$  where:

- $M$  is a mapping on objects (types)  $M : \mathcal{C} \rightarrow \mathcal{C}$
- $\text{unit}$  is a mapping  $\alpha \rightarrow M\alpha$
- $\text{bind}$  is a mapping  $(\alpha \rightarrow M\beta) \rightarrow (M\alpha \rightarrow M\beta)$

such that, for all  $f : \alpha \rightarrow M\beta$  and  $g : \beta \rightarrow M\gamma$ :

$$\begin{array}{ll}
\text{bind unit} = \text{id} & \text{(left identity)} \\
\text{bind } f \circ \text{unit} = f & \text{(right identity)} \\
\text{bind (bind } g \circ f) = (\text{bind } f) \circ (\text{bind } g) & \text{(associativity)}
\end{array}$$

Without providing much details, we note that well known examples of monads include the partiality monad ( $M\alpha = \alpha + \perp$ ) also corresponding to the Maybe type in Haskell, list monad ( $M\tau = \mu\gamma.1 + (\tau \times \gamma)$ ) and other. In programming language semantics, monads can be used in two distinct ways.

### 2.2.1 Effectful languages and meta-languages

Moggi uses monads to define two formal systems. In the first formal system, a monad is used to model the *language* itself. This means that the semantics of a language is given in terms of a one specific monad and the semantics can be used to reason about programs in that language. To quote “When reasoning about programs one has only one monad, because the programming language is fixed, and the main aim is to prove properties of programs” [55, p. 5].

In the second formal system, monads are added to the programming language as type constructors, together with additional constructs corresponding to monadic bind and unit. A single program can use multiple monads, but the key benefit is the ability to reason about multiple languages. To quote “When reasoning about programming languages one has different monads,

one for each programming language, and the main aim is to study how they relate to each other” [55, p. 5].

In this thesis, we generally follow the first approach – this means that we work with an existing programming language without needing to add additional constructs corresponding to the primitives of our semantics (the alternative is discussed in Section 6.3). To clarify the difference, the following two sections show a minimal example of both formal systems. We follow Moggi and start with language where judgements have the form  $x:\tau_1 \vdash e:\tau_2$  with exactly one variable<sup>3</sup>.

**LANGUAGE SEMANTICS.** When using monads to provide semantics of a language, we do not need to extend the language in any way – we assume that the language already contains the effectful primitives (such as the assignment operator  $x \leftarrow e$  or other). A judgement of the form  $x:\tau_1 \vdash e:\tau_2$  is interpreted as a morphism  $\tau_1 \rightarrow M\tau_2$ , meaning that any expression is interpreted as an effectful computation. The semantics of variable access and the application of a primitive function  $f$  is interpreted as follows:

$$\begin{aligned} \llbracket x:\tau_1 \vdash x:\tau_1 \rrbracket &= \text{unit}_M \\ \llbracket x:\tau_1 \vdash f e:\tau_3 \rrbracket &= (\text{bind}_M f) \circ \llbracket e \rrbracket \end{aligned}$$

Variable access is an effect-free computation, that returns the value of the variable, wrapped using  $\text{unit}_M$ . In the second rule, we assume that  $e$  is an expression using the variable  $x$  and producing a value of type  $\tau_2$  and that  $f$  is a (primitive) function  $\tau_2 \rightarrow M\tau_3$ . The semantics lifts the function  $f$  using  $\text{bind}_M$  to a function  $M\tau_2 \rightarrow M\tau_3$  which is compatible with the interpretation of the expression  $e$ .

**META-LANGUAGE INTERPRETATION.** When designing meta-language based on monads, we need to extend the lambda calculus with additional type(s) and expressions that correspond to monadic primitives. Assuming  $\alpha$  is a primitive type:

$$\begin{aligned} \tau &:= \alpha \mid \tau_1 \rightarrow \tau_2 \mid M\tau \\ e &:= x \mid f e \mid \text{return}_M e \mid \text{let}_M x \leftarrow e_1 \text{ in } e_2 \end{aligned}$$

The types consist of the primitive type, function type and a type constructor that represents monadic computations. This means that the expressions in the language can create both effect-free values, such as  $\tau$  and computations  $M\tau$ . The additional expression  $\text{return}_M$  is used to create a monadic computation (with no actual effects) from a value and  $\text{let}_M$  is used to sequence effectful computations. In the semantics, monads are not needed to interpret variable access and application, they are only used in the semantics of additional (monadic) constructs:

$$\begin{aligned} \llbracket x:\tau \vdash x:\tau \rrbracket &= \text{id} \\ \llbracket x:\tau_1 \vdash f e:\tau_3 \rrbracket &= f \circ \llbracket e \rrbracket \\ \llbracket x:\tau_1 \vdash \text{return}_M e:M\tau_2 \rrbracket &= \text{unit}_M \circ \llbracket e \rrbracket \\ \llbracket x:\tau_1 \vdash \text{let}_M y \leftarrow e_1 \text{ in } e_2:M\tau_3 \rrbracket &= \text{bind}_M \llbracket e_2 \rrbracket \circ \llbracket e_1 \rrbracket \end{aligned}$$

In this system, the interpretation of variable access becomes a simple identity function and application is just composition. Monadic computations are constructed explicitly using  $\text{return}_M$  (interpreted as  $\text{unit}_M$ ) and they are

<sup>3</sup> This simplifies the examples as we do not need *strong* monad, but that is an orthogonal issue to the distinction between language semantics and meta-language.



also sequenced explicitly using the  $\text{let}_M$  construct. As noted by Moggi, the first formal system can be easily translated to the latter by inserting appropriate monadic constructs.

Moggi regards the meta-language system as more fundamental, because “its models are more general”. Indeed, this is a valid and reasonable perspective. Yet, we follow the first style, precisely because it is *less general*. As discussed in Chapter 1, our aim is to develop concrete context-aware programming languages (together with their type theory and semantics) rather than to build a general framework for reasoning about languages with context-dependent properties.

### 2.2.2 Marriage of effects and monads

The work on effect systems and monads both tackle the same problem – representing and tracking of computational effects. The two lines of research have been joined by Wadler and Thiemann [103]. This requires extending the categorical structure. A monadic computation  $\tau_1 \rightarrow M\tau_2$  means that the computation has *some* effects while the judgement  $x:\tau_1 \vdash e:\tau_2, \mathbf{r}$  specifies *what* effects the computation has.

To solve this mismatch, Wadler and Thiemann use a *family* of monads  $M^{\mathbf{r}}\tau$  with an annotation that specifies the effects that may be performed by the computation. In their system, an effectful function  $\tau_1 \xrightarrow{\mathbf{r}} \tau_2$  is modelled as a pure function returning monadic computation  $\tau_1 \rightarrow M^{\mathbf{r}}\tau_2$ . Similarly, the semantics of a judgement  $x:\tau_1 \vdash e:\tau_2, \mathbf{r}$  can be given as a function  $\tau_1 \rightarrow M^{\mathbf{r}}\tau_2$ . The precise nature of the family of monads has been later called *indexed monads* by Tate [87] and further developed by Atkey [6] in his work on *parameterized monads* and Katsumata [45].

**THESIS PERSPECTIVE.** The key takeaway for this thesis from the outlined line of research is that, if we want to develop a language with type system that captures context-dependent properties of programs more precisely, the semantics of the language also needs to be a more fine-grained structure (akin to indexed monads). While monads have been used to model effects, an existing research links context-dependence with *comonads* – the categorical dual of monads.

### 2.2.3 Context-dependent languages and meta-languages

The theoretical parts of this thesis extend the work of Uustalu and Vene who use comonads to give the semantics of data-flow computations [94] and more generally, notions of *context-dependent computations* [93]. The computations discussed in the latter work include streams, arrays and containers. This is a more diverse set of examples, but they all mostly represent forms of collections. Ahman et al. [4] discuss the relation between comonads and *containers* [3] in more details.

The utility of comonads has been explored by a number of authors before. Brookes and Geva [14] use *computational* comonads for intensional semantics<sup>4</sup>. In functional programming, Kieburtz [48] proposed to use comonads for stream programming, but also handling of I/O and interoperability.

Biermann and de Paiva used comonads to model the necessity modality  $\Box$  in intuitionistic modal  $S_4$  [11], linking programming languages derived

<sup>4</sup> The structure of a computational comonad has been also used by the author of this thesis to abstract evaluation order of monadic computations [68].

from modal logics to comonads. One such language has been reconstructed by Pfenning and Davies [73]. Nanevski et al. extend this work to Contextual Modal Type Theory (CMTT) [58], which again shows the importance of comonads for *context-dependent* computations.

While Uustalu and Vene use comonads to define the *language semantics* (the first style of Moggi), Nanevski, Pfenning and Davies use comonads as part of meta-language, in the form of  $\Box$  modality, to reason about context-dependent computations (the second style of Moggi). Before looking at the details, we use the following definition of comonad:

**Definition 2.** A comonad over a category  $\mathcal{C}$  is a triple  $(C, \text{counit}, \text{cobind})$  where:

- $C$  is a mapping on objects (types)  $C : \mathcal{C} \rightarrow \mathcal{C}$
- $\text{counit}$  is a mapping  $C\alpha \rightarrow \alpha$
- $\text{cobind}$  is a mapping  $(C\alpha \rightarrow \beta) \rightarrow (C\alpha \rightarrow C\beta)$

such that, for all  $f : C\alpha \rightarrow \beta$  and  $g : C\beta \rightarrow \gamma$ :

$$\text{cobind } \text{counit} = \text{id} \quad (\text{left identity})$$

$$\text{counit} \circ \text{cobind } f = f \quad (\text{right identity})$$

$$\text{cobind } (g \circ \text{cobind } f) = (\text{cobind } g) \circ (\text{cobind } f) \quad (\text{associativity})$$

The definition is similar to a monad with “reversed arrows”. Intuitively, the  $\text{counit}$  operation extracts a value  $\alpha$  from a value that carries additional context  $C\alpha$ . The  $\text{cobind}$  operation turns a context-dependent function  $C\alpha \rightarrow \beta$  into a function that takes a value with context, applies the context-dependent function to value(s) in the context and then propagates the context. The next section makes this intuitive definition more concrete. More detailed discussion about comonads can be found in Orchard’s PhD thesis [62].

**LANGUAGE SEMANTICS.** To demonstrate the approach of Uustalu and Vene, we consider the non-empty list comonad  $C\tau = \mu\gamma. \tau + (\tau \times \gamma)$ . A value of the type is either the last element  $\tau$  or an element followed by another non-empty list  $\tau \times \gamma$  (consisting of the head  $\tau$  and the tail  $\gamma$ ). Note that the list must be non-empty, otherwise  $\text{counit}$  would not be a complete function (it would be undefined on empty list). In the following, we write  $(l_1, \dots, l_n)$  for a list of  $n$  elements:

$$\text{counit } (l_1, \dots, l_n) = l_1$$

$$\text{cobind } f (l_1, \dots, l_n) = (f(l_1, \dots, l_n), f(l_2, \dots, l_n), \dots, f(l_n))$$

The  $\text{counit}$  operation returns the current (first) element of the (non-empty) list. The  $\text{cobind}$  operation creates a new list by applying the context-dependent function  $f$  to the entire list, to the suffix of the list, to the suffix of the suffix and so on. Interestingly, it preserves the *shape* of the list as it turns a list of  $n$  elements into another list of  $n$  elements.

In causal data-flow, we can interpret the list as a list consisting of past values, with the current value in the head. Then, the  $\text{cobind}$  operation calculates the current value of the output based on the current and all past values of the input; the second element is calculated based on all past values and the last element is calculated based just on the initial input ( $l_n$ ). In addition to the operations of comonad, the model also uses some operations that are specific to causal data-flow:

$$\text{prev } (l_1, \dots, l_n) = (l_2, \dots, l_n)$$

The operation drops the first element from the list. In the data-flow interpretation, this means that it returns the previous state of a value.

Now, consider a simple data-flow language with single-variable contexts, variables, primitive built-in functions and a construct `prev e` that returns the previous value of the computation  $e$ . We omit the typing rules, but they are simple – assuming  $e$  has a type  $\tau$ , the expression `prev e` has also type  $\tau$ . The fact that the language models data-flow and values are lists (of past values) is a matter of semantics, which is defined as follows:

$$\begin{aligned} \llbracket x:\tau \vdash x:\tau \rrbracket &= \text{counit}_C \\ \llbracket x:\tau_1 \vdash f e:\tau_3 \rrbracket &= f \circ (\text{cobind}_C \llbracket e \rrbracket) \\ \llbracket x:\tau_1 \vdash \text{prev } e:\tau_2 \rrbracket &= \text{prev} \circ (\text{cobind}_C \llbracket e \rrbracket) \end{aligned}$$

The semantics follows that of effectful computations using monads. A variable access is interpreted using  $\text{counit}_C$  (obtain the value and ignore additional available context); composition uses  $\text{cobind}_C$  to propagate the context to the function  $f$  and `prev` is interpreted using the primitive `prev` (which takes a list and returns a list).

For example, the judgement  $x:\tau \vdash \text{prev} (\text{prev } x):\tau$  represents an expression that expects context with variable  $x$  and returns a stream of values before the previous one. The semantics of the term expresses this behaviour:  $(\text{prev} \circ \text{prev} \circ (\text{cobind}_C \text{counit}_C))$ . Note that the first operation is simply an identity function thanks to the comonad laws discussed earlier.

In the outline presented here, we ignored lambda abstraction. Similarly to monadic semantics, where lambda abstraction requires *strong* monad, the comonadic semantics also requires additional structure called *symmetric (semi)monoidal* comonads. This structure is responsible for the splitting of context-requirements in lambda abstraction. Note that this is what happens in the unusual (*abs*) rule in Figure 4, which distinguishes coeffect systems from effect systems.

We return to this topic when discussing lambda abstraction in Section 3.1.1 and semantics of flat coeffect systems in Section 4.3.

**META-LANGUAGE INTERPRETATION.** To briefly demonstrate the approach that employs comonads as part of a meta-language, we look at an example inspired by the work of Pfenning, Davies and Nanevski. We do not attempt to provide a precise overview of their work. The main purpose of the following discussion is to provide a different intuition behind comonads, and to present an example of a language that includes comonad as a type constructor, together with language primitives corresponding to comonadic operations<sup>5</sup>.

In languages inspired by modal logics, types can have the form  $\Box\tau$ . In the work of Pfenning and Davies, this is the type of a term that is provable with no assumptions. In distributed programming language ML5 by Murphy et al. [56, 57], the  $\Box\tau$  type means *mobile code*, that is code that can be evaluated at any node of a distributed system (the evaluation corresponds to the axiom  $\Box\tau \rightarrow \tau$ ). Finally, Davies and Pfenning [24] consider staged computations and interpret  $\Box\tau$  as a type of unevaluated expressions of type  $\tau$  (with no free variables).

In Contextual Modal Type Theory, the modality  $\Box$  is further annotated with the free variables of the (unevaluated) expression. We write  $\Box^\Psi\tau$  for a type of expressions that requires a context  $\Psi$ . The type is a comonadic counterpart to *indexed monads* used by Wadler and Thiemann when linking

<sup>5</sup> In fact, Pfenning, Davies and Nanevski [73, 58] never mention comonads explicitly. This is done in later work by Gabbay et al. [32], but the connection between the language and comonads is not as direct as in case of monadic or comonadic semantics covered in the previous section.

$$\begin{array}{c}
\text{(eval)} \frac{\Gamma \vdash e : \Box^{\emptyset} \tau}{\Gamma \vdash !e : \tau} \\
\\
\text{(letbox)} \frac{\Gamma \vdash e_1 : \Box^{\Phi, \Psi} \tau_1 \quad \Gamma, x : \Box^{\Phi} \tau_1 \vdash e_2 : \tau_2}{\Gamma \vdash \text{let box } x = e_1 \text{ in } e_2 : \Box^{\Psi} \tau_2}
\end{array}$$

Figure 5: Typing for a comonadic language with contextual staged computations

monads and effect systems and, indeed, it gives rise to a language that tracks context-dependence of computations in a type system.

In staged computation, the type  $C^{\Psi} \tau$  represents an expression that requires the context  $\Psi$  (i.e. the expression is an open term that requires variables  $\Psi$ ). The Figure 5 shows two typing rules for such language. The rules directly correspond to the two operations of a comonad and can be interpreted as follows:

- *(eval)* corresponds to  $\text{counit} : C^{\emptyset} \alpha \rightarrow \alpha$ . It indicates that we can evaluate a closed (unevaluated) term and obtain a value. Interestingly, the rule requires a specific context annotation (empty set of free variables). It is not possible to evaluate an open term.
- *(letbox)* corresponds to  $\text{cobind} : (C^{\Psi} \alpha \rightarrow \beta) \rightarrow C^{\Psi, \Phi} \alpha \rightarrow C^{\Phi} \beta$ . Given a term which requires variable context  $\Psi$ ,  $\Phi$  (expression  $e_1$ ) and a function that turns a term needing  $\Psi$  into an evaluated value (expression  $e_2$ ), we can construct a term that requires just  $\Phi$ .

The fact that the *(eval)* rule requires a specific context is an interesting relaxation from ordinary comonads where  $\text{counit}$  needs to be defined for all values. Here, the indexed  $\text{counit}$  operation needs to be defined *only* on values annotated with  $\emptyset$ .

The annotated  $\text{cobind}$  operation that corresponds to *(letbox)*. An interesting aspect is that it propagates the context-requirements “backwards”. The input expression (second parameter) requires a combination of contexts that are required by the two components – those required by the input of the function (first argument) and those required by the resulting expression (result). This is another key aspect that distinguishes coeffects from effect systems. We return back to the meta-language approach of embedding comonads in Section 6.3.

**THESIS PERSPECTIVE.** As mentioned earlier, we are interested in designing context-dependent languages and so we use comonads for *language semantics*. Uustalu and Vene present a semantics of context-dependent computations in terms of comonads. We provide the rest of the story known from the marriage of monads and effects. We develop coeffect calculus with an type system that tracks the context requirements more precisely (by annotating the types) and we add indexing to comonads and link the two by giving a formal semantics. The indexing allows us to capture applications that do not fit into the model provided by plain comonads.

The *meta-language* approach of Pfenning, Davies and Nanevski is closely related to our work. Most importantly, Contextual Modal Type Theory (CMTT) uses indexed  $\Box$  modality which corresponds to indexed comonads (in a similar way in which effect systems correspond to indexed monads). The relation between CMTT and comonads has been suggested by Gabbay et al. [32], but the meta-language employed by CMTT does not directly correspond to

comonadic operations. For example, our (*letbox*) typing rule from Figure 5 is not a primitive of CMTT and would correspond to  $\text{box}(\Psi, \text{letbox}(e_1, x, e_2))$ . Nevertheless, the indexing in CMTT provides a useful hint for adding indexing to the work of Uustalu and Vene.

### 2.3 THROUGH SUBSTRUCTURAL AND BUNCHED LOGICS

In the coeffect system for tracking resource usage outlined earlier, we associated additional contextual information (set of available resources) with the variable context of the typing judgement:  $\Gamma @ r \vdash e : \tau$ . In other words, our work focuses on “what is happening on the left hand side of  $\vdash$ ”.

In the case of resources, the additional information about the context are simply added to the variable context (as a products), but we will later look at contextual properties that affect how variables are represented. More importantly, *structural coeffects* link additional information to individual variables in the context, rather than the context as a whole.

In this section, we look at type systems that reconsider  $\Gamma$  in a number of ways. First of all, substructural type systems [104] restrict the use of variables in the language. Most famously linear type systems introduced by Wadler [101] can guarantee that a variable is used exactly once. This has interesting implications for memory management and I/O.

In bunched typing developed by O’Hearn [60], the variable context is a tree formed by multiple different constructors (e.g. one that allows sharing and one that does not). Most famously, bunched typing has contributed to the development of separation logic [61] (starting a fruitful line of research in software verification), but it is also interesting on its own.

#### 2.3.1 Substructural type systems.

Traditionally,  $\Gamma$  is viewed as a set of assumptions and typing rules admit (or explicitly include) three transformations that manipulate the variable contexts which are shown in Figure 6. The (*exchange*) rule allows reordering of variables (which is implicit when assumptions are treated as set); (*weakening*) makes it possible to discard an assumption – this has the implication that a variable may be declared but never used. Finally, (*contraction*) makes it possible to use a single variable multiple times (in the rule, this is done explicitly by joining multiple variables into a single one using substitution).

In substructural type systems, the assumptions are typically treated as a list. As a result, they have to be manipulated explicitly. Different systems allow different subsets of the rules. For example, *affine* systems allows exchange and weakening, leading to a system where variable may be used at most once; in *linear* systems, only exchange is permitted and so every variable has to be used exactly once.

When tracking context-dependent properties associated with individual variables, we need to be more explicit in how variables are used. Substructural type systems provide a way to do this. Even if we allow all three operations, we can use a variation on the three rules (exchange, weakening and contraction) to track which variables are used and how (and to track additional contextual information about variables).

$$\begin{array}{c}
\text{(exchange)} \quad \frac{\Gamma, x:\tau_1, y:\tau_2 \vdash e : \gamma}{\Gamma, y:\tau_2, x:\tau_1 \vdash e : \gamma} \\
\\
\text{(weakening)} \quad \frac{\Gamma, \Delta \vdash e : \gamma}{\Gamma, x:\tau, \Delta \vdash e : \gamma} \\
\\
\text{(contraction)} \quad \frac{\Gamma, x:\tau_1, y:\tau_1, \Delta \vdash e : \tau_2}{\Gamma, x:\tau_1, \Delta \vdash e[y \leftarrow x] : \tau_2}
\end{array}$$

Figure 6: Exchange, weakening and contraction typing rules

### 2.3.2 Bunched type systems.

Bunched typing makes one more refinement to how  $\Gamma$  is treated. Rather than having a list of assumptions, the context becomes a tree that contains variable typings (or special identity values) in the leaves and has multiple different types of nodes. The context can be defined, for example, as follows:

$$\Gamma, \Delta, \Sigma := x:\alpha \mid I \mid \Gamma, \Gamma \mid 1 \mid \Gamma; \Gamma$$

The values  $I$  and  $1$  represent two kinds of “empty” contexts. More interestingly, non-empty variable contexts may be constructed using two distinct constructors –  $\Gamma, \Gamma$  and  $\Gamma; \Gamma$  – that have different properties. In particular, weakening and contraction is only allowed for the  $;$  constructor, while exchange is allowed for both.

The structural rules for bunched typing are shown in Figure 7. The syntax  $\Gamma(\Delta)$  is used to mean an assumption tree that contains  $\Delta$  as a sub-tree and so, for example, *(exchange1)* can switch the order of contexts anywhere in the tree. The remaining rules are similar to the rules of linear logic.

One important note about bunched typing is that it requires a different interpretation. The omission of weakening and contraction in linear logic means that variable can be used exactly once. In bunched typing, variables may still be duplicated, but only using the “ $;$ ” separator. The type system can be interpreted as specifying whether a variable may be shared between the body of a function and the context where a function is declared.

The system introduces two distinct function types  $\tau_1 \rightarrow \tau_2$  and  $\tau_1 * \tau_2$  (corresponding to “ $\rightarrow$ ” and “ $*$ ” respectively). The key property is that only the first kind of functions can share variables with the context where a function is declared, while the second restricts such sharing. We do not attempt to give a detailed description here as it is not immediately related to coeffects – for more information, refer to O’Hearn’s introduction [60].

**THESIS PERSPECTIVE.** From the perspective of substructural and bunched types, our work can be viewed as annotating bunches. Such annotations then specify additional information about the context – or, more specifically, about the sub-tree of the context. Although this is not the exact definition used in Chapter 6, we could define contexts as follows:

$$\Gamma, \Delta, \Sigma := x:\alpha \mid 1 \mid \Gamma, \Gamma \mid \Gamma @ r$$

Now we can not only annotate an entire context with some information (as in the simple coeffect system for tracking resources that used judgements of a form  $\Gamma r \vdash e : \tau$ ). We can also annotate individual components. For example, a context containing variables  $x, y, z$  where only  $x$  is used could be written as  $(x:\tau_1 @ \text{used}, (y:\tau_2, z:\tau_3) @ \text{unused})$ .



$$\begin{array}{c}
\text{(exchange1)} \quad \frac{\Gamma(\Delta, \Sigma) \vdash e : \alpha}{\Gamma(\Sigma, \Delta) \vdash e : \alpha} \\
\\
\text{(exchange2)} \quad \frac{\Gamma(\Delta; \Sigma) \vdash e : \alpha}{\Gamma(\Sigma; \Delta) \vdash e : \alpha} \\
\\
\text{(weakening)} \quad \frac{\Gamma(\Delta) \vdash e : \alpha}{\Gamma(\Delta; \Sigma) \vdash e : \alpha} \\
\\
\text{(contraction)} \quad \frac{\Gamma(\Delta; \Sigma) \vdash e : \alpha}{\Gamma(\Delta) \vdash e[\Sigma \leftarrow \Delta] : \alpha}
\end{array}$$


---

Figure 7: Exchange, weakening and contraction rules for bunched typing

For the purpose of this introduction, we ignore important aspects such as how are nested annotations interpreted. The main goal is to show that coeffects can be easily viewed as an extension to the work on bunched logic. Aside from this principal connection, *structural coeffects* also use some of the proof techniques from the work on bunched logics.

## 2.4 CONTEXT ORIENTED PROGRAMMING

The importance of context-aware computations is perhaps most obvious when considering mobile application, client/server web applications or even the internet of things. A pioneering work in the area using functional languages has been done by Serrano [77, 52] (which also inspired the motivating example presented in Chapter 1). His HOP language supports cross-compilation and programs execute in different contexts. However, HOP is not statically type checked.

In the software engineering community, a number of authors have addressed the problem of context-aware computations. Hirschfeld et al. propose *Context-Oriented Programming* (COP) as a methodology [42]. The COP paradigm has been later implemented by programming language features. Costanza [22] develops a domain-specific LISP-like language ContextL and Bardram [7] proposes a Java framework for COP.

Finally, the subject of context-awareness has also been addressed in work focusing on the development of mobile applications [9, 26]. Here, the *context* focuses more on concrete physical context (obtained from the device sensors) than context as an abstract language feature.

We approach the problem from a different perspective, building on the tradition of statically-typed functional programming languages, focusing on type systems as the primary way of capturing contextual properties.

## 2.5 SUMMARY

This chapter presented three different pathways leading to the idea of coeffects. We also introduced the most important related work, although presenting related work was not the primary goal of the chapter. The primary goal was to present the idea of coeffects as a logical follow up to a number of research directions. For this reason, we highlighted only certain aspects of the discussed related work – the remaining aspects as well as important technical details are covered throughout the thesis.

The first pathway follows as a dualization of well-known effect systems. However, this is not simply a syntactic transformation. As we further discuss in the next chapter, coeffect systems treat lambda abstraction differently. The second pathway follows by extending comonadic semantics of context-dependent computations with indexing and building a type system analogous to effect system from the “marriage of effects and monads”. Finally, the third pathway starts with substructural type systems. Coeffect systems naturally arise by annotating bunches in bunched logics with additional information. In this thesis, we mostly follow the first two approaches.



Software developers as well as programming language researchers choose abstractions based not just on how appropriate they are. Other factors include social aspects – how well is the abstraction known, how well is it documented and whether it is a standard tool of the *research programme*<sup>1</sup> that the researcher unconsciously subscribes to.

For tracking of effects, such *standard tools* are well known. When faced with an effectful computation, programming language designers immediately pick monads. For context-aware computations, there are no standard tools. Thus contextual properties may, at first, appear as a set of disconnected examples. Existing systems that capture contextual properties use a wide range of methods including special-purpose type systems, approaches arising from modal logic  $S_4$ , as well as techniques based on abstractions designed for other purpose, most frequently monads.

#### CHAPTER STRUCTURE AND CONTRIBUTIONS

- We start with a characterization of contextual properties. The Section 3.1 explains what is a *coeffect* and contrasts it with a better known notion of *effect*. It explains what is the nature of properties that can be tracked using coeffect systems presented in this theis.
- We describe a number of simple calculi for tracking a wide range of contextual properties. The systems are adapted from diverse sources (type systems, static analyses, logics) and apply to various domains (cross-compilation, liveness, distributed computing, data-flow, security), but share a common structure.
- The uniform presentation of the systems is the key contribution of this chapter. We distinguish between *flat coeffect* systems (Section 3.2) and *structural coeffect* systems (Section 3.3). The fact that we find a common structure in all systems presented here lets us develop unified coeffect calculi in the upcoming three chapters.
- In addition, the coeffect systems for tracking the number of accessed past values in data-flow languages (Sections 3.2.4 and 3.3.3) presents novel results and can be used to optimize data-flow programs.

As mentioned, this chapter may appear as a collection of disconnected examples<sup>2</sup>. But at the end, we will see that they share a common pattern.

#### 3.1 STRUCTURE OF COEFFECT SYSTEMS

When introducing coeffect systems in Section 2.1.2, we related coeffect systems with effect systems. Effect systems track how a program affects the environment, or, in other words capture some *output impurity*. In contrast, coeffect systems track what a program requires from the execution environment, or *input impurity*.

<sup>1</sup> A research programme, as introduced by Lakatos [50], is a network of scientists sharing the same basic assumptions and techniques.

<sup>2</sup> The different properties captured by monads may appear similarly disconnected at first!

$$\begin{array}{c}
\text{(pure)} \quad \frac{\Gamma, x:\tau_1 \vdash e : \tau_2}{\Gamma \vdash \lambda x.e : \tau_1 \rightarrow \tau_2} \\
\\
\text{(effect)} \quad \frac{\Gamma, x:\tau_1 \vdash e : \tau_2 \ \& \ \sigma}{\Gamma \vdash \lambda x.e : \tau_1 \xrightarrow{\sigma} \tau_2 \ \& \ \emptyset}
\end{array}$$

---

Figure 8: Lambda abstraction for pure and effectful computations

Effect systems generally use judgements of the form  $\Gamma \vdash e : \tau \ \& \ \sigma$ , associating effects  $\sigma$  with the output type. We write coeffect systems using judgements of the form  $\Gamma @ \sigma \vdash e : \tau$ , associating the context requirements with  $\Gamma$ . Thus, we extend the traditional notion of free-variable context  $\Gamma$  with richer notions of context. This notation emphasizes the right intuition, but there are more important differences between effects and coeffects.

### 3.1.1 Effectful lambda abstraction

The difference between effects and coeffects becomes apparent when we consider lambda abstraction. The typical lambda abstraction rule for effect systems looks as *(effect)* in Figure 8. Wadler and Thiemann [103] explain how the effect analysis works as follows:

*In the rule for abstraction, the effect is empty because evaluation immediately returns the function, with no side effects. The effect on the function arrow is the same as the effect for the function body, because applying the function will have the same side effects as evaluating the body.*

This is the key property of *output impurity*. The effects are only produced when the function is evaluated and so the effects of the body are attached to the function. A recent work by Tate [87] uses the term *producer* effect systems for such standard systems and characterises them as follows:

*Indeed, we will define an effect as a producer effect if all computations with that effect can be thunked as “pure” computations for a domain-specific notion of purity.*

The thunking is typically performed by a lambda abstraction – given an effectful expression  $e$ , the function  $\lambda x.e$  is an effect free value (thunk) that delays all effects. As shown in the next section, contextual properties do not follow this pattern.

### 3.1.2 Notions of context

We look at three notions of context. The first is the standard free-variable context in  $\lambda$ -calculus. This is well understood and we use it to demonstrate how contextual properties behave. Then we consider two notions of context introduced in this thesis – *flat coeffects* refer to overall properties of the environment and *structural coeffects* refer to properties attached to individual variables. We could track properties associated with values in data structures (e. g. fields of a tuple), but this is left as future work.

**VARIABLE COEFFECTS.** In standard  $\lambda$ -calculus, variable access can be seen as a primitive operation that accesses the context. The variable access expression introduces a context requirement – the expression  $x$  is typeable only in a context that contains  $x : \tau$  for some type  $\tau$ .

The standard lambda abstraction (*pure*), shown in Figure 8, splits the free-variable context of an expression into two parts. At runtime, the value of the parameter has to be provided by the *call site* (dynamic scope) and the remaining values are provided by the *declaration site* (lexical scope). In the type checking, the splitting is determined syntactically – the notation  $\lambda x.e$  names the variable whose value comes from the call site.

Flat and structural coeffects also split context-requirements between the declaration site and the call site. The flat and structural coeffects capture two different ways of doing this.

**FLAT COEFFECTS.** In Section 1.2.1, we used *resources* in a distributed system as an example of flat coeffects. These could be, for example, a database, GPS sensor or access to the current time. We also outlined that such context requirements can be tracked as part of the typing assumption, for example, say we have an expression  $e$  that requires GPS coordinates and the current time. The variable context of such expression will be annotated with a set of required resources, i. e.  $\Gamma @ \{ \text{gps}, \text{time} \}$ .

The interesting case is when we construct a lambda function  $\lambda x.e$ , marshal it and send it to another node. In systems such as Acute [78], the context requirements can be satisfied in a number of ways. When the same resource is available at the target machine (e. g. current time), we can transfer the function with a context requirement and *rebind* the resource. However, if the resource is not available (e. g.. GPS on the server), we need to a capture *remote reference*.

In the example discussed here,  $\lambda x.e$  would require GPS sensor from the declaration site (lexical scope) where the function is declared, which is attached to the current context as  $\Gamma @ \{ \text{gps} \}$ . The current time is required from the caller of the function. So, the context requirement on the call site (dynamic scope) will be  $r = \{ \text{time} \}$ . In coeffect systems, we attach this information to the function, writing  $\tau_1 \xrightarrow{r} \tau_2$ .

We look at resources in distributed programming in more details in Section 3.2.2. The important point here is that in flat coeffect systems, contextual requirements are *split* between the call site and declaration site. Furthermore, there is no syntactic structure that determines how the requirements are split. In the case of distributed programming, the resources can be freely associated with either of the two sites.

**STRUCTURAL COEFFECTS.** On the one hand, variable context provides a *fine-grained tracking* mechanism of how context (variables) are used. On the other hand, flat coeffects let us track *additional information* about the context. The purpose of *structural coeffects* is to reconcile the two and to provide a way for fine-grained tracking of additional information linked to variables in programs. Structural coeffects follow the lexical scoping structure determined by the typing rules.

In Section 1.1.4, we used an example of tracking array access patterns. For every variable, the additional coeffect annotation keeps a range of indices that may be accessed relatively to the current cursor. For example, consider an expression  $x[\text{cursor}] = y[\text{cursor} - 1] + y[\text{cursor} + 1]$ .

Here, the variable context  $\Gamma$  contains two variables, both of type *Arr*. This means  $\Gamma = x:\text{Arr}, y:\text{Arr}$ . For simplicity, we treat *cursor* as a language primitive. The coeffect annotations will be  $(0, 0)$  for  $x$  and  $(-1, 1)$  for  $y$ , denoting that we access only the current value in  $x$ , but we need access to both left and right neighbours in the  $y$  array. In order to unify the flat and structural

notions, we attach this information as a *vector* of annotations associated with a *vector* of variable and write:  $x : \text{Arr}, y : \text{Arr} @ \langle (0, 0), (-1, 1) \rangle$ . The unification is discussed in Chapter 6.

In structural systems, the splitting of context is determined by the name (variable) binding. For example, consider a function that takes  $y$  and contains the above body:  $\lambda y. x[\text{cursor}] = y[\text{cursor} - 1] + y[\text{cursor} + 1]$ . Here, the declaration site contains  $x$  and needs to provide access at least within a range  $(0, 0)$ . The call site provides a value for  $y$ , which needs to be accessible at least within  $(-1, 1)$ . In this way, structural coeffects remove the non-determinism arising from the splitting of requirements in flat coeffect systems.

Before looking at concrete flat and structural systems, we briefly overview some notation used in this thesis. Structural coeffects keep annotations as *vectors* and use a number of operations related to scalars and vectors.

### 3.1.3 Scalars and vectors

The  $\lambda$ -calculus is asymmetric. It maps a context with *multiple* variables to a *single* result. An expression with  $n$  free variables of types  $\tau_i$  can be modelled by a function  $\tau_1 \times \dots \times \tau_n \rightarrow \tau$  with a product on the left, but a single value on the right. In both effect systems and coeffect systems, we *write* the annotation as part of the function arrow. However, in the underlying categorical model, effects are attached to the result  $\tau$ , while coeffects are attached to the context  $\tau_1 \times \dots \times \tau_n$ .

Structural coeffects have one coeffect annotation per each variable. Thus, the annotation consists of multiple values – one belonging to each variable. To distinguish between the overall annotation and individual (per-variable) annotations, we call the overall coeffect a *vector* consisting of *scalar* coeffects. This asymmetry also explains why coeffect systems are not trivially dual to effect systems.

It is useful to clarify how vectors are used in this thesis. Suppose we have a set  $\mathcal{C}$  of *scalars* ranged over by  $r, s, t$ . A vector  $\mathbf{R}$  over  $\mathcal{C}$  is a tuple  $\langle r_1, \dots, r_n \rangle$  of scalars. We use bold face letters like  $\mathbf{r}, \mathbf{s}, \mathbf{t}$  for vectors and normal face  $r, s, t$  for scalars<sup>3</sup>. We also say that a *shape* of a vector  $\text{len}(\mathbf{r})$  (or more generally any container) determines the set of *positions* in a vector. So, a vector of a shape (length)  $n$  has positions  $\{1, 2, \dots, n\}$ . We discuss containers and shapes further in Chapter 6 and also discuss how our use relates to containers of Abbott, Altenkirch and Ghani [3].

Just as in the usual pointwise multiplication of a vector by a scalar, we lift any binary operation on scalars into a scalar-vector one. For a binary operation on scalars  $\circ : \mathcal{C} \times \mathcal{C} \rightarrow \mathcal{C}$ , we define  $\mathbf{s} \circ \mathbf{r} = \langle s \circ r_1, \dots, s \circ r_n \rangle$ . Relations on scalars can be also lifted to vectors. Given two vectors  $\mathbf{r}, \mathbf{s}$  of the same shape with positions  $\{1, \dots, n\}$  and a relation  $\alpha \subseteq \mathcal{C} \times \mathcal{C}$  we define  $\mathbf{r} \alpha \mathbf{s} \Leftrightarrow (r_1 \alpha s_1) \wedge \dots \wedge (r_n \alpha s_n)$ . Finally, we often concatenate vectors – for example, when joining two variable contexts. Given vectors  $\mathbf{r}, \mathbf{s}$  with (possibly different) shapes  $\{1, \dots, n\}$  and  $\{1, \dots, m\}$ , the associative operation for concatenation  $\times$  is defined as  $\mathbf{r} \times \mathbf{s} = \langle r_1, \dots, r_n, s_1, \dots, s_m \rangle$ .

We note that an environment  $\Gamma$  containing  $n$  uniquely named, typed variables is also a vector, but we continue to write ‘,’ for the product, so  $\Gamma_1, x : \tau, \Gamma_2$  should be seen as  $\Gamma_1 \times \langle x : \tau \rangle \times \Gamma_2$ .

<sup>3</sup> For better readability, the thesis also distinguishes different structures using colours. However ignoring the colour does not introduce any ambiguity.

### 3.2 FLAT COEFFECT SYSTEMS

In flat coeffect systems, the additional contextual information are independent of lexically scoped variables. As such, flat coeffects capture properties where the execution environment provides some additional data, resources or information about the execution context.

As mentioned in the introduction, coeffect systems in this chapter may appear as a disconnected set of examples at first. Indeed, this section covers a diverse set of calculi including Haskell’s implicit parameters (Section 3.2.1), distributed computing and cross-compilation (Section 3.2.2), liveness analysis (Section 3.2.3) and data-flow (Section 3.2.4).

For three of the examples, we present a type system and a simple semantics. Although the examples are not new, our novel presentation of the systems (and the fact that they appear side-by-side) makes it possible to see that they share a common structure. The structure is captured by a unified *flat coeffect calculus* in Chapter 4.

#### 3.2.1 Implicit parameters and type classes

Haskell provides two examples of flat coeffects – type class constraints and implicit parameter constraints [102, 51]. Both of the features introduce additional *constraints* on the context requiring that the environment provides certain operations for a type (type classes) or that it provides values for named implicit parameters. In the Haskell type system, constraints  $C$  are attached to the types of top-level declarations, such as let-bound functions. The Haskell notation  $\Gamma \vdash e : C \Rightarrow \tau$  corresponds to our notation  $\Gamma @ C \vdash e : \tau$ .

In this section, we present a type system for implicit parameters in terms of the coeffect typing judgement. We briefly consider type classes, but do not give a full type system.

**IMPLICIT PARAMETERS.** Implicit parameters are a special kind of variables that support dynamic scoping. They make it possible to parameterise a computation (involving a long chain of function calls) without passing parameters explicitly as additional arguments of all involved functions.

The dynamic scoping means that if a function uses a parameter  $?param$  then the caller of the function must set a value of  $?param$  before calling the function. However, implicit parameters also support lexical scoping. If the parameter  $?param$  is available in the lexical scope where a function is defined, then the function will not require a value from the caller.

A simple language with implicit parameters has an expression  $?param$  to read a parameter and an expression<sup>4</sup> `letdyn ?param =  $e_1$  in  $e_2$`  that sets a parameter  $?param$  to the value of  $e_1$  and evaluates  $e_2$  in a context containing  $?param$ .

The fact that implicit parameters support both lexical and dynamic scoping becomes interesting when we consider nested functions. The following function does some pre-processing and then returns a function that builds a formatted string based on two implicit parameters  $?width$  and  $?size$ :

```
let format = λstr →
  let lines = formatLines str ?width in
  (λrest → append lines rest ?width ?size)
```

<sup>4</sup> Haskell uses `let ?p =  $e_1$  in  $e_2$` , but we use a different keyword to avoid confusion.

$$\begin{array}{l}
\text{(var)} \quad \frac{x : \tau \in \Gamma}{\Gamma @ \emptyset \vdash x : \tau} \\
\text{(param)} \quad \frac{}{\Gamma @ \{\text{?param} : \tau\} \vdash \text{?param} : \tau} \\
\text{(sub)} \quad \frac{\Gamma @ \mathbf{r}' \vdash e : \tau}{\Gamma @ \mathbf{r} \vdash e : \tau} \quad (\mathbf{r}' \subseteq \mathbf{r}) \\
\text{(app)} \quad \frac{\Gamma @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \quad \Gamma @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma @ \mathbf{r} \cup \mathbf{s} \cup t \vdash e_1 e_2 : \tau_2} \\
\text{(let)} \quad \frac{\Gamma @ \mathbf{r} \vdash e_1 : \tau_1 \quad \Gamma, x : \tau_1 @ \mathbf{s} \vdash e_2 : \tau_2}{\Gamma @ \mathbf{r} \cup \mathbf{s} \vdash \text{let } x = e_1 \text{ in } e_2 : \tau_2} \\
\text{(abs)} \quad \frac{\Gamma, x : \tau_1 @ \mathbf{r} \cup \mathbf{s} \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{s} \tau_2} \\
\text{(letdyn)} \quad \frac{\Gamma @ \mathbf{r} \vdash e_1 : \tau_1 \quad \Gamma @ \mathbf{s} \vdash e_2 : \tau_2}{\Gamma @ \mathbf{r} \cup (\mathbf{s} \setminus \{\text{?p} : \tau_1\}) \vdash \text{letdyn } \text{?p} = e_1 \text{ in } e_2 : \tau_2}
\end{array}$$

Figure 9: Coeffect rules for tracking implicit parameters

The body of the outer function accesses the parameter `?width`, so it certainly requires a context  $\{\text{?width} : \text{int}\}$ . The nested function (returned as a result) uses the parameter `?width`, but in addition also uses `?size`. Where should the parameters used by the nested function come from?

To keep examples in this chapter uniform, we do not use the Haskell notation and instead write  $\tau_1 \xrightarrow{r} \tau_2$  for a function that requires implicit parameters specified by  $r$ . In a purely dynamically scoped system, they would have to be defined when the user invokes the nested function. However, implicit parameters behave as a combination of lexical and dynamic scoping. This means that the nested function can capture the value of `?width` and require just `?size`. The following shows the two options:

$$\begin{array}{ll}
\text{string} \xrightarrow{\{\text{?width} : \text{int}\}} (\text{string} \xrightarrow{\{\text{?width} : \text{int}, \text{?size} : \text{int}\}} \text{string}) & (\text{dynamic}) \\
\text{string} \xrightarrow{\{\text{?width} : \text{int}\}} (\text{string} \xrightarrow{\{\text{?size} : \text{int}\}} \text{string}) & (\text{mixed})
\end{array}$$

This is not a complete list of possible typings, but it demonstrates the options. The *(dynamic)* case requires the parameter `?width` twice (this may be confusing when the caller provides two different values). In the *(mixed)* case, the nested function captures the `?width` parameter available from the declaration site. Using the latter typing, the function can be called as follows:

```

let formatHello = ( letdyn ?width = 5 in format "Hello")
in ( letdyn ?size = 10 in formatHello "world" )

```

For different typings of `format`, different ways of calling it are valid. This illustrates the point made in Section 3.1.1 – flat coeffect systems may introduce certain non-determinism in the typing. The following section shows how this looks in the type system for implicit parameters.

**TYPE SYSTEM.** Figure 9 shows a type system that tracks the set of expression's implicit parameters. The type system uses judgements of the form  $\Gamma @ \mathbf{r} \vdash e : \tau$  meaning that an expression  $e$  has a type  $\tau$  in a free-variable context  $\Gamma$  with a set of implicit parameters specified by  $\mathbf{r}$ . The annotations

$r, s, t$  are finite partial functions mapping implicit parameter names to types, i. e.  $r, s, t \subseteq \text{Names} \mapsto \text{Types}$ . The expressions include `?param` to read implicit parameter and `letdyn` to bind an implicit parameter. The types are standard, but functions are annotated with the set of implicit parameters that must be available on the call site, i. e.  $\tau_1 \xrightarrow{s} \tau_2$ .

Accessing an ordinary variable (*var*) does not require any implicit parameters. The rule that introduces primitive context requirements is (*param*). Accessing a parameter `?param` of type  $\tau$  requires it to be available in the context. The context may provide more (unused) implicit parameters thanks to the (*sub*) rule.

When we read the rules from the top to the bottom, application (*app*) and let binding (*let*) simply union the context requirements of the sub-expressions. However, lambda abstraction (*abs*) is where the example differs from effect systems. The implicit parameters required by the body  $r \cup s$  can be freely split between the declaration site ( $\Gamma @ r$ ) and the call site ( $\tau_1 \xrightarrow{s} \tau_2$ ). Finally, (*letdyn*) defines an implicit parameter and removes it from the set of requirements.

The union operation  $\cup$  is not a disjoint union, which means that the values for implicit parameters can also be provided by both sites. For example, consider a function with a body  $?a + ?b$ . Assuming that the function takes and returns `int`, the following list shows 4 out of 9 possible valid typing. Full typing derivations can be found in Appendix A.1:

$$\Gamma @ \{?a : \text{int}\} \vdash \lambda x. ?a + ?b : \text{int} \xrightarrow{\{?b : \text{int}\}} \text{int} \quad (1)$$

$$\Gamma @ \{?b : \text{int}\} \vdash \lambda x. ?a + ?b : \text{int} \xrightarrow{\{?a : \text{int}\}} \text{int} \quad (2)$$

$$\Gamma @ \{?a : \text{int}\} \vdash \lambda x. ?a + ?b : \text{int} \xrightarrow{\{?a : \text{int}, ?b : \text{int}\}} \text{int} \quad (3)$$

$$\Gamma @ \emptyset \vdash \lambda x. ?a + ?b : \text{int} \xrightarrow{\{?a : \text{int}, ?b : \text{int}\}} \text{int} \quad (4)$$

The first two examples demonstrate that the system does not have the principal typing property. Both (1) and (2) are valid typings and they may both be desirable in certain contexts where the function is used.

The next typing derivation (3) requires the parameter `?a` from both the declaration site and the call site. This means that, at runtime, two values will be available. Our semantics for the system describes *dynamic rebinding*, meaning that when the caller provides a value for a parameter that is already specified by the declaration site, the new value hides the old one. This means that only the value from the call site is actually used. This (4) gives a more precise typing for this situation.

**SEMANTICS.** Implicit parameters can be implemented by passing around a hidden dictionary that provides values to the implicit parameters. Accessing a parameter then becomes a lookup in the dictionary and the new `letdyn` construct extends the dictionary. To elucidate how such hidden dictionaries are propagated through the program when using lambda abstractions and applications, we present a simple semantics for implicit parameters. The goal here is not to prove properties of the language, but simply to provide a better explanation. A detailed semantics in terms of indexed comonads is shown in Chapter 4.

For simplicity, we assume that all implicit parameters have a type  $\sigma$ . In that setting, coeffect annotations  $r$  are just sets of names, i. e.  $r, s, t \subseteq \text{Names}$ . Given an expression  $e$  of type  $\tau$  that requires free variables  $\Gamma$  and implicit parameters  $r$ , the semantics is a function that takes a product of variables from  $\Gamma$  together with a dictionary of implicit parameters and returns  $\tau$ :



The semantics is defined inductively over the typing derivation:

$$\begin{aligned}
\llbracket \Gamma @ \mathbf{r} \vdash x_i : \tau_i \rrbracket &= \lambda((x_1, \dots, x_n), \_) \rightarrow x_i & (var) \\
\llbracket \Gamma @ \mathbf{r} \vdash ?p : \sigma \rrbracket &= \lambda(\_, f) \rightarrow f ?p & (param) \\
\llbracket \Gamma @ \mathbf{r} \vdash e : \tau \rrbracket &= \lambda(x, f) \rightarrow \llbracket \Gamma @ \mathbf{r}' \vdash e : \tau \rrbracket (x, f|_{\mathbf{r}'}) & (sub) \\
\llbracket \Gamma @ \mathbf{r} \vdash \lambda y. e : \tau_1 \xrightarrow{s} \tau_2 \rrbracket &= \lambda((x_1, \dots, x_n), f) \rightarrow \\
&\quad \lambda(y, g) \rightarrow \llbracket \Gamma, y : \tau_1 @ \mathbf{r} \cup \mathbf{s} \vdash e : \tau_2 \rrbracket ((x_1, \dots, x_n, y), f \uplus g) & (abs) \\
\llbracket \Gamma @ \mathbf{r} \cup \mathbf{s} \cup \mathbf{t} \vdash e_1 e_2 : \tau_2 \rrbracket &= \lambda(x, f) \rightarrow \\
&\quad \text{let } g = \llbracket \Gamma @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \rrbracket (x, f|_{\mathbf{r}}) & (app) \\
&\quad \text{in } g (\llbracket \Gamma @ \mathbf{s} \vdash e_2 : \tau_1 \rrbracket (x, f|_{\mathbf{s}}), f|_{\mathbf{t}}) \\
\llbracket \Gamma @ \mathbf{r} \cup (\mathbf{s} \setminus \{?p : \tau_1\}) \vdash \text{letdyn } ?p = e_1 \text{ in } e_2 : \tau_2 \rrbracket &= \lambda(x, f) \rightarrow \\
&\quad \text{let } v = \llbracket \Gamma @ \mathbf{r} \vdash e_1 : \tau_1 \rrbracket (x, f|_{\mathbf{r}}) & (letdyn) \\
&\quad \text{in } \llbracket \Gamma @ \mathbf{s} \vdash e_2 : \tau_2 \rrbracket (x, f|_{\mathbf{s} \setminus \{?p : \tau_1\}} \uplus \{?p \mapsto v\})
\end{aligned}$$

Here  $\uplus$  and  $f|_{\mathbf{r}}$  are auxiliary definitions:

$$\begin{aligned}
f|_{\mathbf{r}} &= \{(p, v) \mid (p, v) \in f, p \in \mathbf{r}\} \\
f \uplus g &= f|_{\text{dom}(f) \setminus \text{dom}(g)} \cup g
\end{aligned}$$

Figure 10: Semantics of a language with implicit parameters

$$\llbracket x_1 : \tau_1, \dots, x_n : \tau_n @ \mathbf{r} \vdash e : \tau \rrbracket : (\tau_1 \times \dots \times \tau_n) \times (\mathbf{r} \rightarrow \sigma) \rightarrow \tau$$

The dictionary is represented as a function from  $\mathbf{r}$  to  $\sigma$ . This means that it provides a  $\sigma$  value for all implicit parameters that are required according to the typing. Note that the domain of the function is not the set of all possible implicit parameter names, but only the finite subset of names that are required according to the typing.

The dictionary is also attached to the inputs of all functions. That is, a function  $\tau_1 \xrightarrow{s} \tau_2$  is interpreted by a function that takes  $\tau_1$  together with a dictionary that defines values for implicit parameters in  $\mathbf{s}$ :

$$\llbracket \tau_1 \xrightarrow{s} \tau_2 \rrbracket = \tau_1 \times (\mathbf{s} \rightarrow \sigma) \rightarrow \tau_2$$

The definition of the semantics is shown in Figure 10. The *(var)* and *(param)* rules are simple – they project the appropriate variable and implicit parameter, respectively.

When an expression requires implicit parameters  $\mathbf{r}$ , the semantics always provides a dictionary defined *exactly* on  $\mathbf{r}$ . To achieve this, the *(sub)* rule restricts the function to  $\mathbf{r}'$  (which is valid because  $\mathbf{r}' \subseteq \mathbf{r}$ ).

The most interesting rules are *(abs)* and *(app)*. In abstraction, we get two dictionaries  $f$  and  $g$  (from the declaration site and call site, respectively), which are combined and passed to the body of the function. The semantics prefers values from the call site, which is captured by the  $\uplus$  operation. In application, we first evaluate the expression  $e_1$ , then  $e_2$  and finally call the returned function. The three calls use (possibly overlapping) restrictions of the dictionary as required by the static types.

Finally, the *(letdyn)* rule specifies the semantics of the `letdyn` construct, which assigns a value to an implicit parameter. This is similar to *(app)*, because it needs to evaluate the sub-expression first. After evaluating  $e_1$ , the result is added to the dictionary using  $\uplus$ . The semantics of ordinary `let`



binding is omitted, because let binding can be treated as a syntactic sugar for  $(\lambda x.e_2) e_1$ .

Without providing a proof here, we note that the semantics is sounds with respect to the type system – when evaluating an expression, it provides it with a dictionary that is guaranteed to contain values for all implicit parameters that may be accessed. This can be easily checked by examining the semantic rules (and noting that the restriction and union always provide the expected set of parameters).

**MONADIC SEMANTICS.** Implicit parameters are related to the *reader monad*. The type  $\tau_1 \times (\mathbf{r} \rightarrow \sigma) \rightarrow \tau_2$  is equivalent to  $\tau_1 \rightarrow ((\mathbf{r} \rightarrow \sigma) \rightarrow \tau_2)$  through currying. Thus, we can express the function as  $\tau_1 \rightarrow M\tau_2$  for  $M\tau = (\mathbf{r} \rightarrow \sigma) \rightarrow \tau$ . Indeed, the reader monad can be used to model dynamic scoping. However, there is an important distinction from implicit parameters. The usual monadic semantics models fully dynamic scoping, while implicit parameters combine lexical and dynamic scoping.

When using the usual monadic semantics based on the reader monad, the semantics of the (*abs*) rule would be modified as follows:

$$\begin{aligned} \llbracket \Gamma @ \emptyset \vdash \lambda y.e : \tau_1 \xrightarrow{s} \tau_2 \rrbracket &= \lambda((x_1, \dots, x_n), \_) \rightarrow \\ &\lambda(y, g) \rightarrow \llbracket \Gamma, y : \tau_1 @ s \vdash e : \tau_2 \rrbracket ((x_1, \dots, x_n, y), g) \end{aligned}$$

Note that the declaration site dictionary is ignored and the body is called with only the dictionary provided by the call site. This is a consequence of the fact that monadic functions are always pure values created using monadic *unit*, which turns a function  $\tau_1 \rightarrow M^{\mathbf{r}}\tau_2$  into a monadic computation with no side-effects  $M^{\emptyset}\tau_1 \rightarrow M^{\mathbf{r}}\tau_2$ .

As we discuss later in Section 4.6.2, the reader monad can be extended to model rebinding. However, later examples in this chapter, such as liveness in Section 3.2.3 show that other context-aware computations cannot be captured by *any* monad.

**TYPE CLASSES.** Another type of constraints in Haskell that is closely related to implicit parameters are *type class* constraints [102]. They provide a principled form of ad-hoc polymorphism (overloading). When code uses an overloaded operation (e. g. comparison or numeric operators) a constraint is placed on the context in which the operation is used. For example:

```
twoTimes :: Num α ⇒ α → α
twoTimes x = x + x
```

The constraint  $\text{Num } \alpha$  on the function type arises from the use of the  $+$  operator. Similarly to implicit parameters, type classes can be implemented using a hidden dictionary. In the above case, the function `twoTimes` takes an additional dictionary that provides an operation  $+$  of type  $\alpha \times \alpha \rightarrow \alpha$ .

Type classes could be modelled as a coeffect system. The type system would annotate the context with a set of required type classes. The typing of the body of `twoTimes` would look as follows:

$$x : \alpha @ \{\text{Num}_\alpha\} \vdash x + x : \alpha$$

Similarly, the semantics of a language with type class constraints can be defined in a way similar to implicit parameters. The interpretation of the body is a function that takes  $\alpha$  together with a hidden dictionary of operations:  $\alpha \times \text{Num}_\alpha \rightarrow \alpha$ .

Type classes and implicit parameters show two important points about flat coefficient systems. First, the context requirements are associated with some *scope*, such as the body of a function. Second, they are associated with the input. To call a function that takes an implicit parameter or has a type-class constraint, the caller needs to pass a (hidden) parameter together with the function inputs.

**SUMMARY.** Implicit parameters are the simplest example of a system where function abstraction does not “delay” all impurities of the body. Here, the term “delay” refers to the fact that some implicit parameters may be captured (from the declaration site) at the time when the function is defined, but before it is executed. As discussed in Section 3.1.1, this is the defining feature of *coeffect* systems.

In this section, we have seen how this affects both the type system and the semantics of the language. In the type system, the (*abs*) rule places context-requirements on both the declaration site and the call site. For implicit parameters, this rule introduces non-determinism in the type-inference, because the parameters can be split arbitrarily. However, as we show in the next section, this is not always the case. Semantically, lambda abstraction *merges* two parts of context (implicit parameter dictionaries) that are provided by the call site and declaration site.

### 3.2.2 Distributed computing

Distributed programming was used as one of the motivating examples for coeffects in Chapter 1. This section explores the use case. We look at rebinding resources and cross-compilation. The structure of both is very similar to implicit parameters and type class constraints, but they demonstrate that there is a broader use for coeffect systems.

**REBINDABLE RESOURCES.** The need for parameters that support dynamic scoping also arises in distributed computing. To quote an example discussed by Bierman et al. [10]: “Dynamic binding is required in various guises, for example when a marshalled value is received from the network, containing identifiers that must be rebound to local resources.”

Rebindable parameters are identifiers that refer to some specific resource. When a function value is marshalled and sent to another machine, rebinding resources can be handled in two ways. First, if the resource is available on the target machine, the parameter may be *rebound* to the resource on the new machine. This is captured by the dynamic scoping rule. Second, if the resource is not available on the target machine, the resource is either marshalled or a *remote reference* is created. This is captured by the lexical scoping rule.

A practical language that supports rebinding resources is for example Acute [78]. In the following example, we use the construct `access Res` to represent access to a rebinding resource named `Res`. The following simple function accesses a database together with a current date; then it filters from the database based on the date:

```
let recentNews = λ() →
  let db = access News in
  query db "SELECT * WHERE Date > %1" (access Clock)
```

```

// Checks that input is valid; can run on both server and client
let validateInput = λname →
  name ≠ "" && forall isLetter name

// Searches database for a product; must run on the server-side
let retrieveProduct = λname →
  if validateInput name then Some(queryProductDb name)
  else None

// Client-side function to show price or error (for invalid inputs)
let showPrice = λname →
  if validateInput name then
    match (remote retrieveProduct()) with
    | Some p → showPrice (getPrice p)
    | None → showError "Invalid input on the server"
  else showError "Invalid input on the client"

```

---

Figure 11: Sample client-server application with input validation

When `recentNews` is created on the server and sent to the client, a remote reference to the database (available only on the server) must be captured. If the client device supports a clock, then `Clock` can be locally *rebound*, e. g., to accommodate time-zone changes. Otherwise, the date and time needs to be obtained from the server too.

The type system and semantics for rebindable resources are essentially the same as those for implicit parameters. Primitive requirements are introduced by the `access` keyword. Lambda abstraction splits the requirements between declaration site (capturing remote reference) and call site (representing rebinding). For this reason, we do not discuss the system in details and instead look at other uses.

**CROSS-COMPILATION.** A related issue with distributed programming is the need to target increasing number of diverse platforms. Modern applications often need to run on multiple platforms (iOS, Android, Windows or as JavaScript) or multiple versions of the same platform. Many programming languages are capable of targeting multiple different platforms. For example, functional languages that can be compiled to native code and JavaScript include, among others, F#, Haskell and OCaml [97].

Links [21], F# WebTools and WebSharper [84, 67], ML5 and QWeSST [56, 76] and Hop [52] go further and allow including code for multiple distinct platforms in a single source file. A single program is then automatically split and compiled to multiple target runtimes. This poses additional challenges – it is necessary to check where each part of the program can run and statically guarantee that it will be possible to compile code to the required target platform (safe *multi-targeting*).

We demonstrate the problem by looking at input validation. In applications that communicate over an unsecured HTTP channel, user input needs to be validated interactively on the client-side (to provide immediate response) and then again on the server-side (to guarantee safety).

Consider the client-server example in Figure 11. The `retrieveProduct` function represents the server-side, while `showPrice` is called on the client-side and performs a remote call to the server-side function (how this is imple-

a.) Set-based type system for cross-compilation, inspired by Links [21]

$$\begin{aligned}
(sub) \quad & \frac{\Gamma @ \mathbf{r}' \vdash e : \tau}{\Gamma @ \mathbf{r} \vdash e : \tau} \quad (\mathbf{r}' \supseteq \mathbf{r}) \\
(app) \quad & \frac{\Gamma @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \quad \Gamma @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma @ \mathbf{r} \cap \mathbf{s} \cap \mathbf{t} \vdash e_1 \ e_2 : \tau_2} \\
(abs) \quad & \frac{\Gamma, x : \tau_1 @ \mathbf{r} \cup \mathbf{s} \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{s} \tau_2}
\end{aligned}$$

b.) Version-based type system, inspired by Android API level [25]

$$\begin{aligned}
(sub) \quad & \frac{\Gamma @ \mathbf{r}' \vdash e : \tau}{\Gamma @ \mathbf{r} \vdash e : \tau} \quad (\mathbf{r}' \leq \mathbf{r}) \\
(app) \quad & \frac{\Gamma @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \quad \Gamma @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma @ \max\{\mathbf{r}, \mathbf{s}, \mathbf{t}\} \vdash e_1 \ e_2 : \tau_2} \\
(abs) \quad & \frac{\Gamma, x : \tau_1 @ \mathbf{r} \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{r} \tau_2}
\end{aligned}$$

Figure 12: Two variants of coeffect typing rules for cross-compilation

mented is not our concern here). To ensure that the input is valid *both* functions call `validateInput` – however, this is fine, because `validateInput` uses only basic functions and language features that can be cross-compiled to both client-side and server-side.

In Links [21], functions can be annotated as client-side, server-side and database-side. F# WebTools [67] supports cross-compiled (mixed-side) functions similar to `validateInput`. However, these are single-purpose language features and they are not extensible. A practical implementation needs to be able to capture multiple different patterns – sets of environments (client, server, mobile) for distributed computing, but also Android API level [25] to cross-compile for multiple versions of the same platform.

**TYPE SYSTEMS.** Cross-compilation may seem similar to resource tracking (and thus to the tracking of implicit parameters), but it actually demonstrates a couple of new ideas that are important for flat coeffect systems. Unlike with implicit parameters, we will not present a specific existing system in this section – instead we briefly look at two examples that let us explore the range of possibilities.

In the first system, shown in Figure 12 (a), the coeffect annotations are sets of execution environments, i.e.  $\mathbf{r}, \mathbf{s}, \mathbf{t} \subseteq \{\text{client}, \text{server}, \text{database}\}$ . Sub-coeffecting (*sub*) lets us ignore some of the supported execution environments; application (*app*) can be only executed in the *intersection* of the environments required by the two expressions and the function value.

Sub-coeffecting and application are trivially dual to the rules for implicit parameters. We just track supported environments using intersection as opposed to tracking required parameters using union. However, this symmetry does not hold for lambda abstraction (*abs*), which still uses *union*. This models the case when there are two ways of executing the function:

- The function is represented as executable code for an call site environment and is executed there, possibly after it is marshalled and transferred to another machine.
- The function body is compiled for the declaration site environment; the value that is returned is a remote reference to the code and function calls are performed as remote invocations.

This example ignores important considerations – for example, it is likely desirable to make this difference explicit (e.g. using explicit wrapping of unevaluated expressions) and the implementation also needs to be clarified. For a system that does this, see e.g. ML5 [56]). The key point of our brief example is that the algebraic structure of coeffect annotations may be more complex and use, for example,  $\cap$  for application and  $\cup$  for abstraction.

The second system, shown in Figure 12 (b) is inspired by the API level requirements in Android. Coeffect annotations are simply numbers representing the level ( $r, s, t \in \mathbb{N}$ ). Levels are ordered increasingly, so we can always require higher level (*sub*). The requirement on function application (*app*) is the highest level of the levels required by the sub-expressions and the function. The system uses yet another variant of lambda abstraction (*abs*) – the requirements of the body are duplicated and placed on *both* the declaration site and the call site.

The ML5 language [56] mentioned above served as an inspiration for our example. It tracks execution environments using modalities of modal S4 to represent the environment – this approach is similar to coeffects, both from the practical perspective, but also through deeper theoretical links. However, it is based on the *meta-language* style of embedding modalities rather than on the *language-semantics* style (see Section 2.2.1). We return to this topic in Section 6.3.

### 3.2.3 Liveness analysis

Our next example shows the idea of coeffects from a different perspective. Rather than keeping additional information independent of the variable context, we track properties about how variables are used. Nevertheless, we still look at the left-hand side of  $\vdash$  and the structure of the typing rules and semantics will be very similar.

*Live variable analysis* (LVA) [5] is a standard technique in compiler theory. It detects whether a free variable of an expression may be used by a program during its evaluation (it is *live*) or whether it is definitely not needed (it is *dead*). As an optimization, compiler can remove bindings to dead variables as they are never accessed. Wadler [100] describes the property of a variable that is dead as the *absence* of a variable.

**FLAT LIVENESS ANALYSIS.** In this section, we discuss a restricted form of liveness analysis. We do not track liveness of *individual* variables, but of the *entire* variable context. This is not practically useful, but it provides an interesting insight into how flat coeffects work. A per-variable liveness analysis can be captured using structural coeffects and is discussed in Section 3.3.1. Consider the following two examples:

```
let constant42 =  $\lambda x \rightarrow 42$ 
let constant =  $\lambda \text{value} \rightarrow \lambda x \rightarrow \text{value}$ 
```

a.) The operations of a two-point lattice  $\mathcal{L} = \{L, D\}$  where  $D \sqsubseteq L$  are defined as:

$$\begin{array}{llll} L \sqcup L = L & L \sqcup D = D & L \sqcap L = L & L \sqcap D = L \\ D \sqcup L = D & D \sqcup D = D & D \sqcap L = L & D \sqcap D = D \end{array}$$

b.) Sequential composition of (semantic) functions composes annotations using  $\sqcup$ :

$$\begin{array}{lll} f : \tau_1 \xrightarrow{r} \tau_2 & g : \tau_2 \xrightarrow{s} \tau_3 & g \circ f : \tau_1 \xrightarrow{r \sqcup s} \tau_3 \\ f : \tau_1 \xrightarrow{L} \tau_2 & g : \tau_2 \xrightarrow{L} \tau_3 & g \circ f : \tau_1 \xrightarrow{L} \tau_3 \quad (1) \\ f : \tau_1 \xrightarrow{D} \tau_2 & g : \tau_2 \xrightarrow{L} \tau_3 & g \circ f : \tau_1 \xrightarrow{D} \tau_3 \quad (2) \\ f : \tau_1 \xrightarrow{L} \tau_2 & g : \tau_2 \xrightarrow{D} \tau_3 & g \circ f : \tau_1 \xrightarrow{D} \tau_3 \quad (3) \\ f : \tau_1 \xrightarrow{D} \tau_2 & g : \tau_2 \xrightarrow{D} \tau_3 & g \circ f : \tau_1 \xrightarrow{D} \tau_3 \quad (4) \end{array}$$

c.) Pointwise composition of (semantic) functions composes annotations using  $\sqcap$ :

$$\begin{array}{lll} f : \tau_1 \xrightarrow{r} \tau_2 & h : \tau_1 \xrightarrow{s} \tau_3 & \langle f, h \rangle : \tau_1 \xrightarrow{r \sqcap s} \tau_2 \times \tau_3 \\ f : \tau_1 \xrightarrow{D} \tau_2 & h : \tau_1 \xrightarrow{D} \tau_3 & \langle f, h \rangle : \tau_1 \xrightarrow{D} \tau_2 \times \tau_3 \quad (1) \\ f : \tau_1 \xrightarrow{D} \tau_2 & h : \tau_1 \xrightarrow{L} \tau_3 & \langle f, h \rangle : \tau_1 \xrightarrow{L} \tau_2 \times \tau_3 \quad (2) \\ f : \tau_1 \xrightarrow{L} \tau_2 & h : \tau_1 \xrightarrow{D} \tau_3 & \langle f, h \rangle : \tau_1 \xrightarrow{L} \tau_2 \times \tau_3 \quad (3) \\ f : \tau_1 \xrightarrow{L} \tau_2 & h : \tau_1 \xrightarrow{L} \tau_3 & \langle f, h \rangle : \tau_1 \xrightarrow{L} \tau_2 \times \tau_3 \quad (4) \end{array}$$

Figure 13: Liveness annotations with sequential and pointwise composition

The body of the first function is just a constant 42 and so the context of the body is marked as *dead*. The parameter (call site) of the function is not used and can also be marked as dead. Similarly, no variables from the declaration site are used and so they are also marked as dead.

In contrast, the body of the second function accesses a variable value and so the body of the function is marked as *live*. In the flat system, we do not track *which* variable was used and so we have to mark both the call site and the declaration site as live (this will be refined in a structural version of the system).

**FORWARD VS. BACKWARD & MAY VS. MUST.** Static analyses can be classified as either *forward* or *backward* (depending on how they propagate information) and as either *must* or *may* (depending on what properties they guarantee). Liveness is a *backward* analysis – the requirements are propagated from variables to their declarations. The distinction between *must* and *may* is apparent when we look at an example with conditionals:

```
let defaultArg = λcond → λinput →
  if cond then 42 else input
```

Liveness analysis is a *may* analysis meaning that it marks variable as live when it *may* be used and as dead if it is *definitely* not used. This means that the variable `input` is *live* in the example above. A *must* analysis would mark the variable only if it was used in both of the branches (this is sometimes called *neededness* or *very busy* variable/expression).

The distinction between *may* and *must* analyses demonstrates the importance of interaction between contextual properties and certain language constructs such as conditionals.

$$\begin{array}{l}
\text{(var)} \quad \frac{x : \tau \in \Gamma}{\Gamma @ \mathbf{L} \vdash x : \tau} \\
\text{(const)} \quad \frac{c : \tau \in \Delta}{\Gamma @ \mathbf{D} \vdash c : \tau} \\
\text{(sub)} \quad \frac{\Gamma @ \mathbf{r}' \vdash e : \tau}{\Gamma @ \mathbf{r} \vdash e : \tau} \quad (\mathbf{r}' \sqsubseteq \mathbf{r}) \\
\text{(app)} \quad \frac{\Gamma @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{\mathbf{t}} \tau_2 \quad \Gamma @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma @ \mathbf{r} \sqcup (\mathbf{s} \sqcap \mathbf{t}) \vdash e_1 e_2 : \tau_2} \\
\text{(let)} \quad \frac{\Gamma @ \mathbf{r} \vdash e_1 : \tau_1 \quad \Gamma, x : \tau_1 @ \mathbf{s} \vdash e_2 : \tau_2}{\Gamma @ \mathbf{s} \vdash \text{let } x = e_1 \text{ in } e_2 : \tau_2} \\
\text{(abs)} \quad \frac{\Gamma, x : \tau_1 @ \mathbf{r} \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{r}} \tau_2}
\end{array}$$

Figure 14: Coeffect rules for tracking whole-context liveness

**TYPE SYSTEM.** A type system that captures whole-context liveness annotates the context with value of a two-point lattice  $\mathcal{L} = \{\mathbf{L}, \mathbf{D}\}$ . The annotation  $\mathbf{L}$  marks the context as *live* and  $\mathbf{D}$  stands for a *dead* context. Figure 13 (a) defines the ordering  $\sqsubseteq$ , meet  $\sqcap$  and join operations  $\sqcup$  of the lattice.

The typing rules for tracking whole-context liveness are shown in Figure 14. The language now includes constants  $c : \tau \in \Delta$ . Accessing a constant (*const*) annotates the context as dead using  $\mathbf{D}$ . This contrasts with variable access (*var*), which marks the context as live using  $\mathbf{L}$ . A dead context (definitely not needed) can be treated as live context using the (*sub*) rule. This captures the *may* nature of the analysis.

The (*app*) rule is best understood by discussing its semantics. The semantics uses *sequential composition* to compose the semantics of  $e_2$  with the function obtained as the result of  $e_1$ . However, we need more than just sequential composition. The same input context is passed to the expression  $e_1$  (in order to get the function value) and to the sequentially composed function (to evaluate  $e_2$  followed by the function call). This is captured by *pointwise composition*.

Consider first *sequential composition* of (semantic) functions  $f, g$  annotated with  $\mathbf{r}, \mathbf{s}$ . The composed function  $g \circ f$  is annotated with  $\mathbf{r} \sqcup \mathbf{s}$  as shown in Figure 13 (b). The argument of the function  $g \circ f$  is live only when the arguments of both  $f$  and  $g$  are live (1). When the argument of  $f$  is dead, but  $g$  requires  $\tau_2$  (2), we can evaluate  $f$  without any input and obtain  $\tau_2$ , which is then passed to  $g$ . When  $g$  does not require its argument (3,4), we can just evaluate  $g$ , without evaluating  $f$ . Here, the semantics *implements* the dead code elimination optimization.

Secondly, a *pointwise composition* passes the same argument to  $f$  and  $h$ . The parameter is live if either the parameter of  $f$  or  $h$  is live. The pointwise composition is written as  $\langle f, h \rangle$  and it combines annotations using  $\sqcap$  as shown in Figure 13 (c). Here, the argument is not needed only when both  $f$  and  $h$  do not need it (1). In all other cases, the parameter is needed and is then used either once (2,3) or twice (4). The rule for function application (*app*) combines the two operations. The context  $\Gamma$  is live if it is needed by  $e_1$  (which always needs to be evaluated) *or* when it is needed by the function value *and* by  $e_2$ .



The *(abs)* rule duplicates the annotation of the body, similarly to the cross-compilation example in Figure 12. When the body accesses any variables, it requires both the argument and the variables from declaration site. When it does not use any variables, it marks both as dead. Finally, the *(let)* rule annotates the composed expression with the liveness of the expression  $e_2$  – if the context of  $e_2$  is live, then it also requires variables from  $\Gamma$ ; if it is dead, then it does not require  $\Gamma$  or  $x$ . As further discussed later in Section 4.3.3, the *(let)* rule is again just a syntactic sugar for  $(\lambda x. e_2) e_1$ . This follows from the simple observation that  $r \sqcup (s \sqcap r) = r$ .

**EXAMPLES.** Before looking at the semantics, we consider a number of simple examples to demonstrate the key aspects of the system. Full typing derivations are shown in Appendix A.2:

- |                       |     |
|-----------------------|-----|
| $(\lambda x. 42) y$   | (1) |
| $\text{twoTimes } 42$ | (2) |
| $(\lambda x. x) 42$   | (3) |

In the first case, the context is dead. In (1), the function’s parameter is dead and so the overall context is dead, even though the argument uses a variable  $y$  – the semantics evaluates the function without passing it an actual argument. In the second case (2), the function is a variable that needs to be obtained and so the context is live. In the last case (3), the function accesses a variable and so its declaration site is marked as requiring the context (*abs*). This is where structural coeffect analysis would be more precise – the system shown here cannot capture the fact that  $x$  is a bound variable.

**SEMANTICS.** As showed in the examples, the type system for the liveness coeffect calculus marks the context of an expression  $(\lambda x. 42) y$  as dead. This means that the semantics of the above expression must not evaluate the argument  $y$ . In other words, the type system is only sound if the semantics includes dead code elimination.

To capture dead code elimination in the semantics, we add a special empty value and pass it as an argument to a function whose argument is not needed, so  $(\lambda x. 42)$  will be called with an empty value as argument (because it does not need its argument).

We can represent such empty values using the option type (known as *Maybe* in Haskell). We use the notation  $\tau + 1$  to denote option types. Given a context with variables  $x_i$  of type  $\tau_i$ , the semantics is a function taking  $(\tau_1 \times \dots \times \tau_n) + 1$ . When the context is live, it will be called with the left value (product of variable assignments); when the context is dead, it will be called with the right value (containing no information).

However, ordinary option type is not sufficient. We need to capture the fact that the representation depends on the annotation – in other words, the type is *indexed* by the coeffect annotation. The indexing is discussed in details in Section 4.3.3. For now, it suffices to define the semantics using two separate rules:

$$\begin{aligned} \llbracket x_1 : \tau_1, \dots, x_n : \tau_n @ L \vdash e : \tau \rrbracket & : (\tau_1 \times \dots \times \tau_n) \rightarrow \tau \\ \llbracket x_1 : \tau_1, \dots, x_n : \tau_n @ D \vdash e : \tau \rrbracket & : 1 \rightarrow \tau \end{aligned}$$



The semantics of functions is defined similarly. When the argument of a function is live, the function takes the input value; when the argument is dead, the semantic function takes a unit as its argument:

$$\begin{aligned} \llbracket \tau_1 \xrightarrow{L} \tau_2 \rrbracket &= \tau_1 \rightarrow \tau_2 \\ \llbracket \tau_1 \xrightarrow{D} \tau_2 \rrbracket &= 1 \rightarrow \tau_2 \end{aligned}$$

Unlike with implicit parameters, the coeffect system for liveness tracking cannot be modelled using monads. Any monadic semantics would express functions as  $\tau_1 \rightarrow M \tau_2$ . Unless laziness is already built-in, there is no way to call such function without first obtaining a value  $\tau_1$ . The above semantics makes this possible by taking a unit 1 when the argument is not live.

In Figure 15, we define the semantics directly. We write  $()$  for the only value of type 1. This appears, for example, in  $(const)$  which takes  $()$  as the input and returns a constant using a global dictionary  $\delta$ . In  $(var)$ , the context is live and so the semantics performs a projection. Sub-coeffecting is captured by two rules. A dead context can be treated as live using  $(abs-1)$ ; in other cases, the annotation is not changed  $(abs-2)$ .

Lambda abstraction can be annotated in just two ways. When the body requires context  $(abs-1)$ , the value of a bound variable  $y$  is added to the context  $\Gamma$  before passing it to the body. When the body does not require context  $(abs-2)$ , it is called with  $()$  as the input.

For application, there are 8 possible combinations of annotations. The semantics of some of them is the same, so we only need to show 3 cases. The rules should be read as ML-style pattern matching, where the last rule handles all cases not covered by the first two. In  $(app-1)$ , we handle the case when the function  $g$  does not require its argument –  $e_2$  is not used and instead, the function is called with  $()$  as the argument. The case  $(app-2)$  covers the case when the expression  $e_1$  does not require a context, but  $e_1$  does. Finally, in  $(app-3)$ , the same input (which may be either tuple of variables or unit) is propagated uniformly to both  $e_1$  and  $e_2$ .

**SUMMARY.** Unlike with implicit parameters, lambda abstraction for liveness analysis does not introduce non-determinism. It simply duplicates the context requirements. However, this still matches the property of coeffects that impurities cannot be delayed or thunked and attached just to the function arrow – we place requirements on both call site and declaration site.

The semantics of liveness reveals three interesting properties. Firstly, the coeffect calculus for liveness cannot be modelled as a monadic computation of the form  $\tau_1 \rightarrow M \tau_2$ . Secondly, the system would not work without the coeffect annotations. The shape of the semantic function depends on the annotation (the input is either 1 or  $\tau$ ) and is *indexed* by the annotation.

Finally, we discussed how the semantics of application arises from *sequential* and *pointwise* composition. This is an important aspect of coeffect systems – categorical semantics typically builds on *sequential* composition, but to model full  $\lambda$  calculus it needs more. For coeffects, we need *pointwise* composition where the same context is shared by multiple sub-expressions.

### 3.2.4 Data-flow languages

We used implicit parameters as our first example, because they show the simplest form of coeffects. Liveness requires a richer coeffect annotation structure, but the flat version is not practical. In this section, we look at a system with a structure similar to liveness that is not a toy example.

$$\begin{aligned}
\llbracket \Gamma @ L \vdash x_i : \tau_i \rrbracket &= \lambda(x_1, \dots, x_n) \rightarrow x_i & (var) \\
\llbracket \Gamma @ D \vdash c_i : \tau_i \rrbracket &= \lambda() \rightarrow \delta(c_i) & (const) \\
\\
\llbracket \Gamma @ L \vdash e : \tau \rrbracket &= \lambda x \rightarrow \llbracket \Gamma @ D \vdash e : \tau \rrbracket () & (sub-1) \\
\llbracket \Gamma @ r \vdash e : \tau \rrbracket &= \lambda x \rightarrow \llbracket \Gamma @ r \vdash e : \tau \rrbracket x & (sub-2) \\
\\
\llbracket \Gamma @ L \vdash \lambda y. e : \tau_1 \xrightarrow{L} \tau_2 \rrbracket &= \lambda(x_1, \dots, x_n) \rightarrow & (abs-1) \\
&\quad \lambda y \rightarrow \llbracket \Gamma, y : \tau_1 @ L \vdash e : \tau_2 \rrbracket (x_1, \dots, x_n, y) \\
\llbracket \Gamma @ D \vdash \lambda y. e : \tau_1 \xrightarrow{D} \tau_2 \rrbracket &= \lambda() \rightarrow & (abs-2) \\
&\quad \lambda() \rightarrow \llbracket \Gamma, y : \tau_1 @ D \vdash e : \tau_2 \rrbracket () \\
\\
\llbracket \Gamma @ r \vdash e_1 e_2 : \tau_2 \rrbracket &= \lambda x \rightarrow & (app-1) \\
&\quad \text{let } g = \llbracket \Gamma @ r \vdash e_1 : \tau_1 \xrightarrow{D} \tau_2 \rrbracket x \text{ in } g () \\
\llbracket \Gamma @ L \vdash e_1 e_2 : \tau_2 \rrbracket &= \lambda x \rightarrow & (app-2) \\
&\quad \text{let } g = \llbracket \Gamma @ L \vdash e_1 : \tau_1 \xrightarrow{L} \tau_2 \rrbracket x \text{ in } g (\llbracket \Gamma @ D \vdash e_2 : \tau_1 \rrbracket ()) \\
\llbracket \Gamma @ r \sqcup (s \sqcap t) \vdash e_1 e_2 : \tau_2 \rrbracket &= \lambda x \rightarrow & (app-3) \\
&\quad \text{let } g = \llbracket \Gamma @ r \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \rrbracket x \text{ in } g (\llbracket \Gamma @ s \vdash e_2 : \tau_1 \rrbracket x)
\end{aligned}$$

Figure 15: Semantics that implements dead code elimination for  $\lambda$ -calculus

The Section 1.1.4 briefly demonstrated that we can treat array access as an operation that accesses a context. In case of arrays, the context is neighbourhood of a current location in the array specified by a cursor. In this section, we make the example more concrete, using a simpler and better studied programming model, data-flow languages.

Lucid [99] is a declarative data-flow language designed by Wadge and Ashcroft. In Lucid, variables represent streams and programs are written as transformations over streams. A function application  $\text{square}(x)$  represents a stream of squares calculated from the stream of values  $x$ .

The data-flow approach has been successfully used in domains such as development of real-time embedded application where many *synchronous languages* [8] build on the data-flow paradigm. The following example is inspired by the Lustre [37] language and implements a program to count the number of edges on a Boolean stream:

```

let edge = false fby (input && not (prev input))

let edgeCount =
  0 fby ( if edge then 1 + (prev edgeCount)
        else prev edgeCount )

```

The construct  $\text{prev } x$  returns a stream consisting of previous values of the stream  $x$ . The second value of  $\text{prev } x$  is first value of  $x$  (and the first value is undefined). The construct  $y \text{ fby } x$  returns a stream whose first element is the first element of  $y$  and the remaining elements are values of  $x$ . Note that in Lucid, the constants such as `false` and `0` are constant streams.

$$\begin{array}{c}
\text{(var)} \quad \frac{x : \tau \in \Gamma}{\Gamma @ 0 \vdash x : \tau} \\
\text{(prev)} \quad \frac{\Gamma @ n \vdash e : \tau}{\Gamma @ n + 1 \vdash \text{prev } e : \tau} \\
\text{(sub)} \quad \frac{\Gamma @ n' \vdash e : \tau}{\Gamma @ n \vdash e : \tau} \quad (n' \leq n) \\
\text{(app)} \quad \frac{\Gamma @ m \vdash e_1 : \tau_1 \xrightarrow{p} \tau_2 \quad \Gamma @ n \vdash e_2 : \tau_1}{\Gamma @ \max(m, n + p) \vdash e_1 e_2 : \tau_2} \\
\text{(let)} \quad \frac{\Gamma @ m \vdash e_1 : \tau_1 \quad \Gamma, x : \tau_1 @ n \vdash e_2 : \tau_2}{\Gamma @ n + m \vdash \text{let } x = e_1 \text{ in } e_2 : \tau_2} \\
\text{(abs)} \quad \frac{\Gamma, x : \tau_1 @ n \vdash e : \tau_2}{\Gamma @ n \vdash \lambda x. e : \tau_1 \xrightarrow{n} \tau_2}
\end{array}$$

Figure 16: Coeffect rules for tracking context-usage in data-flow language

Formally, the constructs are defined as follows (writing  $x_n$  for  $n$ -th element of a stream  $x$ ):

$$(\text{prev } x)_n = \begin{cases} \text{nil} & \text{if } n = 0 \\ x_{n-1} & \text{if } n > 0 \end{cases} \quad (y \text{ fby } x)_n = \begin{cases} y_0 & \text{if } n = 0 \\ x_n & \text{if } n > 0 \end{cases}$$

When reading data-flow programs, we do not need to think about variables in terms of streams – we can see them as simple values. Most of the operations perform calculation just on the *current* value of the stream. However, the operation **fb**y and **prev** are different. They require additional *context* which provides past values of variables (for **prev**) and information about the current location in the stream (for **fb**y).

The semantics of Lucid-like languages can be captured using a number of mathematical structures. Wadge [98] originally defined a monadic semantics, while Uustalu and Vene later used comonads [92]. In Chapter 4, we extend the latter approach. The present chapter presents a sketch of a concrete data-flow semantics defined directly on streams.

In the introductory example with array access patterns, we used coeffects to track the range of values accessed. In this section, we look at a simpler example – we only consider the **prev** operation and track the maximal number of *past values* needed. This is an important information for efficient implementation of data-flow languages. When we can guarantee that at most  $x$  past values are accessed, the values can be stored in a pre-allocated buffer rather than using e. g. on-demand computed lazy streams.

**TYPE SYSTEM.** We can use a coeffect type system to track the maximal number of accessed past values. Here, the context is annotated with a single integer. The current value is always present, so 0 means that no past values are needed, but the current value is still available. The typing rules of the system are shown in Figure 16.

Variable access (*var*) annotates the context with 0; sub-coeffecting (*sub*) allows us to require more values than is actually needed. Primitive context-requirements are introduced in (*prev*), which increments the number of past values by one. Thus, for example, **prev** (**prev**  $x$ ) requires 2 past values.

The (*app*) rule follows the same intuition as for liveness. It combines *sequential* and *pointwise* composition of semantic functions. In case of data-flow, the operations combine annotations using  $+$  and *max* operations:

$$\begin{array}{lll} f : \tau_1 \xrightarrow{m} \tau_2 & g : \tau_2 \xrightarrow{n} \tau_3 & g \circ f : \tau_1 \xrightarrow{m+n} \tau_3 \\ f : \tau_1 \xrightarrow{m} \tau_2 & h : \tau_1 \xrightarrow{n} \tau_3 & \langle f, h \rangle : \tau_1 \xrightarrow{\max(m,n)} \tau_2 \times \tau_3 \end{array}$$

Sequential composition adds the annotations. The function  $f$  needs  $m$  past values to produce a single  $\tau_2$  value. To produce two  $\tau_2$  values, we thus need  $m + 1$  past values of  $\tau_1$ ; to produce three  $\tau_2$  values, we need  $m + 2$  past values of  $\tau_1$ , and so on. To produce  $n$  past values that are required as the input of  $g$ , we need  $m + n$  past values of type  $\tau_1$ . The pointwise composition is simpler. It uses the same stream to evaluate functions requiring  $m$  and  $n$  past values, and so it needs maximum of the two at most.

In summary, function application (*app*) requires maximum of the values needed to evaluate  $e_1$  and the number of values needed to evaluate the argument  $e_2$ , sequentially composed with the function.

In function abstraction (*abs*), the requirements of the body are duplicated on the declaration site and the call site as in liveness analysis. If the body requires  $n$  past values, it may access  $n$  values of any variables – including those available in  $\Gamma$ , as well as the parameter  $x$ . Finally, the (*let*) rule simply adds the two requirements. This corresponds to the sequential composition operation, but it is also a rule that we obtain by treating let-binding as a syntactic sugar for  $(\lambda x. e_2) e_1$ .

**EXAMPLE.** As with the liveness example, the application rule might require more explanation. The following example is somewhat arbitrary, but it demonstrates the rule well. We assume that *counter* is a stream of positive integers (starting from zero) and *tick* flips between 0 and 1. The full typing derivation is shown in Appendix A.3:

```
( if (prev tick) = 0
  then (λx → prev x)
  else (λx → x) ) (prev counter)
```

The left-hand side of the application returns a function depending on the *previous* value of *tick*. The resulting stream of functions flips between a function returning a current value and a function returning the previous value. If the current tick is 0, and the function is applied to a stream  $\langle \dots, 4, 3, 2, 1 \rangle$  (where 1 is the current value), it yields the stream  $\langle \dots, 4, 4, 2, 2 \rangle$ .

To obtain the function, we need one past value from the context (for *prev tick*). The returned function needs either none or one past value (thus a subtyping rule is required to type it as requiring one past value). So, the annotations for (*app*) are  $m = 1, p = 1$ . The function is called with *prev counter* as an argument, meaning that the result is either the first or second past element. Given *counter* =  $\langle \dots, 5, 4, 3, 2, 1 \rangle$ , the argument is  $\langle \dots, 5, 4, 3, 2 \rangle$  and so the overall result is a stream  $\langle \dots, 5, 5, 3, 3 \rangle$ . From the argument, we get the requirement  $n = 1$ .

Using the (*app*) rule, we get that the overall number of past elements needed is  $\max(1, 1 + 1) = 2$ . This should match the intuition about the code – when the first function is applied to the argument, the computation will first access *prev tick* (using one past value) and then *prev (prev counter)* (using two past values).

$$\begin{aligned}
\llbracket \Gamma @ 0 \vdash x_i : \tau_i \rrbracket &= \lambda \langle x_0, \dots, x_n \rangle \rightarrow x_i & (var) \\
\llbracket \Gamma @ n + 1 \vdash \text{prev } e : \tau \rrbracket &= \lambda \langle v_0, \dots, v_{n+1} \rangle \rightarrow \\
&\quad \llbracket \Gamma @ n \vdash e : \tau \rrbracket \langle v_1, \dots, v_{n+1} \rangle & (prev) \\
\llbracket \Gamma @ n \vdash e : \tau \rrbracket &= \lambda \langle v_0, \dots, v_n \rangle \rightarrow \\
&\quad \llbracket \Gamma @ n' \vdash e : \tau \rrbracket \langle v_0, \dots, v_{n'} \rangle & (sub) \\
\llbracket \Gamma @ n \vdash \lambda y. e : \tau_1 \xrightarrow{n} \tau_2 \rrbracket &= \lambda \langle v_0, \dots, v_n \rangle \rightarrow \\
&\quad \lambda (y, g) \rightarrow \llbracket \Gamma, y : \tau_1 @ n \vdash e : \tau_2 \rrbracket \langle (v_0, y_0), \dots, (v_n, y_n) \rangle & (abs) \\
\llbracket \Gamma @ \max(m, n + p) \vdash e_1 \ e_2 : \tau_2 \rrbracket &= \lambda \langle v_0, \dots, v_{\max(m, n + p)} \rangle \rightarrow \\
&\quad \text{let } g = \llbracket \Gamma @ m \vdash e_1 : \tau_1 \xrightarrow{p} \tau_2 \rrbracket \langle v_0, \dots, v_m \rangle \\
&\quad \text{in } g \ ( \llbracket \Gamma @ n \vdash e_2 : \tau_1 \rrbracket \langle v_0, \dots, v_n \rangle, \dots, \\
&\quad \quad \llbracket \Gamma @ n \vdash e_2 : \tau_1 \rrbracket \langle v_p, \dots, v_{n+p} \rangle ) & (app)
\end{aligned}$$

Figure 17: Semantics showing how past values are accessed in a data-flow language

**SEMANTICS.** The sample language discussed in this section is a *causal* data-flow language. This means that a computation can access *past* values of the stream (but not future values). In the semantics, we again need richer structure over the input.

Uustalu and Vene [93] model causal data-flow computations using a non-empty list  $\text{NeList } \tau = \tau \times (\text{NeList } \tau + 1)$  over the input. A function  $\tau_1 \rightarrow \tau_2$  is thus modelled as  $\text{NeList } \tau_1 \rightarrow \tau_2$ . This model is difficult to implement efficiently, as it creates unbounded lists of past elements.

The coeffect system tracks maximal number of past values and so we can define the semantics using a list of fixed length. As with liveness, this is a data structure *indexed* by the coeffect annotation. We write  $\tau^n$  for a list containing  $n$  elements (which can be also viewed as an  $n$ -element product  $\tau \times \dots \times \tau$ ).

As with the previous examples, our semantics interprets a judgement using a (semantic) function; functions in the language are modelled as functions taking a list of inputs:

$$\begin{aligned}
\llbracket x_1 : \tau_1, \dots, x_n : \tau_n @ n \vdash e : \tau \rrbracket &: (\tau_1 \times \dots \times \tau_n)^{n+1} \rightarrow \tau \\
\llbracket \tau_1 \xrightarrow{n} \tau_2 \rrbracket &: \tau_1^{n+1} \rightarrow \tau_2
\end{aligned}$$

Note that the semantics requires one more value than is the number of past values. This is because the first value is the current value and has to be always available, even when the annotation is zero as in *(var)*.

The rules defining the semantics are shown in Figure 17. The semantics of the context is a *list of products*. To make the rules easier to follow, we write  $\langle v_1, \dots, v_n \rangle$  for an  $n$ -element list containing products. Products that model the entire context such as  $v_1$  are written in bold. When we access individual variables, we write  $v = (x_1, \dots, x_m)$  where  $x_i$  denote individual variables of the context.

In *(var)*, the context is a singleton-list containing a product of variables, from which we project the right one. In *(prev)* and *(sub)*, we drop some of the elements from the history (from the front and end, respectively) and then evaluate the original expression.

Lambda abstractions *(abs)* receives two lists of the same size – one containing values of the variables (list of products) from the declaration site

$\langle v_0, \dots, v_n \rangle$  and one containing the argument (list of values) provided by the call site  $\langle y_0, \dots, y_n \rangle$ . The semantics applies the well-known *zip* operation on the lists and passes the result to the body.

Finally, application (*app*) uses the input context in two ways, which gives rise to the two requirements combined using *max*. First, it evaluates the expression  $e_1$  which is called with the past  $m$  values. The resulting function  $g$  is then sequentially composed with the semantics of  $e_2$ . To call the function, we need to evaluate  $e_2$  repeatedly – namely,  $p + 1$  times, which results in the overall requirement for  $n + p$  past values.

**SUMMARY.** Type systems have been used in the context of data-flow languages, for example to check initialization properties [20], but to our knowledge, not for checking the maximal number of required past values. Thus this section serves not just as an example, but also shows how coeffects can lead to novel results.

The most interesting point about the data-flow system is that it is remarkably similar to our earlier liveness example. In the type system, abstraction (*abs*) duplicates the context requirements and application (*abs*) arises from sequential and pointwise composition. We capture this striking similarity in Chapter 4. Before doing that, we look at one more example and then explore the *structural* class of systems.

### 3.2.5 Permissions and safe locking

In the implicit parameters and data-flow examples, the context provides additional resources or values that may be accessed at runtime. However, coeffects can also track *permissions* or *capabilities* to perform some operation. We can invert the intuition behind liveness and use it as a trivial example – when the context is live, it contains a *permission* to access variables. In this section, we briefly consider a system for safe locking of Flanagan and Abadi [29] as one, more advanced example. Calculus of capabilities of Cray et al. [23] is discussed later in Section 3.4.

**SAFE LOCKING.** The system for safe locking prevents race conditions (by only allowing access to mutable state under a lock) and avoids deadlocks (by imposing strict partial order on locks). The following program uses a mutable state under a lock:

```
newlock l : ρ in
let state = refρ 10 in
sync l (!state)
```

The declaration `newlock` creates a lock  $l$  protecting memory region  $\rho$ . We can then allocate mutable variables in that memory region (second line). An access to one or more mutable variables is only allowed in scope that is protected by a lock. This is done using the `sync` keyword, which locks a lock and evaluates an expression in a context that contains permission to access memory region of the lock ( $\rho$  in the above example).

The type system for safe locking associates a list of acquired locks with the context. Interestingly, the original presentation of the system by Flanagan and Abadi [29] uses a coeffect-style judgements of a form  $\Gamma; p \vdash e : \tau$  where  $p$  is a list of accessible regions (protected by an acquired lock). Using our notation, the rule for `sync` looks as follows:

$$(sync) \frac{\Gamma @ \mathbf{p} \vdash e_1 : m \quad \Gamma @ \mathbf{p} \cup \{m\} \vdash e_2 : \tau}{\Gamma @ \mathbf{p} \vdash \mathbf{sync} \ e_1 \ e_2 : \tau}$$

The rule requires that  $e_1$  yields a value of a singleton type  $m$ . The type is added as an indicator of the locked region to the context  $\mathbf{p} \cup \{m\}$  which is then used to evaluate the expression  $e_2$ .

**SUMMARY.** Despite attaching annotations to the variable context, the system for safe locking uses effect-style lambda abstraction. Lambda abstraction associates all requirements with the call site – a lambda function created under a lock cannot access protected memory available at the time of creation. It will be executed later and can only access the memory available then. This suggests that safe locking is better seen as an effect system.

Another interesting aspect is the extension to avoid deadlocks. In that case, the type system needs to reject programs that acquire locks in an invalid order. One way to model this is to replace  $\mathbf{p} \cup \{m\}$  with a *partial* operation  $\mathbf{p} \uplus \{m\}$  which is only defined when the lock  $m$  can be added to the set  $\mathbf{p}$ . Supporting partial operations on coeffect annotations is an interesting future extension for coeffect systems.

### 3.3 STRUCTURAL COEFFECT SYSTEMS

In structural coeffect systems, the additional information is associated with individual variables. This is very often information about how the variables are used, or, in which contexts they are used. In Chapter 1, we introduced the idea using an example that tracks array access patterns. Each variable is annotated with a range specifying which elements of the corresponding array may be accessed.

In this section, we look at three examples in detail – we revisit liveness and show a practically useful structural version of the system; we consider an example inspired by linear logic; finally, we revisit data-flow to get a more precise analysis. Although quite different, the common pattern among these three examples is somewhat easier to see, because they all track information about variable usage. We finish the section with a brief outline of several other applications.

#### 3.3.1 Liveness analysis revisited

The flat system for liveness analysis presented in Section 3.2.3 is interesting from a theoretical perspective, but it is not practically useful. Here, we revisit the problem and define a structural system that tracks liveness per-variable.

**STRUCTURAL LIVENESS.** Recall two examples discussed earlier where the flat liveness analysis marked the whole context as (syntactically) live, despite the fact part of it was (semantically) dead:

```
let constant = λy → λx → y
let answer = (λx → x) 42
```

In the first case, the variable  $x$  is dead, but was marked as live. In the second example, the declaration site of the `answer` value is dead, but was marked as live. This is because in both of the expressions, *some* variable is accessed. However, the *(abs)* rule of flat liveness has no way of determining *which*



variables are used by the body – and, in particular, whether the accessed variable is the *bound* variable or some of the *free* variables.

As discussed earlier, we can resolve this by attaching a *vector* of liveness annotations to a *vector* of variables. In the first example, the available variables are  $y$  and  $x$ , so the variable context  $\Gamma$  is a vector  $\langle y:\tau, x:\tau \rangle$ . Only the variable  $y$  is used and so the annotated context is:  $y:\tau, x:\tau @ \langle L, D \rangle$ . When writing the contexts, we omit angle brackets around variables, but it should still be viewed as a vector. There are two important points:

- The fact that variables are now a vector means that we cannot freely reorder them. This guarantees that  $x:\tau, y:\tau @ \langle L, D \rangle$  can not be confused with  $y:\tau, x:\tau @ \langle L, D \rangle$ . We need to define the type system in a way that is similar to sub-structural systems (discussed in Section 2.3) and add explicit rules for manipulating the context.
- We choose to attach a vector of annotations to a vector of variables, rather than attaching individual annotations to individual variables. This lets us unify and combine flat and structural systems as discussed in Section 6.1, but the alternative is briefly explored in Section 6.3.

**TYPE SYSTEM.** The structural system for liveness uses the same two-point lattice of annotations  $\mathcal{L} = \{L, D\}$  that was used by the flat system. We also use the  $\sqcup, \sqcap$  and  $\sqsubseteq$  operators that are defined in Figure 13.

The rules of the system are split into two groups. Figure 18 (a) shows the standard syntax-driven rules plus sub-coeffecting. In *(var)*, the context contains just the single accessed variable, which is annotated as live. Unused variables can be introduced using weakening. A constant (*const*) is accessed in an empty context, which also carries no annotations. The sub-coeffecting rule (*sub*) uses a pointwise extension of the  $\sqsubseteq$  relation over two vectors as defined in Section 3.1.3.

In the (*abs*) rule, the variable context of the body  $\Gamma, x:\tau_1$  is annotated with a vector  $\mathbf{r} \times \langle \mathbf{s} \rangle$ , where the vector  $\mathbf{r}$  corresponds to  $\Gamma$  and the singleton annotation  $\mathbf{s}$  corresponds to the variable  $x$ . Thus, the function is annotated with  $\mathbf{s}$ . Note that the free-variable context is annotated with vectors, but functions take only a single input and so are annotated with primitive annotations.

The (*app*) rule is similar to function applications in flat systems, but there is an important difference. In structural systems, the two sub-expressions have separate variable contexts  $\Gamma_1$  and  $\Gamma_2$ . Therefore, the composed expression just concatenates the variables and their corresponding annotations. (We can still use the same variable in both sub-expressions thanks to the structural contraction rule.)

The context  $\Gamma_1$  is used to evaluate  $e_1$  and is thus annotated with  $\mathbf{r}$ . The annotation for  $\Gamma_2$  is more interesting. It is a result of sequential composition of two semantic functions – the first one takes the (multi-variable) context  $\Gamma_2$  and evaluates  $e_2$ ; the second takes the result of type  $\tau_1$  and passes it to the function  $\tau_1 \xrightarrow{t} \tau_2$ . The composition is defined as follows:

$$g : \tau_1 \times \dots \times \tau_n \xrightarrow{\mathbf{s}} \sigma \quad f : \sigma \xrightarrow{t} \tau \quad f \circ g : \tau_1 \times \dots \times \tau_n \xrightarrow{t \sqcup \mathbf{s}} \tau$$

This definition is only for illustration and is revised in Chapter 5. The function  $g$  takes a product of multiple variables (and is annotated with a vector). The function  $f$  takes just a single value and is annotated with the scalar. As in the flat system, sequential composition is modelled using  $\sqcup$  – but here, we use a scalar-vector extension of the operation. Finally, the (*let*) rule follows similar reasoning (and also corresponds to the typing of  $(\lambda x. e_2) e_1$ ).



a.) Ordinary, syntax-driven rules along with sub-coeffecting

$$\begin{aligned}
(\text{var}) \quad & \frac{}{x : \tau @ \langle L \rangle \vdash x : \tau} \\
(\text{const}) \quad & \frac{c : \tau \in \Delta}{() @ \langle \rangle \vdash c : \tau} \\
(\text{abs}) \quad & \frac{\Gamma, x : \tau_1 @ \mathbf{r} \times \langle s \rangle \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{s} \tau_2} \\
(\text{app}) \quad & \frac{\Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \quad \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma_1, \Gamma_2 @ \mathbf{r} \times (\mathbf{t} \sqcup \mathbf{s}) \vdash e_1 e_2 : \tau_2} \\
(\text{let}) \quad & \frac{\Gamma_1, x : \tau_1 @ \mathbf{r} \times \langle t \rangle \vdash e_1 : \tau_2 \quad \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma_1, \Gamma_2 @ \mathbf{r} \times (\mathbf{t} \sqcup \mathbf{s}) \vdash \text{let } x = e_2 \text{ in } e_1 : \tau_2} \\
(\text{sub}) \quad & \frac{\Gamma @ \mathbf{r} \vdash e : \tau}{\Gamma @ \mathbf{r}' \vdash e : \tau} \quad \mathbf{r} \sqsubseteq \mathbf{r}'
\end{aligned}$$

b.) Structural rules for context manipulation

$$\begin{aligned}
(\text{weak}) \quad & \frac{\Gamma @ \mathbf{r} \vdash e : \sigma}{\Gamma, x : \tau @ \mathbf{r} \times \langle D \rangle \vdash e : \sigma} \\
(\text{exch}) \quad & \frac{\Gamma_1, x : \tau', y : \tau, \Gamma_2 @ \mathbf{r} \times \langle s, t \rangle \times \mathbf{q} \vdash e : \sigma}{\Gamma_1, y : \tau, x : \tau', \Gamma_2 @ \mathbf{r} \times \langle t, s \rangle \times \mathbf{q} \vdash e : \sigma} \quad \begin{array}{l} \text{len}(\Gamma_1) = \text{len}(\mathbf{r}) \\ \text{len}(\Gamma_2) = \text{len}(\mathbf{s}) \end{array} \\
(\text{contr}) \quad & \frac{\Gamma_1, y : \tau, z : \tau, \Gamma_2 @ \mathbf{r} \times \langle s, t \rangle \times \mathbf{q} \vdash e : \sigma}{\Gamma_1, x : \tau, \Gamma_2 @ \mathbf{r} \times \langle s \sqcap t \rangle \times \mathbf{q} \vdash e[z \leftarrow x][y \leftarrow x] : \sigma} \quad \begin{array}{l} \text{len}(\Gamma_1) = \text{len}(\mathbf{r}) \\ \text{len}(\Gamma_2) = \text{len}(\mathbf{s}) \end{array}
\end{aligned}$$

Figure 18: Structural coeffect liveness analysis

**STRUCTURAL TYPING RULES.** The structural typing rules are shown in Figure 18 (b). They mirror the rules known from sub-structural type systems (Section 2.3). Weakening (*weak*) extends the context with a single unused variable  $x$  and adds the  $D$  annotation to the vector of coeffects.

The variable is always added to the end as in the (*abs*) rule. However, the exchange rule (*exch*) lets us arbitrarily reorder variables. It flips the variables  $x$  and  $x'$  and their corresponding coeffect annotations in the vector. This is done by requiring that the lengths of the remaining, unchanged, parts of the vectors match.

Finally, contraction (*contr*) makes it possible to use a single variable multiple times. Given a judgement that contains variables  $y$  and  $z$ , we can derive a judgement for an expression where both  $z$  and  $y$  are replaced by a single variable  $x$ . Their annotations  $s, t$  are combined into  $s \sqcap t$ , which means that  $x$  is live if either  $z$  or  $y$  were live in the original expression.

**EXAMPLE.** To demonstrate how the system works, we consider the expression  $(\lambda x \rightarrow v) y$ . This is similar to an example where flat liveness mistakenly marks the entire context as live. Despite the fact that the variable  $y$  is accessed (syntactically), it is not live – because the function that takes it as an argument always returns  $v$ .

The typing derivation for the body uses  $(var)$  and  $(abs)$ . However, we also need  $(weak)$  to add the unused variable  $x$  to the context:

$$(weak) \frac{\frac{\frac{}{v:\tau @ \langle L \rangle \vdash v:\tau} (var)}{v:\tau, x:\tau @ \langle L, D \rangle \vdash v:\tau} (abs)}{v:\tau @ \langle L \rangle \vdash (\lambda x \rightarrow v) : \tau \xrightarrow{D} \tau}$$

The interesting part is the use of the  $(app)$  rule in the next step. Although the variable  $y$  is live in the expression  $y$ , it is marked as dead in the overall expression, because the function is annotated with  $D$ :

$$(app) \frac{\frac{v:\tau @ \langle L \rangle \vdash (\lambda x \rightarrow v) : \tau \xrightarrow{D} \tau \quad \frac{}{y:\tau @ \langle L \rangle \vdash y:\tau} (var)}{v:\tau, y:\tau @ \langle L \rangle \times (D \sqcup \langle L \rangle) \vdash (\lambda x \rightarrow v) y : \tau}}{v:\tau, y:\tau @ \langle L, D \rangle \vdash (\lambda x \rightarrow v) y : \tau}$$

The application is written in two steps – the first one directly applies the  $(app)$  rule and the second one simplifies the coefficient annotation. The key part is the use of the scalar-vector operator  $D \sqcup \langle L \rangle$ . Using the definition of the scalar-vector extension, this equals  $\langle D \sqcup L \rangle$  which is  $\langle D \rangle$ .

**SEMANTICS.** When defining the semantics of flat liveness calculus, we used an indexed form of the option type  $1 + \tau$  (which is  $1$  for dead contexts and  $\tau$  for live contexts). In the semantics of expressions, the type constructor was applied to the entire context, i.e.  $1 + (\tau_1 \times \dots \times \tau_n)$ . In the structural version, the semantics applies the option type constructor to individual elements of the free-variable context pair:  $(1 + \tau_1) \times \dots \times (1 + \tau_n)$ . For each variable, the type is indexed by the corresponding annotation:

$$\llbracket x_1:\tau_1, \dots, x_n:\tau_n @ \langle r_1, \dots, r_n \rangle \vdash e : \tau \rrbracket : (\tau'_1 \times \dots \times \tau'_n) \rightarrow \tau$$

$$\text{where } \tau'_i = \begin{cases} \tau_i & (r_i = L) \\ 1 & (r_i = D) \end{cases}$$

Note that the product of the free variables is not an ordinary tuple of our language, but a special construction (we return to this topic in Section 5.3). This follows from the asymmetry of  $\lambda$ -calculus, as discussed in Section 3.1.3. Functions take just a single input and so they are interpreted in the same way as in flat calculus:

$$\llbracket \tau_1 \xrightarrow{L} \tau_2 \rrbracket = \tau_1 \rightarrow \tau_2 \quad \llbracket \tau_1 \xrightarrow{D} \tau_2 \rrbracket = 1 \rightarrow \tau_2$$

The rules that define the semantics are shown in Figure 19. To make the definition simpler, we are somewhat vague when working with products. We write variables of product type such as  $\mathbf{v}$  in bold-face and individual values like  $x$  in normal face. We freely re-associate products and so  $(\mathbf{v}, x)$  should not be seen as a nested product, but simply as a product containing all variables from the product  $\mathbf{v}$  together with one additional variable  $x$  at the end. We shall be more precise in Chapter 5.

In  $(var)$ , the context contains just a single variable and so we do not even need to apply projection;  $(cosnt)$  receives no variables and uses global constant lookup function  $\delta$ . In  $(abs)$ , we obtain two parts of the context and combine them into  $(\mathbf{v}, x)$ . This works the same way regardless of whether the variables are live or dead. For simplicity, we omit sub-coeffecting, which just turns some of the available values  $v_i$  to unit values  $()$ .

As dictated by the semantics, the application again needs to “implement” dead code elimination (otherwise the type system would be unsound). When

a.) Semantics of ordinary expressions

$$\llbracket x : \tau @ \langle L \rangle \vdash x : \tau \rrbracket = \lambda(x) \rightarrow x \quad (var)$$

$$\llbracket () @ \langle \rangle \vdash c : \tau \rrbracket = \lambda() \rightarrow \delta(c) \quad (const)$$

$$\begin{aligned} \llbracket \Gamma @ \mathbf{r} \vdash \lambda y. e : \tau_1 \xrightarrow{s} \tau_2 \rrbracket &= \lambda \mathbf{v} \rightarrow \\ \lambda y \rightarrow \llbracket \Gamma, y : \tau_1 @ \mathbf{r} \times \langle s \rangle \vdash e : \tau_2 \rrbracket (\mathbf{v}, y) \end{aligned} \quad (abs)$$

$$\begin{aligned} \llbracket \Gamma_1, \Gamma_2 @ \mathbf{r} \times (\mathbf{L} \sqcup \mathbf{s}) \vdash e_1 e_2 : \tau_2 \rrbracket &= \lambda(\mathbf{v}_1, \mathbf{v}_2) \rightarrow \\ \text{let } g &= \llbracket \Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{L} \tau_2 \rrbracket \mathbf{v}_1 \\ \text{in } g &(\llbracket \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1 \rrbracket \mathbf{v}_2) \end{aligned} \quad (app-1)$$

$$\begin{aligned} \llbracket \Gamma_1, \Gamma_2 @ \mathbf{r} \times (\mathbf{D} \sqcup \mathbf{s}) \vdash e_1 e_2 : \tau_2 \rrbracket &= \lambda(\mathbf{v}_1, \mathbf{v}_2) \rightarrow \\ \text{let } g &= \llbracket \Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{D} \tau_2 \rrbracket \mathbf{v}_1 \\ \text{in } g &() \end{aligned} \quad (app-2)$$

b.) Semantics of structural context manipulation

$$\llbracket \Gamma, x : \tau @ \mathbf{r} \times \langle D \rangle \vdash e : \sigma \rrbracket = \lambda(\mathbf{v}, ()) \rightarrow \llbracket \Gamma @ \mathbf{r} \vdash e : \sigma \rrbracket \mathbf{v} \quad (weak)$$

$$\begin{aligned} \llbracket \Gamma_1, y : \tau, x : \tau', \Gamma_2 @ \mathbf{r} \times \langle t, s \rangle \times \mathbf{q} \vdash e : \sigma \rrbracket &= \lambda(\mathbf{v}_1, y, x, \mathbf{v}_2) \rightarrow \\ \llbracket \Gamma_1, x : \tau', y : \tau, \Gamma_2 @ \mathbf{r} \times \langle s, t \rangle \times \mathbf{q} \vdash e : \sigma \rrbracket &(\mathbf{v}_1, x, y, \mathbf{v}_2) \end{aligned} \quad (exch)$$

$$\begin{aligned} \llbracket \Gamma_1, x : \tau, \Gamma_2 @ \mathbf{r} \times \langle D \rangle \times \mathbf{q} \vdash e[z \leftarrow x][y \leftarrow x] : \sigma \rrbracket &= \lambda(\mathbf{v}_1, (), \mathbf{v}_2) \rightarrow \\ \llbracket \Gamma_1, y : \tau, z : \tau, \Gamma_2 @ \mathbf{r} \times \langle D, D \rangle \times \mathbf{q} \vdash e : \sigma \rrbracket &(\mathbf{v}_1, (), (), \mathbf{v}_2) \end{aligned} \quad (contr-1)$$

$$\begin{aligned} \llbracket \Gamma_1, x : \tau, \Gamma_2 @ \mathbf{r} \times \langle L \rangle \times \mathbf{q} \vdash e[z \leftarrow x][y \leftarrow x] : \sigma \rrbracket &= \lambda(\mathbf{v}_1, x, \mathbf{v}_2) \rightarrow \\ \left\{ \begin{array}{l} \llbracket \Gamma_1, y : \tau, z : \tau, \Gamma_2 @ \mathbf{r} \times \langle L, L \rangle \times \mathbf{q} \vdash e : \sigma \rrbracket (\mathbf{v}_1, x, x, \mathbf{v}_2) \\ \llbracket \Gamma_1, y : \tau, z : \tau, \Gamma_2 @ \mathbf{r} \times \langle D, L \rangle \times \mathbf{q} \vdash e : \sigma \rrbracket (\mathbf{v}_1, (), x, \mathbf{v}_2) \\ \llbracket \Gamma_1, y : \tau, z : \tau, \Gamma_2 @ \mathbf{r} \times \langle L, D \rangle \times \mathbf{q} \vdash e : \sigma \rrbracket (\mathbf{v}_1, x, (), \mathbf{v}_2) \end{array} \right. & \quad (contr-2) \end{aligned}$$

Figure 19: Semantics of structural liveness

the input parameter of the function  $g$  is live (*app-1*), we first evaluate  $e_2$  and then pass the result to  $g$ . When the parameter is dead (*app-2*), we do not need to evaluate  $e_2$  and so all values in  $\mathbf{v}_2$  can be dead, i.e.  $()$ .

In the structural rules, (*weak*) receives context containing a dead variable as the last one. It drops the  $()$  value and evaluates the expression in a context  $\mathbf{v}$ . Exchange (*exch*) simply swaps two variables. In contraction, we either duplicate a dead value (*contr-1*), or a live value (*contr-2*). In the latter, one of the duplicates may be dead and so we need to consider three separate cases.

**SUMMARY.** The structural liveness calculus is a typical example of a system that tracks per-variable annotations. In a number of ways, the system is simpler than the flat coeffect calculi. In lambda abstraction, we simply annotate function with the annotation of a matching variable (this rule is the same for all upcoming systems). In application, the *pointwise* composition is no longer needed, because the sub-expressions use separate contexts. On the other hand, we had to add weakening, contraction and exchange rules to let us manipulate contexts.

The semantics of weakening demonstrates an important point about coeffects that may be quite confusing. When we read the *typing rule* from top to bottom, weakening adds a variable to the context. When we read the *semantic rule*, weakening drops a variable value from the context! This duality is caused by the fact that coeffects talk about context – they describe how

to build the context required by the sub-expressions and so the semantics implements transformation from the context in the (typing) conclusion to the (typing) assumption. How should coeffects be understood, in general, is discussed further in Section 4.2.2.

The structural systems discussed in the upcoming sections are remarkably similar to the one shown here. We discuss two more examples to explore the design space, but we shall omit details that are shared with the system in this section.

### 3.3.2 Bounded variable use

Liveness analysis checks whether a variable is used or unused. With structural coeffects, we can go further and track how many times is the variable accessed. Girard et al. [35] coined this idea as *bounded linear logic* and use it to restrict well-typed programs to polynomial-time algorithms. We first introduce the system in our, coeffect, style and then relate it with the original formulation.

**BOUNDED VARIABLE USE.** The system discussed in this section tracks the number of times a variable is accessed in the call-by-name evaluation. Although we look at an example that tracks *variable usage*, the same system could be used to track access to resources that are always passed as a reference (and behave effectively as call-by-name) and so the system is relevant for call-by-value languages too. To demonstrate the idea, consider the following term:

$$(\lambda v. x + v + v) (x + y)$$

When evaluated, the body of the function directly accesses  $x$  once and then twice indirectly, via the function argument. Similarly,  $y$  is accessed twice indirectly. Thus, the overall expression uses  $x$  three times and  $y$  twice.

As discussed in Chapter 5, the system preserves type and coeffect annotations under the  $\beta$ -reduction. Reducing the expression in this case gives  $x + (x + y) + (x + y)$ . This has the same bounds as the original expression –  $x$  is used three times and  $y$  twice.

**TYPE SYSTEM.** The type system in Figure 20 annotates contexts with vectors of integers. The rules have the same structure as those of the system for liveness analysis. The only difference is how annotations are combined. Here, we use integer multiplication ( $*$ ) for sequential composition and addition ( $+$ ) for point-wise composition.

Variable access (*var*) annotates a variable with 1, meaning that it has been used once. An unused variable (*weak*) is annotated with 0. Multiple occurrences of the same variable are introduced by contraction (*contr*), which adds the numbers of the two contracted variables.

As previously, application (*app*) and let binding (*let*) combine two separate contexts. The second part applies a function that uses its parameter  $t$ -times to an argument that uses variables in  $\Gamma_2$  at most  $s$ -times (here,  $s$  is a vector of integers with an annotations for each variable in  $\Gamma_2$ ). The sequential composition (modelling call-by-name) multiplies the uses, meaning that the total number of uses is  $(t * s)$  (where  $*$  is a point-wise multiplication of a vector by a scalar). This models the fact that for each use of the function parameter, we replicate the variable uses in  $e_2$ .

a.) Ordinary, syntax-driven rules along with sub-coeffecting

$$\begin{aligned}
(var) & \frac{}{x : \tau @ \langle 1 \rangle \vdash x : \tau} \\
(abs) & \frac{\Gamma, x : \tau_1 @ \mathbf{r} \times \langle \mathbf{s} \rangle \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{s}} \tau_2} \\
(app) & \frac{\Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{\mathbf{t}} \tau_2 \quad \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma_1, \Gamma_2 @ \mathbf{r} \times (\mathbf{t} * \mathbf{s}) \vdash e_1 e_2 : \tau_2} \\
(let) & \frac{\Gamma_1, x : \tau_1 @ \mathbf{r} \times \langle \mathbf{t} \rangle \vdash e_1 : \tau_2 \quad \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma_1, \Gamma_2 @ \mathbf{r} \times (\mathbf{t} * \mathbf{s}) \vdash \text{let } x = e_2 \text{ in } e_1 : \tau_2} \\
(sub) & \frac{\Gamma @ \mathbf{r} \vdash e : \tau}{\Gamma @ \mathbf{r}' \vdash e : \tau} \quad \mathbf{r} \leq \mathbf{r}'
\end{aligned}$$

b.) Structural rules for context manipulation

$$\begin{aligned}
(weak) & \frac{\Gamma @ \mathbf{r} \vdash e : \sigma}{\Gamma, x : \tau @ \mathbf{r} \times \langle 0 \rangle \vdash e : \sigma} \\
(exch) & \frac{\Gamma_1, x : \tau', y : \tau, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}, \mathbf{t} \rangle \times \mathbf{q} \vdash e : \sigma}{\Gamma_1, y : \tau, x : \tau', \Gamma_2 @ \mathbf{r} \times \langle \mathbf{t}, \mathbf{s} \rangle \times \mathbf{q} \vdash e : \sigma} \quad \begin{array}{l} \text{len}(\Gamma_1) = \text{len}(\mathbf{r}) \\ \text{len}(\Gamma_2) = \text{len}(\mathbf{s}) \end{array} \\
(contr) & \frac{\Gamma_1, y : \tau, z : \tau, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}, \mathbf{t} \rangle \times \mathbf{q} \vdash e : \sigma}{\Gamma_1, x : \tau, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} + \mathbf{t} \rangle \times \mathbf{q} \vdash e[z \leftarrow x][y \leftarrow x] : \sigma} \quad \begin{array}{l} \text{len}(\Gamma_1) = \text{len}(\mathbf{r}) \\ \text{len}(\Gamma_2) = \text{len}(\mathbf{s}) \end{array}
\end{aligned}$$

Figure 20: Structural coeffect bounded reuse analysis

Finally, the sub-coeffecting rule (*sub*) safely overapproximates the number of accesses using the pointwise  $\leq$  relation. We can view any variable as being used a greater number of times than it actually is.

EXAMPLE. To type check the expression  $(\lambda v. x + v + v) (x + y)$  discussed earlier, we need to use abstraction, application, but also the contraction rule. Assuming the type judgement for the body, abstractions yields:

$$(abs) \frac{x : \mathbb{Z}, v : \mathbb{Z} @ \langle 1, 2 \rangle \vdash x + v + v : \mathbb{Z}}{x : \mathbb{Z} @ \langle 1 \rangle \vdash (\lambda v. x + v + v) : \mathbb{Z} \xrightarrow{2} \mathbb{Z}}$$

To type-check the application, the contexts of  $e_1$  and  $e_2$  need to contain disjoint variables. For this reason, we  $\alpha$ -rename  $x$  to  $x'$  in the argument  $(x + y)$  and later join  $x$  and  $x'$  using the contraction rule. Assuming  $(x' + y)$  is checked in a context that marks  $x'$  and  $y$  as used once, the application rule yields a judgement that is simplified as follows:

$$(contr) \frac{\frac{x : \mathbb{Z}, x' : \mathbb{Z}, y : \mathbb{Z} @ \langle 1 \rangle \times (2 * \langle 1, 1 \rangle) \vdash (\lambda v. x + v + v) (x' + y) : \mathbb{Z}}{x : \mathbb{Z}, x' : \mathbb{Z}, y : \mathbb{Z} @ \langle 1, 2, 2 \rangle \vdash (\lambda v. x + v + v) (x' + y) : \mathbb{Z}}}{x : \mathbb{Z}, y : \mathbb{Z} @ \langle 3, 2 \rangle \vdash (\lambda v. x + v + v) (x + y) : \mathbb{Z}}$$

The first step performs scalar multiplication, producing the vector  $\langle 1, 2, 2 \rangle$ . In the second step, we use contraction to join variables  $x$  and  $x'$  from the function and argument terms respectively.

SEMANTICS. In the previous examples, we defined the semantics – somewhat informally – using a simple  $\lambda$ -calculus language to encode the model. More formally, this could be a Cartesian-closed category. In that model, we

can reuse variables arbitrarily and so it is not a good fit for modelling bounded reuse. Girard et al. [35] model their bounded linear logic in an (ordinary) linear logic where variables can be used at most once.

Following the same approach, we could model a variable  $\tau$ , annotated with  $r$  as a product containing  $r$  copies of  $\tau$ , that is  $\tau^r$ :

$$\llbracket x_1:\tau_1, \dots, x_n:\tau_n @ \langle r_1, \dots, r_n \rangle \vdash e:\tau \rrbracket : (\tau_1^{r_1} \times \dots \times \tau_n^{r_n}) \rightarrow \tau$$

where  $\tau_i^{r_i} = \underbrace{\tau_i \times \dots \times \tau_i}_{r_i\text{-times}}$

The functions are interpreted similarly. A function  $\tau_1 \xrightarrow{t} \tau_2$  is modelled as a function taking  $t$ -element product of  $\tau_1$  values:  $\tau_1^t \rightarrow \tau_2$ .

The rules that define the semantics of bounded calculus are mostly the same as (or easy to adapt from) the semantic rules of liveness in Figure 19. The ones that differ are those that use sequential composition (application and let binding) and the contraction rule, which represents pointwise composition.

In the following, we use vector names  $\mathbf{v}_i$  for contexts containing multiple variables i.e. have a type  $\tau_1^{r_1} \times \dots \times \tau_m^{r_m}$ . Each vector contains multiple copies of each variable, to model the fact that variables are used in an affine way (at most once). We do not explicitly write the sizes of these vectors (number of variables in a context; number of instances of a variable) as these are clear from the coefficient annotations. We assume that  $\Gamma_2$  contains  $n$  variables and that  $\mathbf{s} = \langle s_1, \dots, s_n \rangle$ :

$$\begin{aligned} \llbracket \Gamma_1, x:\tau, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} + \mathbf{t} \rangle \times \mathbf{q} \vdash e[z \leftarrow x][y \leftarrow x] : \sigma \rrbracket = \\ \lambda(\mathbf{v}_1, (x_1, \dots, x_{s+t}), \mathbf{v}_2) \rightarrow \\ \llbracket \Gamma_1, y:\tau, z:\tau, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}, \mathbf{t} \rangle \times \mathbf{q} \vdash e : \sigma \rrbracket \\ (\mathbf{v}_1, (x_1, \dots, x_s), (x_{s+1}, \dots, x_{s+t}), \mathbf{v}_2) \end{aligned} \quad (contr)$$

$$\begin{aligned} \llbracket \Gamma_1, \Gamma_2 @ \mathbf{r} \times (\mathbf{t} * \mathbf{s}) \vdash e_1 e_2 : \tau_2 \rrbracket = \\ \lambda(\mathbf{v}_1, ((x_{1,1}, \dots, x_{1,t*s_1}), \dots, (x_{n,1}, \dots, x_{n,t*s_n}))) \rightarrow \\ \text{let } g = \llbracket \Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \rrbracket \mathbf{v}_1 \\ \text{let } \mathbf{y}_1 = ((x_{1,1}, \dots, x_{1,s_1}), \dots, (x_{n,1}, \dots, x_{n,s_n})) \\ \text{let } \dots \\ \text{let } \mathbf{y}_t = ((x_{1,(t-1)*s_1+1}, \dots, x_{1,t*s_1}), \dots, \\ (x_{n,(t-1)*s_n+1}, \dots, x_{n,t*s_n})) \\ \text{in } g(\llbracket \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1 \rrbracket \mathbf{y}_1, \dots, \llbracket \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1 \rrbracket \mathbf{y}_t) \end{aligned} \quad (app)$$

In the *(contr)* rule, the semantic function is called with  $\mathbf{s} + \mathbf{t}$  copies of a value for the  $x$  variable. The values are split between  $\mathbf{s}$  and  $\mathbf{t}$  separate copies of variables  $y$  and  $z$ , respectively.

The *(app)* rule is similar in that it needs to split the input variable context. However, it needs to split values of multiple variables – in  $x_{i,j}$ , the index  $i$  stands for an index of the variable while  $j$  is an index of one of multiple copies of the value. In the semantic function, the second part of the context consists of  $n$  variables where the multiplicity of each value is specified by the annotation  $s_i$  multiplied by  $t$ . The rule needs to evaluate the argument  $e_2$   $t$ -times and each call requires  $s_i$  copies of the  $i^{\text{th}}$  variable. To do this, we create contexts  $\mathbf{y}_1$  to  $\mathbf{y}_t$ , each containing  $s_i$  copies of the variable (and so we require  $s_i * t$  copies of each variable). Note that the contexts are created such that each value is used exactly once.

It is worth noting that the *(var)* rule requires exactly one copy of a variable and so the system tracks precisely the number of uses. However, the *(sub)* rule lets us ignore additional copies of a value. Thus, permitting *(sub)* rule is only possible if the underlying model is *affine* rather than *linear*.

**BOUNDED LINEAR LOGIC.** The system presented in this section is based on the idea of bounded linear logic (BLL) [35], but it is adapted to follow the structure of other coeffect systems discussed in this chapter. This elucidates the connection between BLL and coeffects.

The big difference, using the terminology from Section 2.2.3, is that our system is written in *language semantics* style, while BLL is written in *meta-language* style. We briefly consider the original BLL formulation.

The terms and types of our system are the terms and types of an ordinary  $\lambda$ -calculus, with the only difference that functions carry coeffect annotations. In BLL, the language of types is extended with a type constructor  $!_k A$  (where  $A$  is a proposition, corresponding to a type  $\tau$  in our system). The type denotes a value  $A$  that can be used at most  $k$  times.

As a result, BLL does not need to attach additional annotation to the variable context as a whole. The requirements are attached to individual variables and so our context  $\tau_1, \dots, \tau_n @ \langle k_1, \dots, k_n \rangle$  corresponds to a BLL assumption  $!_{k_1} A_1, \dots, !_{k_n} A_n$ . Using the formulation of bounded logic (and omitting the terms), the weakening and contraction rules are written as follows:

$$\begin{array}{c} \text{(weak)} \quad \frac{\Gamma \vdash B}{\Gamma, !_0 A \vdash B} \qquad \text{(contr)} \quad \frac{\Gamma, !_n A, !_m A \vdash B}{\Gamma, !_{n+m} A \vdash B} \end{array}$$

The system captures the same idea as the structural coeffect system presented above. Variable access in bounded linear logic is simply an operation that produces a value  $!_n A$  and so the system further introduces *dereliction* rule which lets us treat  $!_1 A$  as a value  $A$ . We further explore difference between *language semantics* and *meta-language* in Section 6.3.

**SUMMARY.** Comparing the structural coeffect calculus for tracking liveness and for bounded variable reuse reveals which parts of the systems differ and which parts are shared. In particular, both systems use the same vector operations ( $\times$ ,  $\langle - \rangle$ ) and also share the lambda abstraction rule (*abs*). They differ in the primitive values used to annotate used and unused variables ( $L$ ,  $D$  and  $1$ ,  $0$ , respectively) and in the operators used for sequential composition and contraction ( $\sqcup$ ,  $\sqcap$  and  $*$ ,  $+$ , respectively). The algebraic structure capturing these operators is developed in Chapter 5.

The brief overview of bounded linear logic shows an alternative approach to tracking properties related to individual variables – we could attach annotations to the variables themselves rather than attaching a *vector* of annotations to the entire context. The main benefit of our approach is that it lets us unify flat and structural systems (Chapter 6).

### 3.3.3 Data-flow languages revisited

When discussing data-flow languages in an earlier section, we said that the context provides past values of variables. In Section 3.2.4, we tracked this as a *flat* property, which gives us a system that keeps the same number of past values for all variables. However, data-flow can also be adapted to a structural system which keeps the number of required past values individually for each variable. Consider the following example:



$$\begin{array}{c}
\text{(var)} \quad \frac{}{x : \tau @ \langle 0 \rangle \vdash x : \tau} \\
\text{(prev)} \quad \frac{\Gamma @ \mathbf{r} \vdash e : \tau}{\Gamma @ 1 + \mathbf{r} \vdash \text{prev } e : \tau} \\
\text{(app)} \quad \frac{\Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \quad \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma_1, \Gamma_2 @ \mathbf{r} \times (\mathbf{t} + \mathbf{s}) \vdash e_1 e_2 : \tau_2} \\
\text{(weak)} \quad \frac{\Gamma @ \mathbf{r} \vdash e : \sigma}{\Gamma, x : \tau @ \mathbf{r} \times \langle 0 \rangle \vdash e : \sigma} \\
\text{(contr)} \quad \frac{\Gamma_1, y : \tau, z : \tau, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}, \mathbf{t} \rangle \times \mathbf{q} \vdash e : \sigma}{\Gamma_1, x : \tau, \Gamma_2 @ \mathbf{r} \times \langle \max(\mathbf{s}, \mathbf{t}) \rangle \times \mathbf{q} \vdash e[z \leftarrow x][y \leftarrow x] : \sigma}
\end{array}$$

Figure 21: Structural coeffect bounded reuse analysis

`let` offsetAdd = left + `prev` right

The value offsetAdd adds values of left with previous values of right. To evaluate a current value of the stream, we need the current value of left and one past value of right. Flat system is not able to capture this level-of-detail and simply requires 1 past values of both streams in the variable context.

Turning a flat data-flow system to a structural data-flow system is a change similar to the one between flat and structural liveness. In case of liveness analysis, we included the flat system only as an illustration (it is not practically useful). For data-flow, the flat system is less precise, but still practically useful (simplicity may outweigh precision).

**TYPE SYSTEM.** The type system in Figure 21 annotates the variable context with a vector of integers. This is similar as in the bounded reuse system, but the integers *mean* a different thing. Consequently, they are also calculated differently. We omit rules that are the same for all structural coeffect systems (exchange, lambda abstraction).

In data-flow, we annotate both used variables (*var*) and unused variables (*weak*) with 0, meaning that no past values are required. This is the same as in flat data-flow, but different from bounded reuse and liveness (where difference between using and not using a variable matters). Primitive requirements are introduced by the (*prev*) rule, which increments the annotation of all variables in the context.

In flat data-flow, we identified sequential composition and pointwise composition as two primitive operations that were used in the (flat) application. In the structural system, these are used in (*app*) and (*contr*), respectively. Thus application combines coeffect annotations using + and contraction using *max*. This contrasts with bounded reuse, which uses \* and +, respectively.

**EXAMPLE.** As an example, consider a function  $\lambda x. \text{prev } (y + x)$  applied to an argument `prev (prev y)`. The body of the function accesses the past value of two variables, one free and one bound. The (*abs*) rule splits the annotations between the declaration site and call site of the function:

$$\text{(abs)} \quad \frac{y : \mathbb{Z}, x : \mathbb{Z} @ \langle 1, 1 \rangle \vdash \text{prev } (y + x) : \mathbb{Z}}{y : \mathbb{Z} @ \langle 1 \rangle \vdash \lambda x. \text{prev } (y + x) : \mathbb{Z} \xrightarrow{1} \mathbb{Z}}$$



The expression always requires the previous value of  $y$  and adds it to a previous value of the parameter  $x$ . Evaluating the value of the argument  $\text{prev}(\text{prev } y)$  requires two past values of  $y$  and so the overall requirement for the (free) variable  $y$  is 3 past values. In order to use the contraction rule, we rename  $y$  to  $y'$  in the argument:

$$\frac{\frac{y:\mathbb{Z}@\langle 1 \rangle \vdash \lambda x. (\dots) : \mathbb{Z} \xrightarrow{1} \mathbb{Z} \quad x:\mathbb{Z}@\langle 2 \rangle \vdash (\text{prev}(\text{prev } y')) : \mathbb{Z}}{y:\mathbb{Z}, y':\mathbb{Z}@\langle 1,3 \rangle \vdash (\lambda x. \text{prev}(y+x)) (\text{prev}(\text{prev } y')) : \mathbb{Z}}}{y:\mathbb{Z}@\langle 3 \rangle \vdash (\lambda x. \text{prev}(y+x)) (\text{prev}(\text{prev } y)) : \mathbb{Z}}$$

The derivation uses (*app*) to get requirements  $\langle 1,3 \rangle$  and then (*contr*) to take the maximum, showing three past values are sufficient.

Note that we get the same requirements when we perform  $\beta$  reduction of the expression. Substituting the argument for  $x$  yields the expression  $\text{prev}(y + (\text{prev}(\text{prev } y)))$ . Semantically, this performs stream lookups  $y[1]$  and  $y[3]$  where the indices are the number of enclosing  $\text{prev}$  constructs.

**SEMANTICS.** To define the semantics of our structural data-flow language, we can use the same approach as when adapting flat liveness to structural liveness. Rather than wrapping the whole context in a type constructor (list or option), we now wrap the individual components of the product representing the variables in the context.

The result is similar as the structure used for bounded reuse. The only difference is that, given a variable annotated with  $r$ , we need  $1 + r$  values. That is, we need the current value, followed by  $r$  past values:

$$\begin{aligned} \llbracket x_1:\tau_1, \dots, x_n:\tau_n @ \langle r_1, \dots, r_n \rangle \vdash e : \tau \rrbracket &: (\tau_1^{(r_1+1)} \times \dots \times \tau_n^{(r_n+1)}) \rightarrow \tau \\ \llbracket \tau_1 \xrightarrow{s} \tau_2 \rrbracket &= \tau_1^{(s+1)} \rightarrow \tau_2 \end{aligned}$$

Despite the similarity with the semantics for bounded reuse, the values here *represent* different things. Rather than providing multiple copies of a value (out of which each can be used just once), the pair provides past values (that can be reused and freely accessed). To illustrate the behaviour we consider the semantics of the  $\text{prev}$  construct and of the structural contraction rule:

$$\begin{aligned} \llbracket \Gamma @ \langle s_1 + 1, \dots, s_n + 1 \rangle \vdash \text{prev } e : \tau \rrbracket &= \\ \lambda((x_{1,0}, \dots, x_{1,s_1+1}), \dots, (x_{n,0}, \dots, x_{n,s_n+1})) &\rightarrow \llbracket \Gamma @ \langle s_1, \dots, s_n \rangle \vdash e : \tau \rrbracket \\ &((x_{1,0}, \dots, x_{1,s_1}), \dots, (x_{n,0}, \dots, x_{n,s_n})) \end{aligned} \quad (\text{prev})$$

$$\begin{aligned} \llbracket \Gamma_1, x:\tau, \Gamma_2 @ \mathbf{r} \times \langle \max(s, t) \rangle \times \mathbf{q} \vdash e[z \leftarrow x][y \leftarrow x] : \sigma \rrbracket &= \\ \lambda(\mathbf{v}_1, (x_0, x_1, \dots, x_{\max(s,t)}), \mathbf{v}_2) &\rightarrow \llbracket \Gamma_1, y:\tau, z:\tau, \Gamma_2 @ \mathbf{r} \times \langle s, t \rangle \times \mathbf{q} \vdash e : \sigma \rrbracket \\ &(\mathbf{v}_1, (x_0, \dots, x_s), (x_0, \dots, x_t), \mathbf{v}_2) \end{aligned} \quad (\text{contr})$$

In (*prev*), the semantic function is called with an argument that stores values of  $n$  variables, such that a variable  $x_i$  has values ranging from  $x_{i,0}$  to  $x_{i,s_i+1}$ . Thus, there is one current value, followed by  $s_i + 1$  past values. The expression  $e$  nested under  $\text{prev}$  requires only  $s_i$  past values and so the semantics simply drops the last value.

In the (*contr*) rule, the semantic function receives  $\max(s, t)$  values of a specific variable  $x$ . It needs to produce values for two separate variables,  $y$  and  $z$  that require  $s$  and  $t$  past values. Both of these numbers are certainly smaller than (or equal to) the number of values available. Thus we simply take the first values. Unlike in the contraction for BLL, the values are duplicated and the same values are used for both variables.

**SUMMARY.** Two of the structural examples shown so far (liveness and data-flow) extend an earlier flat version of a similar system. We discuss this relation in general later. However, a flat system can generally be turned into a structural one – although this only gives a useful system when the flat version captures statically scoped properties, i. e. related to variables.

The data-flow example demonstrates that the a flat system can also be turned into structural system. In general, this only works for systems where lambda abstraction duplicates context requirements (as in Figure 14).

### 3.3.4 Security, tainting and provenance

Tainting is a mechanism where variables coming from potentially untrusted sources are marked (*tainted*) and the use of such variables is disallowed in contexts where untrusted input can cause security issues or other problems. Tainting can be done dynamically using a runtime mark (e. g. in the Perl language) or using a static type system. Tainting can be viewed as a special case of *provenance tracking*, known from database systems [16], where values are annotated with more detailed information about their source.

Static typed systems based on tainting have been use to prevent cross-site scripting attacks [95] and SQL injection attacks [39, 38]. In the latter case, we want to check that SQL commands cannot be directly constructed from, potentially dangerous, inputs provided by the user. Consider the type checking of the following expression in a context containing variables `id` and `msg`:

```
let name = query("SELECT Name WHERE Id = %1", id)
msg + name
```

In this example, `id` must not come directly from a user input, because `query` requires untainted string. Otherwise, the attacker could specify values such as `"1; DROP TABLE Users"`. The variable `msg` may or may not be tainted, because it is not used in protected context (i.e. to construct an SQL query).

In runtime checking, all (string) values need to be wrapped in an object with a Boolean flag (for tainting) or more complex data (for provenance). In static checking, the information need to be associated with the variables in the variable context.

**CORE DEPENDENCY CALCULUS.** Taint checking is a special case of checking of the *non-interference* property in *secure information flow*. There, the aim is to guarantee that sensitive information (such as credit card number) cannot be leaked to contexts with low secrecy (e. g. sent via an unsecured network channel). Volpano et al. [96] provide the first (provably) sound type system that guarantees non-inference and Sabelfeld et al. [75] surveys more recent work. Information flow checking has been also integrated (as a single-purpose extension) in the FlowCaml [79] language. Finally, Russo et al. and Swamy et al. [74, 81] show that such properties can be checked using a monadic library.

Systems for secure information flow typically define a lattice of security classes  $(S, \leq)$  where  $S$  is a finite set of classes and an ordering. For example a set  $\{L, H\}$  represents low and high secrecy, respectively with  $L \leq H$  meaning that low security values can be treated as high security (but not the other way round).

IMPLICIT FLOWS. An important aspect of secure information flow is called *implicit flows*. Consider the following example which returns either  $y$  or zero, depending on the value of  $x$ :

```
let z = if x > 0 then y else 0
```

If the value of  $y$  is high-secure, then  $z$  becomes high-secure after the assignment (this is an *explicit* flow). However, if  $x$  is high-secure, then the value of  $z$  becomes high-secure, regardless of the security level of  $y$ , because the fact whether an assignment is performed or not performed leaks information in its own (this is an *implicit* flow).

Although we do not describe a coefficient calculus for information flow checking, it is worth noting that Abadi et al. [1] realized that there is a number of analyses similar to secure information flow and unified them using a single model called Dependency Core Calculus (DCC). This would be a useful basis for coefficient-based information flow checking.

The DCC captures other cases where some information about expression relies on properties of variables in the context where it executes. This includes, for example, *binding time analysis* [89], which detects which parts of programs can be partially evaluated (do not depend on user input) and *program slicing* [90] that identifies parts of programs that contribute to the output of an expression.

COEFFICIENT SYSTEMS. The work outlined in this section is another area where coefficient systems could be applied. We do not develop coefficient systems for taint tracking, security and provenance in detail, but briefly mention some examples in the upcoming chapters.

The systems work in the same way as the examples discussed already. For example, consider the tainting example with the query function calling an SQL database. To capture such tainting, we annotate variables with  $T$  for *tainted* and with  $U$  for *untainted*. Accessing a variable marks it as untainted, but using an expression that depends on some variable in certain dangerous contexts – such as in arguments of query – does introduce a taint on all the variables contributing to the expression. This is captured using the standard application rule (*app*):

$$(\text{app}) \frac{\Gamma @ \mathbf{r} \vdash \text{query} : \text{string} \xrightarrow{T} \text{Table} \quad \text{id} : \text{string} @ \langle \mathbf{U} \rangle \vdash \text{id} : \text{string}}{\Gamma, \text{id} : \text{string} @ \mathbf{r} \times \langle \mathbf{T} \rangle \vdash \text{query}(\dots, \text{id}) : \text{Table}}$$

The derivation assumes that `query` is a standard function that requires the parameters to be tainted (it does not have to be a built-in language construct). The argument is a variable and so it is not tainted in the assumptions.

In the conclusion, we need to derive an annotation for the variable `id`. To do this, we combine  $T$  (from the function) and  $U$  (from the argument). In case of tainting, the variable is tainted whenever it is already tainted *or* the function marks it as tainted. For different kinds of annotations, the composition would work differently – for example, for provenance, we could union the *set* of possible data sources, or even combine *probability distributions* modelling the influence of different sources on the value. However, expanding such ideas is beyond the scope of this thesis.

### 3.4 BEYOND PASSIVE CONTEXTS

In both flat and structural systems discussed so far, the context provides additional data (resources, implicit parameters, historical values) or meta-

data (security, provenance). However, *within* the language, it is impossible to write a function that modifies the context. We use the term *passive* context for such applications.

There is a number of systems that also capture contextual properties, but that make it possible to *change* the context – not just by evaluating certain code block in a locally modified context (e.g. by wrapping it in `prev` in data-flow), but also by calling a function that, for example, acquires new capabilities and returns those to the caller. Such actions appear to be closer to effects than to coeffects. While this thesis focuses on systems with passive contexts, we briefly consider the most important examples of the *active* variant.

**CALCULUS OF CAPABILITIES.** Crary et al. [23] introduced the Calculus of Capabilities to provide a sound system with region-based memory management for low-level code that can be easily compiled to assembly language. They build on the work of Tofte and Talpin [91] who developed an effect system (as discussed in Section 2.2.2) that uses lexically scoped *memory regions* to provide an efficient and controlled memory management.

In the work of Tofte and Talpin, the context is *passive*. They extend a simple functional language with the `letrgn` construct that defines a new memory region, evaluates an expression (possibly) using memory in that region and then deallocates the memory of the region:

```
let calculate = λinput →
  letrgn ρ in
  let x = refρ input in
  x := !x + 1; !x
```

The memory region  $\rho$  is a part of the context, but only in the scope of the body of `letrgn`. It is only available to the last two lines which allocate a memory cell in the region, increment a value in the region and then read it. The region is de-allocated when the execution leaves its lexical scope – there is no way to allocate a region inside a function and pass it back to the caller.

The calculus of capabilities differs in two ways. First, it allows explicit allocation and deallocation of memory regions (and so region lifetimes do not necessarily follow strict LIFO ordering). Second, it uses continuation-passing style. We ignore the latter aspect. The following example is almost identical to the previous one:

```
let calculate = λinput →
  letrgn ρ in
  let x = refρ input in
  x := !x + 1; x
```

The difference is that the example does not return the *value* of a reference using `!x`, but returns the reference `x` itself. The reference is allocated in a newly created region  $\rho$ . Together with the value, the function returns a *capability* to access the region  $\rho$ .

This is where systems with active context differ from systems with passive context. To type check such programs, we do not only need to know what context is required to call `calculate` (i.e. context on the left-hand side of  $\vdash$ ). We also need to know what effects an expression has on the context when it evaluates and the current context needs to be updated after a function call. This is an effectful property that would appear on the right-hand side of  $\vdash$ .

ACTIVE CONTEXTS. In a systems with passive contexts, we only need an annotation that specifies the required context. In semantics, this is reflected by having some structure (data type)  $\mathcal{C}$  over the *input* of the function. Without giving any details, the semantics generally has the following structure (with comonad to model coeffects on the left):

$$\llbracket \tau_1 \xrightarrow{r} \tau_2 \rrbracket = \mathcal{C}^r \tau_1 \rightarrow \tau_2$$

Systems with active contexts require two annotations – one that specifies the context required before the call is performed and one that specifies how the context changes after the call (this could be either a *new* context or *update* to the original context). Thus the structure of the semantics would look as follows (with comonad to model coeffects on the left and monad to model effects on the right):

$$\llbracket \tau_1 \xrightarrow{r,s} \tau_2 \rrbracket = \mathcal{C}^r \tau_1 \rightarrow \mathcal{M}^s \tau_2$$

In case of Calculus of Capabilities, both of the structures could be the same and they could carry a set of available memory regions. In this thesis, we focus only on passive contexts. However, capturing active contexts is an interesting future work.

SOFTWARE UPDATING. Another example of a system that uses contextual information actively is dynamic software updating (DSU) [31, 41]. The DSU systems have the ability to update programs at runtime without stopping them. For example, Proteus developed by Stoyle et al. [80] investigates what language support is needed to enable safe dynamic software updating in C-like languages. The work is based on capabilities and follows a structure similar to the Calculus of Capabilities [23].

The system distinguishes between *concrete* uses and *abstract* uses of a value. When a value is used concretely, the program examines its representation (and so it is not safe to change the representation during an update). An abstract use of a value does not examine the representation and so updating the value does not break the program.

The Proteus system uses capabilities to restrict what types may be used concretely after any point in the program. All other types, not listed in the capability, can be dynamically updated as this will not change concrete representation of types accessed later in the evaluation.

Similarly to Capability Calculus, capabilities in DSU can be changed by a function call. For example, calling a function that may update certain types makes it impossible to use those types concretely following the function call. This means that DSU uses the context *actively* and not just *passively*.

### 3.5 SUMMARY

This chapter served two purposes. The first aim was to present existing work on programming languages and systems that include some notion of *context*. Because there was no well-known abstraction capturing contextual properties, the languages use a wide range of formalisms – including principled approaches based on comonads and modal S4, ad-hoc type system extensions and static analyses as well as approaches based on monads. We looked at a number of applications including Haskell’s implicit parameters and type classes, data-flow languages such as Lucid, liveness analysis and also a number of security properties.

The second aim of this chapter was to re-formulate the existing work in a more uniform style and thus reveal that all *context-dependent* languages share a common structure. In the upcoming three chapters, we identify the common structure more precisely and develop three calculi to capture it. We will then be able to re-create many of the examples discussed in this chapter just by instantiating our unified calculi.

This chapter was divided into two major sections. First, we looked at *flat* systems, which track whole-context properties. Next, we look at *structural* systems, which track per-variable properties. Both of the variants are useful and important – for example, implicit parameters can only be expressed as *flat* system, but liveness analysis is only useful as *structural*. For this reason, we explore both of these variants in this thesis (Chapter 4 and Chapter 5, respectively). We can, however, unify the two variants into a single system discussed in Chapter 6.

Successful programming language abstractions need to generalize a wide range of recurring problems while capturing the key commonalities. These two aims are typically in opposition – more general abstractions are less powerful, while less general abstractions cannot be used as often.

In the previous chapter, we outlined a number of systems that capture how computations access the environment in which they are executed. We identified two kinds of systems – *flat systems* capturing whole-context properties and *structural systems* capturing per-variable properties. As we show in Chapter 6, the systems can be unified using a single abstraction. This is useful when implementing and composing the systems, but such abstraction is *less powerful* – i. e. its generality hides useful properties that we can see when we consider the systems separately. For this reason, we discuss *flat coeffects* and *structural coeffects* separately in this and the next chapter.

#### CHAPTER STRUCTURE AND CONTRIBUTIONS

- We present a *flat coeffect calculus* as a type system that is parameterized by a *flat coeffect algebra* (Section 4.2). We show that the system can be instantiated to obtain three of the systems discussed in Section 3.2, namely implicit parameters, liveness and data-flow.
- We give the equational theory of the calculus and discuss type-preservation for call-by-name and call-by-value reduction (Section 4.4). We also extend the calculus with subtyping and pairs (Section 4.5). These two sections motivate the laws of the flat coeffect algebra.
- We present the semantics of the calculus in terms of *indexed comonads*, which are a generalization of comonads, a category-theoretical dual of monads (Section 4.3). This provides a deeper insight into how (and why) the calculus works and shows an intriguing link with effects.

#### 4.1 INTRODUCTION

In the previous chapter, we looked at three important examples of systems that track whole-context properties. The type systems for whole-context liveness (Section 3.2.3) and whole-context data-flow (Section 3.2.4) have a very similar structure. First, their lambda abstraction duplicates the requirements. Given a body with context requirements  $\mathbf{r}$ , the declaration site context *as well as* the function arrow are annotated with  $\mathbf{r}$ . Second, their application arises from the combination of *sequential* and *pointwise* composition.

The system for tracking of implicit parameters and similar (Section 3.2.1) differ in two ways. In lambda abstraction, they split the context requirements between the declaration site and the call site and they use only a single operator on the indices, typically  $\cup$ .

Despite the differences, the systems fit the same unified framework. This becomes apparent when we consider the categorical structure (Section 4.3). However, rather than starting from the semantics, we first explain how the systems can be unified syntactically (Section 4.2.1) and then provide the semantics as a justification.



#### 4.1.1 Related work

The development in this chapter can be seen as a counterpart to the well-known development of *effect systems* [33] and the use of *monads* [55] in programming languages. The syntax and type system of the flat coeffect calculus follows a similar style as effect systems [53, 86], but differs in the structure, as explained in the previous chapter, most importantly in lambda abstraction (the relationship with monads is further discussed in Section 4.6).

Wadler and Thiemann [103] famously showed a correspondence between effect systems to monads (a topic that has been also discussed by Atkey [6] and recently revisited by Vazou and Leijen [59]). This line of work relates effectful functions  $\tau_1 \xrightarrow{\sigma} \tau_2$  to monadic computations  $\tau_1 \rightarrow M^\sigma \tau_2$ . In this chapter, we show a similar correspondence between *coeffect systems* and *comonads*. However, due to the asymmetry of  $\lambda$ -calculus, defining the semantics in terms of comonadic computations is not a simple mechanical dualization of the work on effect systems and monads.

The main purpose of the comonadic semantics presented in this chapter is to provide a motivation for the typing of the flat coeffect calculus. Our approach is inspired by the work of Uustalu and Vene [93] who present the semantics of contextual computations (mainly for data-flow) in terms of comonadic functions  $C\tau_1 \rightarrow \tau_2$ . Our *indexed comonads* annotate the structure with information about the required context, i.e.  $C^r\tau_1 \rightarrow \tau_2$ . This is similar to the recent work on *parameterized monads* by Katsumata [45].

## 4.2 FLAT COEFFECT CALCULUS

The flat coeffect calculus is defined in terms of *flat coeffect algebra*, which defines the structure of context annotations, such as  $r, s, t$ . These can be sets of implicit parameters, versions represented as integers or other values. The expressions of the calculus are those of the  $\lambda$ -calculus with *let* binding; assuming  $\alpha$  ranges over base types, the types of the calculus are defined as:

$$\begin{aligned} e &::= x \mid \lambda x. e \mid e_1 \ e_2 \mid \text{let } x = e_1 \text{ in } e_2 \\ \tau &::= \alpha \mid \tau_1 \xrightarrow{r} \tau_2 \end{aligned}$$

We discuss subtyping and pairs in Section 4.5. The type  $\tau_1 \xrightarrow{r} \tau_2$  represents a function from  $\tau_1$  to  $\tau_2$  that requires additional context  $r$ . It can be viewed as a pure function that takes  $\tau_1$  *with* or *wrapped in* a context  $r$ .

In the categorical semantics, the function  $\tau_1 \xrightarrow{r} \tau_2$  is modelled by a morphism  $C^r\tau_1 \rightarrow \tau_2$ . However, the object  $C^r$  does not itself exist as a syntactical value. This is because we use comonads to define the *semantics* rather than *embedding* them into the language as in the meta-language approaches (the distinction has been discussed in Section 2.2.1). The annotations  $r$  are formed by an algebraic structure discussed next.

#### 4.2.1 Reconciling lambda abstraction

Recall the lambda abstraction rules for the implicit parameters system (annotating the context with sets of required parameters) and the data-flow system (annotating the context with the number of past required values):

$$\begin{array}{c} \text{(param)} \quad \frac{\Gamma, x : \tau_1 @ r \cup s \vdash e : \tau_2}{\Gamma @ r \vdash \lambda x. e : \tau_1 \xrightarrow{s} \tau_2} \quad \text{(df1)} \quad \frac{\Gamma, x : \tau_1 @ n \vdash e : \tau_2}{\Gamma @ n \vdash \lambda x. e : \tau_1 \xrightarrow{n} \tau_2} \end{array}$$



In order to capture both systems using a single calculus, we need a way of unifying the two systems. For the data-flow system, this can be achieved by over-approximating the number of required past elements:

$$(df2) \frac{\Gamma, x : \tau_1 @ \min(\mathbf{n}, \mathbf{m}) \vdash e : \tau_2}{\Gamma @ \mathbf{n} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{m}} \tau_2}$$

The rule  $(df1)$  is admissible in a system that includes the  $(df2)$  rule. Furthermore, if we include sub-typing rule (on annotations of functions) and sub-coeffecting rule (on annotations of contexts), then the reverse is also true – because  $\min(\mathbf{n}, \mathbf{m}) \leq \mathbf{m}$  and  $\min(\mathbf{n}, \mathbf{m}) \leq \mathbf{n}$ . In other words  $(df1)$  is more precise, but  $(df2)$  gives a sound over-approximation with a structure that can be unified with  $(param)$ .

#### 4.2.2 Understanding flat coeffects

Before looking at the type system in Figure 22, let us clarify how the rules should be understood. The coeffect calculus provides both analysis of context dependence (type system) and semantics for context (how it is propagated). These two aspects provide different ways of reading the judgements  $\Gamma @ \mathbf{r} \vdash e : \tau$  and the typing rules used to define it.

- ANALYSIS OF CONTEXT DEPENDENCE. Syntactically, coeffect annotations  $\mathbf{r}$  model *context requirements*. This means we can over-approximate them and require more than is actually needed at runtime.

Syntactically, the typing rules should be read top-down. In function application, the context requirements of multiple assumptions (arising from two sub-expressions) are *merged*; in lambda abstraction, the requirements of a single expression (the body) are split between the declaration site and the call site.

- SEMANTICS OF CONTEXT PASSING. Semantically, coeffect annotations  $\mathbf{r}$  model *contextual capabilities*. This means that we can throw away capabilities, if a sub-expression requires fewer than we currently have.

Semantically, the typing rules should be read bottom-up. In application, the capabilities provided to the term  $e_1 \ e_2$  are *split* between the two sub-expressions; in abstraction, the capabilities provided by the call site and declaration site are *merged* and passed to the body.

The reason for this asymmetry follows from the fact that the context appears in a *negative position* in the semantic model (Section 4.3). It means that we need to be careful about using the words *split* and *merge*, because they can be read as meaning exactly the opposite things. To disambiguate, we always use the term *context requirements* when using the syntactic view, especially in the rest of Section 4.2, and *context capabilities* or just *available context* when using the semantic view, especially in Section 4.3.

#### 4.2.3 Flat coeffect algebra

To make the flat coeffect system general enough, the algebra consists of three operations. Two of them,  $\otimes$  and  $\oplus$ , represent the *sequential* and *point-wise* composition, which are mainly used in function application. The third operator,  $\wedge$  is used in lambda abstraction and represents *splitting* of context requirements (or, semantically, *merging* of available context capabilities).

In addition to the three operations, we also require two special values used to annotate variable access and constant access and a relation that defines the ordering.

**Definition 3.** A *flat coeffect algebra*  $(\mathcal{C}, \otimes, \oplus, \wedge, \text{use}, \text{ign}, \leq)$  is a set  $\mathcal{C}$  together with elements  $\text{use}, \text{ign} \in \mathcal{C}$ , relation  $\leq$  and binary operations  $\otimes, \oplus, \wedge$  such that  $(\mathcal{C}, \otimes, \text{use})$  is a monoid,  $(\mathcal{C}, \oplus, \text{ign})$  is an idempotent monoid,  $(\mathcal{C}, \leq)$  is a pre-order and  $(\mathcal{C}, \wedge)$  is a band (idempotent semigroup). That is, for all  $r, s, t \in \mathcal{C}$ :

$$\begin{aligned} r \otimes (s \otimes t) &= (r \otimes s) \otimes t & \text{use} \otimes r &= r = r \otimes \text{use} \\ r \oplus (s \oplus t) &= (r \oplus s) \oplus t & r \oplus r &= r & \text{ign} \oplus r &= r = r \oplus \text{ign} \\ r \wedge (s \wedge t) &= (r \wedge s) \wedge t & r \wedge r &= r \\ \text{if } r \leq s \text{ and } s \leq t &\text{ then } r \leq t & t &\leq t \end{aligned}$$

In addition, the following distributivity axioms hold:

$$\begin{aligned} (r \oplus s) \otimes t &= (r \otimes t) \oplus (s \otimes t) \\ t \otimes (r \oplus s) &= (t \otimes r) \oplus (t \otimes s) \end{aligned}$$

In two of the three systems, some of the operators of the flat coeffect algebra coincide, but in the data-flow system all three are distinct. Similarly, the two special elements coincide in some, but not all systems. The required laws are motivated by the aim to capture common properties of the three examples, without unnecessarily restricting the system:

- The monoid  $(\mathcal{C}, \otimes, \text{use})$  represents *sequential* composition of (semantic) functions. The laws of a monoid are required in order to form a category structure in the semantics (Section 4.3).
- The idempotent monoid  $(\mathcal{C}, \oplus, \text{ign})$  represents *pointwise* composition, i. e. the case when the same context is passed to multiple (independent) computations. The monoid laws guarantee that usual syntactic transformations on tuples and the unit value (Section 4.5) preserve the coeffect. Idempotence holds for all our examples and allows us to unify the flat and structural systems in Chapter 6.
- For the  $\wedge$  operation, we require associativity and idempotence. The idempotence requirement makes it possible to duplicate the coeffects and place the same requirement on both call site and declaration site. Using the example from Section 4.2.1, this guarantees that the rule (df1) is not a special case, but can always be derived from (df2). In some cases, the operator forms a monoid with the unit being the greatest element of the set  $\mathcal{C}$ .

It is worth noting that, in some of the systems, the operators  $\oplus$  and  $\wedge$  are the least upper bound and the greatest lower bounds of a lattice. For example, in data-flow computations, they are *max* and *min* respectively. However, this duality does not hold for implicit parameters (we discuss lattice-based formulation of coeffects in Section 6.2.3).

Using the syntactic reading, the two operators represent *merging* and *splitting* of context requirements – in the (*abs*) rule,  $\wedge$  appears in the assumption and the combined context requirements of the body are split between two positions in the conclusions; in the (*app*) rule,  $\oplus$  appears in the conclusion and combines two context requirements from the assumptions.

$$\begin{array}{c}
\text{(var)} \quad \frac{x : \tau \in \Gamma}{\Gamma @ \text{use} \vdash x : \tau} \\
\text{(const)} \quad \frac{c : \tau \in \Delta}{\Gamma @ \text{ign} \vdash c : \tau} \\
\text{(sub)} \quad \frac{\Gamma @ r' \vdash e : \tau}{\Gamma @ r \vdash e : \tau} \quad (r' \leq r) \\
\text{(app)} \quad \frac{\Gamma @ r \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \quad \Gamma @ s \vdash e_2 : \tau_1}{\Gamma @ r \oplus (s \otimes t) \vdash e_1 e_2 : \tau_2} \\
\text{(abs)} \quad \frac{\Gamma, x : \tau_1 @ r \wedge s \vdash e : \tau_2}{\Gamma @ r \vdash \lambda x. e : \tau_1 \xrightarrow{s} \tau_2} \\
\text{(let)} \quad \frac{\Gamma @ r \vdash e_1 : \tau_1 \quad \Gamma, x : \tau_1 @ s \vdash e_2 : \tau_2}{\Gamma @ s \oplus (s \otimes r) \vdash \text{let } x = e_1 \text{ in } e_2 : \tau_2}
\end{array}$$

Figure 22: Type system for the flat coeffect calculus

**ORDERING.** The flat coeffect algebra requires a pre-order relation  $\leq$ , which is used to define sub-coeffecting rule of the type system. When the idempotent monoid  $(\mathcal{C}, \oplus, \text{ign})$  also has the commutative property (i. e. forms a semi-lattice), the  $\leq$  relation can be defined as the ordering of the semi-lattice:

$$r \leq s \iff r \oplus s = s$$

This definition is consistent with all three examples that motivate flat coeffect calculus, but it cannot be used with the structural coeffects (where it fails for the bounded reuse calculus) and so we choose not to use it.

Furthermore, the **use** coeffect is often the top or the bottom element of the semi-lattice. As discussed in Section 4.4, when this is the case, we are able to prove certain syntactic properties of the calculus.

#### 4.2.4 Flat coeffect types

The type system for flat coeffect calculus is shown in Figure 22. Variables (*var*) and constants (*const*) are annotated with special values provided by the coeffect algebra. Following the top-down syntactic reading, the (*sub*) rule allows us to treat an expression with fewer context requirements as an expression with more context requirements.

The (*abs*) rule is defined as discussed in Section 4.2.1. The body is annotated with context requirements  $r \wedge s$ , which are then split between the context-requirements on the declaration site  $r$  and context-requirements on the call site  $s$ . Examples of the  $\wedge$  operator are discussed in the next section.

In function application (*app*), context requirements of both expressions and the function are combined. As discussed in Chapter 3, the pointwise composition  $\oplus$  is used to combine the context requirements of the expression representing a function  $r$  and the context requirements of the argument, sequentially composed with the context-requirements of the function  $s \otimes t$ .

The type system also includes a rule for let-binding. The rule is *not* equivalent to the typing derivation for  $(\lambda x. e_2) e_1$ , but it corresponds to *one* possible typing derivation. As we show in 4.2.6, the typing used in (*let*) is more precise than the general rule that can be derived from  $(\lambda x. e_2) e_1$ . Additional constructs such as pairs are covered in Section 4.5.

#### 4.2.5 Examples of flat coeffects

The flat coeffect calculus generalizes the flat systems discussed in Section 3.2 of the previous chapter. We can instantiate it to a specific use just by providing a flat coeffect algebra. The following summary defines the systems for implicit parameters, liveness and data-flow. For the latter two, the general calculus has a lambda abstraction that is compatible with those discussed in Chapter 4, but includes implicit sub-coffecting.

**Example 1** (Implicit parameters). *Assuming  $\text{ld}$  is a set of implicit parameter names, the flat coeffect algebra is formed by  $(\mathcal{P}(\text{ld}), \cup, \cup, \cup, \emptyset, \emptyset, \subseteq)$ .*

For simplicity, we assume that all parameters have the same type  $\rho$  and so the annotations only track sets of names. The definition uses set union for all three operations. Both variables and constants are annotated with  $\emptyset$  and the ordering is defined by  $\subseteq$ . The definition satisfies the flat coeffect algebra laws because  $(S, \cup, \emptyset)$  is an idempotent, commutative monoid. The system has a single additional typing rule for accessing the value of a parameter:

$$(param) \frac{?p \in c}{\Gamma @ c \vdash ?p : \rho}$$

The rule specifies that the accessed parameter  $?p$  needs to be in the set of required parameters  $c$ . As discussed earlier, we use the same type  $\rho$  for all parameters, but it is easy to define an extension that uses mappings from names to types.

**Example 2** (Liveness). *Let  $\mathcal{L} = \{L, D\}$  be a two-point lattice such that  $D \sqsubseteq L$  with a join  $\sqcup$  and meet  $\sqcap$ . The flat coeffect algebra for liveness is then formed by  $(\mathcal{L}, \sqcap, \sqcup, \sqcap, \sqcup, L, D, \sqsubseteq)$ .*

As in Section 3.2.3, sequential composition  $\circledast$  is modelled by the meet operation  $\sqcap$  and pointwise composition  $\oplus$  is modelled by join  $\sqcup$ . The two-point lattice is a commutative, idempotent monoid. Distributivity  $(r \sqcup s) \sqcap t = (r \sqcap t) \sqcup (s \sqcap t)$  does not hold for *every* lattice, but it trivially holds for the two-point lattice used here.

The definition uses join  $\sqcup$  for the  $\wedge$  operator that is used by lambda abstraction. This means that, when the body is live  $L$ , both declaration site and call site are marked as live  $L$ . When the body is dead  $D$ , the declaration site and call site can be marked as dead  $D$ , or as live  $L$ . The latter is less precise, but permissible over-approximation, which could otherwise be obtained via sub-typing.

**Example 3** (Data-flow). *In data-flow, context is annotated with natural numbers and the flat coeffect algebra is formed by  $(\mathbb{N}, +, \max, \min, 0, 0, \leq)$ .*

As discussed earlier, sequential composition  $\circledast$  is represented by  $+$  and pointwise composition  $\oplus$  uses  $\max$ . For data-flow, we need a third separate operator for lambda abstraction. Annotating the body with  $\min(r, s)$  ensures that both call site and declaration site annotations are equal or greater than the annotation of the body. As with liveness, this allows over-approximation.

As required by the laws,  $(\mathbb{N}, +, 0)$  and  $(\mathbb{N}, \max, 0)$  form monoids and  $(\mathbb{N}, \min)$  forms a band. Note that data-flow is our first example where  $\circledast$  is not idempotent. The distributivity laws require the following to be the case:  $\max(r, s) + t = \max(r + t, s + t)$ , which is easy to see. Finally, a simple data-flow language includes an additional rule for [prev](#):

$$(prev) \frac{\Gamma @ c \vdash e : \tau}{\Gamma @ c + 1 \vdash \text{prev } e : \tau}$$

As a further example that was not covered earlier, it is also possible to combine liveness analysis and data-flow. In the above calculus,  $0$  denotes that we require the current value, but no previous values. However, for constants, we do not even need the current value.

**Example 4** (Optimized data-flow). *In optimized data-flow, context is annotated with natural numbers extended with the  $\perp$  element, that is  $\mathbb{N}_\perp = \{\perp, 0, 1, 2, 3, \dots\}$  such that  $\forall n \in \mathbb{N}. \perp \leq n$ . The flat coeffect algebra is  $(\mathbb{N}_\perp, +, \max, \min, 0, \perp, \leq)$  where  $m + n$  is  $\perp$  whenever  $m = \perp$  or  $n = \perp$  and  $\min, \max$  treat  $\perp$  as the least element.*

Note that  $(\mathbb{N}_\perp, +, 0)$  is a monoid for the extended definition of  $+$ ; for the bottom element  $0 + \perp = \perp$  and for natural numbers  $0 + n = n$ . The structure  $(\mathbb{N}, \max, \perp)$  is also a monoid, because  $\perp$  is the least element and so  $\max(n, \perp) = n$ . Finally,  $(\mathbb{N}, \min)$  is a band (the extended  $\min$  is still idempotent and associative) and the distributivity law also holds for  $\mathbb{N}_\perp$ .

#### 4.2.6 Typing of let binding

Recall the (*let*) rule in Figure 22. It annotates the expression `let  $x = e_1$  in  $e_2$`  with context requirements  $s \oplus (s \otimes r)$ . This is a special case of typing an expression  $(\lambda x. e_2) e_1$ , using the idempotence of  $\wedge$  as follows:

$$(app) \frac{\Gamma @ r \vdash e_1 : \tau_1 \quad \frac{\Gamma, x : \tau_1 @ s \vdash e_2 : \tau_2}{\Gamma @ s \vdash \lambda x. e_2 : \tau_1 \xrightarrow{s} \tau_2} (abs)}{\Gamma @ s \oplus (s \otimes r) \vdash (\lambda x. e_2) e_1 : \tau_2}$$

This design decision is similar to ML value restriction, but it works the other way round. Our *let* binding is more restrictive than the typing of abstraction-application, rather than being more general. The choice is motivated by the fact that the typing obtained using the special rule for let-binding is more precise (with respect to sub-coeffecting) for all the examples consider in this chapter. Table 1 shows how the coeffect annotations are simplified for our examples.

	Definition	Simplified
Implicit parameters	$s \cup (s \cup r)$	$s \cup r$
Liveness	$s \sqcap (s \sqcup r)$	$s$
Data-flow	$\max(s, s + r)$	$s + r$

Table 1: Simplified annotation for let binding in sample flat calculi instances

The simplified annotations directly follow from the definitions of particular flat coeffect algebras. It is perhaps somewhat unexpected that the annotation can be simplified in different ways for different examples.

To see that the simplified annotations are more precise, assume that we used arbitrary splitting  $s = s_1 \wedge s_2$  rather than idempotence. The “Definition” column would use  $s_1$  and  $s_2$  for the first and second  $s$ , respectively. The corresponding simplified annotation would have  $s_1 \wedge s_2$  instead of  $s$ . For all our systems, the simplified annotation (on the right) is more precise than the original (on the left):

$$\begin{aligned} s_1 \cup (s_2 \cup r) &\supseteq (s_1 \cup s_2) \cup r && \text{(implicit parameters)} \\ s_1 \sqcap (s_2 \sqcup r) &\supseteq (s_1 \sqcap s_2) && \text{(liveness)} \\ \max(s_1, s_2 + r) &\supseteq \min(s_1, s_2) + r && \text{(data-flow)} \end{aligned}$$

In other words, the inequality states that using idempotence, we get a more precise typing. This can be expressed using the abstract operators of the flat coeffect algebra as:

$$s_1 \oplus (s_2 \otimes r) \geq (s_1 \wedge s_2) \oplus ((s_1 \wedge s_2) \otimes r)$$

This property cannot be proved from other properties of the flat coeffect algebra. To make the flat coeffect system as general as possible, we do not *in general* require it as an additional axiom, although the above examples provide reasonable basis for requiring that the specialized annotation for let binding is the least possible annotation for the expression  $(\lambda x.e_2) e_1$ .

### 4.3 CATEGORICAL MOTIVATION

The type system of flat coeffect calculus arises syntactically, as a generalization of the examples discussed in Chapter 3, but we can also obtain it by looking at the categorical semantics of context-dependent computations. This is a direction that we explore in this section. Although the development presented here is interesting in its own, our main focus is *using* categorical semantics to motivate and explain the design of flat coeffect calculus.

#### 4.3.1 Categorical semantics

As discussed in Section 2.2, categorical semantics interprets terms as morphisms in some category. For typed calculi, the semantics defined by  $\llbracket - \rrbracket$  usually interprets typing judgements  $x_1 : \tau_1 \dots x_n : \tau_n \vdash e : \tau$  as morphisms  $\llbracket \tau_1 \times \dots \times \tau_n \rrbracket \rightarrow \llbracket \tau \rrbracket$ .

As a best known example, Moggi [55] showed that the semantics of various effectful computations can be captured uniformly using (*strong*) *monads*. In that approach, computations are interpreted as  $\tau_1 \times \dots \times \tau_n \rightarrow M\tau$  for some monad  $M$ . For example,  $M\alpha = \alpha \cup \{\perp\}$  models partiality (maybe monad),  $M\alpha = \mathcal{P}(\alpha)$  models non-determinism (list monad) and side-effects can be modelled using  $M\alpha = S \rightarrow (\alpha \times S)$  (state monad). Here, the structure of a strong monad provides necessary “plumbing” for composing monadic computations – sequential composition and strength for lifting free variables into the body of computation under a lambda abstraction.

Following similar approach to Moggi, Uustalu and Vene [93] showed that (*monoidal*) *comonads* uniformly capture the semantics of various kinds of context-dependent computations [93]. For example, data-flow computations over non-empty lists are modelled using the non-empty list comonad  $NEList \alpha = \alpha + (\alpha \times NEList \alpha)$ .

The monadic and comonadic model outlined here represents at most a binary analysis of effects or context-dependence. A function  $\tau_1 \rightarrow \tau_2$  performs *no* effects (requires no context) whereas  $\tau_1 \rightarrow M\tau_2$  performs *some* effects and  $C\tau_1 \rightarrow \tau_2$  requires *some* context<sup>1</sup>.

In the next section, we introduce *indexed comonads*, which provide a more precise analysis and let us model computations with context requirements  $r$  as functions  $C^r\tau_1 \rightarrow \tau_2$  using an *indexed comonad*  $C^r$ .

<sup>1</sup> This is an over-simplification as we can use e.g. stacks of monad transformers and model functions with two different effects using  $\tau_1 \rightarrow M_1(M_2 \tau_2)$ . However, monad transformers require defining complex system of lifting to be composable. Consequently, they are usually used for capturing different kinds of impurities (exceptions, non-determinism, state), but not for capturing fine-grained properties (e.g. a set of memory regions that may be accessed by a stateful computation).

## 4.3.2 Introducing comonads

In category theory, *comonad* is a dual of *monad*. As already outlined in Section 2.2, we get a definition of a comonad by taking a definition of a monad and “reversing the arrows”. More formally, one of the equivalent definitions of comonad looks as follows:

**Definition 4.** A comonad over a category  $\mathcal{C}$  is a triple  $(C, \text{counit}, \text{cobind})$  where:

- $C$  is a mapping on objects (types)  $C : \mathcal{C} \rightarrow \mathcal{C}$
- $\text{counit}$  is a mapping  $C\alpha \rightarrow \alpha$
- $\text{cobind}$  is a mapping  $(C\alpha \rightarrow \beta) \rightarrow (C\alpha \rightarrow C\beta)$

such that, for all  $f : C\alpha \rightarrow \beta$  and  $g : C\beta \rightarrow \gamma$ :

$$\text{cobind } \text{counit} = \text{id} \quad (\text{left identity})$$

$$\text{counit} \circ \text{cobind } f = f \quad (\text{right identity})$$

$$\text{cobind } (g \circ \text{cobind } f) = (\text{cobind } g) \circ (\text{cobind } f) \quad (\text{associativity})$$

From the functional programming perspective, we can see  $C$  as a parametric data type such as `NEList`. The `counit` operation extracts a value  $\alpha$  from a value that carries additional context  $C\alpha$ . The `cobind` operation turns a context-dependent function  $C\alpha \rightarrow \beta$  into a function that takes a value with context, applies the context-dependent function to value(s) in the context and then propagates the context.

As mentioned earlier, Uustalu and Vene [93] use comonads to model data-flow computations. They describe infinite (coinductive) streams and non-empty lists as example comonads.

**Example 5** (Non-empty list). A non-empty list is a recursive data-type defined as  $\text{NEList } \alpha = \alpha + (\alpha \times \text{NEList } \alpha)$ . We write `inl` and `inr` for constructors of the left and right cases, respectively. The type `NEList` forms a comonad together with the following `counit` and `cobind` mappings:

$$\begin{aligned} \text{counit } l &= h && \text{when } l = \text{inl } h \\ \text{counit } l &= h && \text{when } l = \text{inr } (h, t) \\ \text{cobind } f \, l &= \text{inl } (f \, l) && \text{when } l = \text{inl } h \\ \text{cobind } f \, l &= \text{inr } (f \, l, \text{cobind } f \, t) && \text{when } l = \text{inr } (h, t) \end{aligned}$$

The `counit` operation returns the head of the non-empty list. Note that it is crucial that the list is *non-empty*, because we always need to be able to obtain a value. The `cobind` operation defined here returns a list of the same length as the original where, for each element, the function  $f$  is applied on a *suffix* list starting from the element. Using a simplified notation for list, the result of applying `cobind` to a function that sums elements of a list gives the following behaviour:

$$\text{cobind } \text{sum } (7, 6, 5, 4, 3, 2, 1, 0) = (28, 21, 15, 10, 6, 3, 1, 0)$$

The fact that the function  $f$  is applied to a *suffix* is important in order to satisfy the *left identity* law, which requires that  $\text{cobind } \text{counit } l = l$ .

It is also interesting to examine some data types that do *not* form a comonad. As already mentioned, list  $\text{List } \alpha = 1 + (\alpha \times \text{List } \alpha)$  is not a comonad, because the `counit` operation is not defined for the value `inl ()`. The `Maybe` data type defined as  $1 + \alpha$  is not a comonad for the same reason. However, if we consider flat coeffect calculus for liveness, it appears natural to model computations as functions  $\text{Maybe } \tau_1 \rightarrow \tau_2$ . To use such a model, we need to generalise comonads to *indexed comonads*.



### 4.3.3 Generalising to indexed comonads

The flat coeffect algebra includes a monoid  $(\mathcal{C}, \otimes, \text{use})$ , which defines the behaviour of sequential composition, where the annotation  $\text{use}$  represents a variable access. An indexed comonad is formed by a data type (object mapping)  $C^r \alpha$  where the annotation  $r$  determines what context is required.

**Definition 5.** Given a monoid  $(\mathcal{C}, \otimes, \text{use})$  with binary operator  $\otimes$  and unit  $\text{use}$ , an indexed comonad over a category  $\mathcal{C}$  is a triple  $(C^r, \text{counit}_{\text{use}}, \text{cobind}_{r,s})$  where:

- $C^r$  for all  $r \in \mathcal{C}$  is a family of object mappings
- $\text{counit}_{\text{use}}$  is a mapping  $C^{\text{use}} \alpha \rightarrow \alpha$
- $\text{cobind}_{r,s}$  is a mapping  $(C^r \alpha \rightarrow \beta) \rightarrow (C^{r \otimes s} \alpha \rightarrow C^s \beta)$

such that, for all  $f : C^r \alpha \rightarrow \beta$  and  $g : C^s \beta \rightarrow \gamma$ :

$$\text{cobind}_{\text{use},s} \text{counit}_{\text{use}} = \text{id} \quad (\text{left identity})$$

$$\text{counit}_{\text{use}} \circ \text{cobind}_{r,\text{use}} f = f \quad (\text{right identity})$$

$$\text{cobind}_{r \otimes s, t} (g \circ \text{cobind}_{r,s} f) = (\text{cobind}_{s,t} g) \circ (\text{cobind}_{r,s \otimes t} f) \quad (\text{associativity})$$

Rather than defining a single mapping  $C$ , we are now defining a family of mappings  $C^r$  indexed by a monoid structure. Similarly, the  $\text{cobind}_{r,s}$  operation is now formed by a *family* of mappings for different pairs of indices  $r, s$ . To be fully precise,  $\text{cobind}$  is a family of natural transformations and we should include  $\alpha, \beta$  as indices, writing  $\text{cobind}_{r,s}^{\alpha,\beta}$ . For the purpose of this thesis, it is sufficient to omit the superscripts and treat  $\text{cobind}$  just as a family of mappings (rather than natural transformations). When this does not introduce ambiguity, we also occasionally omit the subscripts.

The  $\text{counit}$  operation is not defined for all  $r \in \mathcal{C}$ , but only for the unit  $\text{use}$ . We still include the unit as an index writing  $\text{counit}_{\text{use}}$ , but this is merely for symmetry and as a useful reminder to the reader. Crucially, this means that the operation is defined only for special contexts.

If we look at the indices in the laws, we can see that the left and right identity require  $\text{use}$  to be the unit of  $\otimes$ . Similarly, the associativity law implies the associativity of the  $\otimes$  operator.

**COMPOSITION.** The co-Kleisli category that models sequential composition is formed by the unit arrow (provided by  $\text{counit}$ ) together with the (associative) composition operation that composes computations with contextual requirements as follows:

$$\begin{aligned} - \hat{\circ} - & : (C^r \tau_1 \rightarrow \tau_2) \rightarrow (C^s \tau_2 \rightarrow \tau_3) \rightarrow (C^{r \otimes s} \tau_1 \rightarrow \tau_3) \\ g \hat{\circ} f & = g \circ (\text{cobind}_{r,s} f) \end{aligned}$$

The composition  $\hat{\circ}$  best expresses the intention of indexed comonads. Given two functions with contextual requirements  $r$  and  $s$ , their composition is a function that requires  $r \otimes s$ . The contextual requirements propagate *backwards* and are attached to the input of the composed function.

**EXAMPLES.** Any comonad can be turned into an indexed comonad using a trivial monoid. However, indexed comonads are more general and can be used with other data types, including indexed Maybe.

**Example 6 (Comonads).** Any comonad  $C$  is an indexed comonad with an index provided by a trivial monoid  $(\{1\}, *, 1)$  where  $1 * 1 = 1$ . The mapping  $C^1$  is the mapping  $C$  of the underlying comonad. The operations  $\text{counit}_1$  and  $\text{cobind}_{1,1}$  are defined by the operations  $\text{counit}$  and  $\text{cobind}$  of the comonad.



**Example 7** (Indexed option). *The indexed option comonad is defined over a monoid  $(\{L, D\}, \sqcup, L)$  where  $\sqcup$  is defined as earlier, i.e.  $L = \mathbf{r} \sqcup \mathbf{s} \iff \mathbf{r} = \mathbf{s} = L$ . Assuming  $1$  is the unit type inhabited by  $()$ , the mappings are defined as follows:*

$$\begin{array}{ll} C^L \alpha = \alpha & \text{cobind}_{\mathbf{r}, \mathbf{s}} : (C^{\mathbf{r}} \alpha \rightarrow \beta) \rightarrow (C^{\mathbf{r} \sqcup \mathbf{s}} \alpha \rightarrow C^{\mathbf{s}} \beta) \\ C^D \alpha = 1 & \text{cobind}_{L, L} f \, x = f \, x \\ & \text{cobind}_{L, D} f \, () = () \\ \text{counit}_L : C^L \alpha \rightarrow \alpha & \text{cobind}_{D, L} f \, () = f \, () \\ \text{counit}_L v = v & \text{cobind}_{D, D} f \, () = () \end{array}$$

The *indexed option comonad* models the semantics of the liveness coeffect system discussed in Section 3.2.3, where  $C^L \alpha = \alpha$  models a live context and  $C^D \alpha = 1$  models a dead context which does not contain a value. The counit operation extracts a value from a live context. As in the direct model discussed in Chapter B, the cobind operation can be seen as an implementation of dead code elimination. The definition only evaluates  $f$  when the result is marked as live and is thus required, and it only accesses  $x$  if the function  $f$  requires its input.

The indexed family  $C^{\mathbf{r}}$  in the above example is analogous to the Maybe (or option) data type  $\text{Maybe } \alpha = 1 + \alpha$ . As mentioned earlier, this type does not permit (non-indexed) comonad structure, because  $\text{counit } ()$  is not defined. This is not a problem with indexed comonads, because live contexts are distinguished by the (type-level) coeffect annotation and counit only needs to be defined on live contexts.

**Example 8** (Indexed product). *The semantics of implicit parameters is modelled by an indexed product comonad. We use a monoid  $(\mathcal{P}(\text{Id}), \cup, \emptyset)$  where  $\text{Id}$  is the set of (implicit parameter) names. As previously, all parameters have the type  $\rho$ . The data type  $C^{\mathbf{r}} \alpha = \alpha \times (\mathbf{r} \rightarrow \rho)$  represents a value  $\alpha$  together with a function that associates a parameter value  $\rho$  with every implicit parameter name in  $\mathbf{r} \subseteq \text{Id}$ . The cobind and counit operations are defined as:*

$$\begin{array}{ll} \text{counit}_{\emptyset} : C^{\emptyset} \alpha \rightarrow \alpha & \text{cobind}_{\mathbf{r}, \mathbf{s}} : (C^{\mathbf{r}} \alpha \rightarrow \beta) \rightarrow (C^{\mathbf{r} \cup \mathbf{s}} \alpha \rightarrow C^{\mathbf{s}} \beta) \\ \text{counit}_{\emptyset} (a, g) = a & \text{cobind}_{\mathbf{r}, \mathbf{s}} f (a, g) = (f(a, g|_{\mathbf{r}}), g|_{\mathbf{s}}) \end{array}$$

In the definition, we use the notation  $(a, g)$  for a pair containing the value of type  $\alpha$  together with  $g$ , which is a function  $\mathbf{r} \rightarrow \rho$ . The counit operation takes a value and a function (with empty set as a domain), ignores the function and extracts the value. The cobind operation uses the restriction operation  $g|_{\mathbf{r}}$  to restrict the domain of  $g$  to implicit parameters  $\mathbf{r}$  and  $\mathbf{s}$  in order to get implicit parameters required by the argument of  $f$  and by the resulting computation, respectively (i.e. semantically, it *splits* the available context capabilities). The function  $g$  passed to cobind is initially defined on  $\mathbf{r} \cup \mathbf{s}$  and so the restriction is valid in both cases.

The structure of *indexed comonads* is sufficient to model sequential composition of computations that use a single variable (as discussed in Section 2.2). To model full  $\lambda$ -calculus with lambda abstraction and multiple-variable contexts, we need additional operations introduced in the next section.

#### 4.3.4 Flat indexed comonads

Because of the asymmetry of  $\lambda$ -calculus (discussed in Section 3.1), the duality between monads and comonads can no longer help us with defining the additional structure required to model full  $\lambda$ -calculus. In comonadic compu-

tations, additional information is attached to the context. In application and lambda abstraction, the context is propagated differently than in effectful computations.

To model the effectful  $\lambda$ -calculus, Moggi [55] requires a *strong* monad which has an additional operation  $\text{strength} : \alpha \times M\beta \rightarrow M(\alpha \times \beta)$ . This allows lifting of free variables into an effectful computation. In Haskell, strength can be expressed in the host language and so it is implicit.

To model  $\lambda$ -calculus with contextual properties, Uustalu and Vene [93] require *lax semi-monoidal* comonad. This structure requires an additional monoidal operation:

$$m : C\alpha \times C\beta \rightarrow C(\alpha \times \beta)$$

The  $m$  operation is needed in the semantics of lambda abstraction. It represents merging of contexts and is used to merge the context of the declaration site (containing free variables) and the call site (containing bound variable). For example, for implicit parameters, this combines the additional parameters defined in the two contexts.

The semantics of flat coeffect calculus requires operations for *merging*, but also for *splitting* of contexts. In addition, we also need a lifting operation (similar to  $\iota$  from Definition 7) to model sub-coeffecting.

**Definition 6.** Given a flat coeffect algebra  $(\mathcal{C}, \otimes, \oplus, \wedge, \text{use}, \text{ign}, \leq)$ , a flat indexed comonad is an indexed comonad over the monoid  $(\mathcal{C}, \otimes, \text{use})$  equipped with families of operations  $\text{merge}_{\mathbf{r}, \mathbf{s}}$ ,  $\text{split}_{\mathbf{r}, \mathbf{s}}$  and  $\text{lift}_{\mathbf{r}', \mathbf{r}}$  where:

- $\text{merge}_{\mathbf{r}, \mathbf{s}}$  is a family of mappings  $C^{\mathbf{r}}\alpha \times C^{\mathbf{s}}\beta \rightarrow C^{\mathbf{r} \wedge \mathbf{s}}(\alpha \times \beta)$
- $\text{split}_{\mathbf{r}, \mathbf{s}}$  is a family of mappings  $C^{\mathbf{r} \oplus \mathbf{s}}(\alpha \times \beta) \rightarrow C^{\mathbf{r}}\alpha \times C^{\mathbf{s}}\beta$
- $\text{lift}_{\mathbf{r}', \mathbf{r}}$  is a family of mappings  $C^{\mathbf{r}'}\alpha \rightarrow C^{\mathbf{r}}\alpha$  for all  $\mathbf{r}', \mathbf{r}$  such that  $\mathbf{r} \leq \mathbf{r}'$

The  $\text{merge}_{\mathbf{r}, \mathbf{s}}$  operation is the most interesting one. Given two comonadic values with additional contexts specified by  $\mathbf{r}$  and  $\mathbf{s}$ , it combines them into a single value with additional context  $\mathbf{r} \wedge \mathbf{s}$ . The  $\wedge$  operation often represents *greatest lower bound*<sup>2</sup>, elucidating the fact that merging may result in the loss of some parts of the contexts  $\mathbf{r}$  and  $\mathbf{s}$ . We look at examples of this operation in the next section.

The  $\text{split}_{\mathbf{r}, \mathbf{s}}$  operation splits a single comonadic value (containing a tuple) into two separate values. Note that this does not simply duplicate the value, because the additional context is also split. To obtain coeffects  $\mathbf{r}$  and  $\mathbf{s}$ , the input needs to provide *at least*  $\mathbf{r}$  and  $\mathbf{s}$ , so the tags are combined using the  $\oplus$ , which is often the *least upper-bound*<sup>1</sup>.

Finally,  $\text{lift}_{\mathbf{r}', \mathbf{r}}$  is a family of operations that “forget” some part of a context. This models the sub-coeffecting operation and lets us, for example, forget some of the available implicit parameters, or turn a live context (containing a value) into a dead context (empty).

**ALTERNATIVE DEFINITION.** Although we do not demand this as a general law, in all our systems, it is the case that  $\mathbf{r} \leq \mathbf{r} \oplus \mathbf{s}$  and  $\mathbf{s} \leq \mathbf{r} \oplus \mathbf{s}$ . This special case allows a simpler definition of *indexed flat comonad* by expressing the split operation in terms of lifting (sub-coeffecting) as follows:

$$\begin{aligned} \text{map}_{\mathbf{r}} f &= \text{cbind}_{\mathbf{r}, \mathbf{r}} (f \circ \text{counit}_{\text{use}}) \\ \text{split}_{\mathbf{r}, \mathbf{s}} c &= (\text{map}_{\mathbf{r}} \text{fst} (\text{lift}_{\mathbf{r} \oplus \mathbf{s}, \mathbf{r}} c), \text{map}_{\mathbf{s}} \text{snd} (\text{lift}_{\mathbf{r} \oplus \mathbf{s}, \mathbf{s}} c)) \end{aligned}$$

<sup>2</sup> The  $\wedge$  and  $\oplus$  operations are the greatest and least upper bounds in the liveness and data-flow examples, but not for implicit parameters. However, they remain useful as an informal analogy.

The  $\text{map}_r$  operation is the mapping on arrows that corresponds to the object mapping  $C^r$ . The definition is dual to the standard definition of  $\text{map}$  for monads in terms of  $\text{bind}$  and  $\text{unit}$ . The functions  $\text{fst}$  and  $\text{snd}$  are first and second projections from a two-element pair. To define the  $\text{split}_{r,s}$  operation, we use the argument  $c$  twice, use lifting to throw away additional parts of the context and then transform the values in the context.

This alternative is valid for our examples, but we do not use it for two reasons. Firstly, this would be the only place where our semantics uses a variable *twice* (in this case  $c$ ). Thus using an explicit split means that the structure required by our semantics does not need to provide variable duplication and our model could be embedded in linear or affine category. Secondly, explicit split is similar to the definition that is needed for structural coefficients in Chapter 5 and so it makes the connection between the two system easier to see.

**EXAMPLES.** All the examples of *indexed comonads* discussed in Section 4.3.3 can be extended into *flat indexed comonads*. Note that this is not a *general* statement, because each example requires us to define additional operations, specific for the example.

**Example 9** (Monoidal comonads). *Just like indexed comonads generalise comonads, the additional structure of flat indexed comonads generalises symmetric semimonoidal comonads of Uustalu and Vene [93]. The flat coefficient algebra is defined as  $(\{1\}, *, *, *, 1, 1, =)$  where  $1 * 1 = 1$  and  $1 = 1$ . The additional operation  $\text{merge}_{1,1}$  is provided by the monoidal operation called  $m$  by Uustalu and Vene. The  $\text{split}_{1,1}$  operation is defined by duplication and  $\text{lift}_{1,1}$  is the identity function.*

**Example 10** (Indexed option). *Flat coefficient algebra for liveness defines  $\oplus$  and  $\wedge$  as  $\sqcup$  and  $\sqcap$ , respectively and specifies that  $D \sqsubseteq L$ . Recall also that the object mapping is defined as  $C^L \alpha = \alpha$  and  $C^D \alpha = 1$ . The additional operations of a flat indexed comonad are defined as follows:*

$$\begin{array}{lll} \text{merge}_{L,L} (a, b) = (a, b) & \text{split}_{L,L} (a, b) = (a, b) & \text{lift}_{L,D} v = () \\ \text{merge}_{L,D} (a, ()) = () & \text{split}_{L,D} (a, b) = (a, ()) & \text{lift}_{L,L} v = v \\ \text{merge}_{D,L} ((), b) = () & \text{split}_{D,L} (a, b) = ((), b) & \text{lift}_{D,D} () = () \\ \text{merge}_{D,D} ((), ()) = () & \text{split}_{D,D} () = ((), ()) & \end{array}$$

Without the indexing, the merge operations implements *zip* on option values, returning an option only when both values are present. The behaviour of the split operation is partly determined by the indices. When the input is *dead*, both values have to be dead (this is also the only solution of  $D = r \sqcap s$ ), but when the input is *live*, the operation can perform implicit sub-coeffecting and drop one of the values.

Explicit sub-coeffecting using the (*sub*) rule is modelled by the lift operation. This can turn a *live* value  $v$  into a dead value  $()$ , or it can behave as identity. The behaviour is, again, determined by the index.

**Example 11** (Indexed product). *For implicit parameters, both  $\wedge$  and  $\oplus$  are the  $\cup$  operation and the relation  $\leq$  is formed by the subset relation  $\subseteq$ . Recall that the data type  $C^r \alpha$  is  $\alpha \times (r \rightarrow \rho)$  where  $\rho$  is the type of implicit parameter values. The additional operations are defined as:*

$$\begin{array}{ll} \text{split}_{r,s} ((a, b), g) = ((a, g|_r), (b, g|_s)) & \text{where } f \uplus g = \\ \text{merge}_{r,s} ((a, f), (b, g)) = ((a, b), f \uplus g) & f|_{\text{dom}(f) \setminus \text{dom}(g)} \cup g \\ \text{lift}_{r',r} (a, g) = (a, g|_r) & \end{array}$$

The split operation splits the tuple and restricts the function (representing available implicit parameters) to the required sub-sets. This corresponds to the definition in terms of lift, which performs just the restriction. The merge operation is more interesting. It uses the  $\uplus$  operation that we defined when introducing implicit parameters in Section 3.2.1. It merges the values, preferring the definitions from the right-hand side (call site) over left-hand side (declaration site). Thus the operation is not symmetric.

**Example 12** (Indexed list). *Our last example provides the semantics of data-flow computations. The flat coeffect algebra is formed by  $(\mathbb{N}, +, \max, \min, 0, 0, \leq)$ . In a non-indexed version, the semantics is provided by a non-empty list. In the indexed semantics, the index represents the number of available past values. The data type is then a pair of the current value, followed by  $n$  past values. The mappings that form the flat indexed comonad are defined as follows:*

$$\begin{aligned}
 \text{counit}_0 \langle a_0 \rangle &= a_0 & C^n \alpha &= \underbrace{\alpha \times \dots \times \alpha}_{(n+1)\text{-times}} \\
 \text{cobind}_{m,n} f \langle a_0, \dots, a_{m+n} \rangle &= \langle f \langle a_0, \dots, a_m \rangle, \dots, f \langle a_n, \dots, a_{m+n} \rangle \rangle \\
 \text{merge}_{m,n} (\langle a_0, \dots, a_m \rangle, \langle b_0, \dots, b_n \rangle) &= \langle (a_0, b_0), \dots, (a_{\min(m,n)}, b_{\min(m,n)}) \rangle \\
 \text{split}_{m,n} (\langle a_0, b_0 \rangle, \dots, \langle a_{\max(m,n)}, b_{\max(m,n)} \rangle) &= \langle \langle a_0, \dots, a_m \rangle, \langle b_0, \dots, b_n \rangle \rangle \\
 \text{lift}_{n',n} \langle a_0, \dots, a_{n'} \rangle &= \langle a_0, \dots, a_n \rangle \quad (\text{when } n \leq n')
 \end{aligned}$$

The reader is invited to check that the number of required past elements in each of the mappings matches the number specified by the indices. The index specifies the number of *past* elements and so the list always contains at least one value. Thus counit returns the element of a singleton list.

The  $\text{cobind}_{m,n}$  operation requires  $m + n$  elements in order to generate  $n$  past results of the  $f$  function, which itself requires  $m$  past values. When combining two lists,  $\text{merge}_{m,n}$  behaves as *zip* and produces a list that has the length of the shorter argument. When splitting a list,  $\text{split}_{m,n}$  needs the maximum of the required lengths. Finally, the lifting operation just drops some number of elements from a list.

#### 4.3.5 Properties and related notions

The flat indexed comonad structure is all we need to give the semantics of the flat coeffect calculus. Before doing so in Section 4.3.6, we briefly consider additional properties and other categorical structures that have been proposed mainly in the context of monads and effects and we look how they relate to indexed comonads.

**SHAPE PRESERVATION.** Ordinary comonads have the *shape preservation* property [65]. Intuitively, this means that the core comonad structure does not provide a way of modeling computations where the additional context changes during the computation. For example, in the NEList comonad, the length of the list stays the same after applying cobind.

Indexed comonads are not restricted by this property of comonads. For example, given the indexed product comonad, in the computation  $\text{cobind}_{\mathbf{r},\mathbf{s}} f$ , the shape of the context changes from providing implicit parameters  $\mathbf{r} \cup \mathbf{s}$  to providing just implicit parameters  $\mathbf{s}$ .

**FAMILIES OF MONADS.** When linking effect systems and monads, Wadler and Thiemann [55] propose a *family of monads* as the categorical structure. The dual structure, *family of comonads*, is defined as follows.

**Definition 7.** A family of comonads is formed by triples  $(C^{\mathbf{r}}, \text{cobind}_{\mathbf{r}}, \text{count}_{\mathbf{r}})$  for all  $\mathbf{r}$  such that each triple forms a comonad. Given  $\mathbf{r}, \mathbf{r}'$  such that  $\mathbf{r} \leq \mathbf{r}'$ , there is also a mapping  $\iota_{\mathbf{r}',\mathbf{r}} : C^{\mathbf{r}'} \rightarrow C^{\mathbf{r}}$  satisfying certain coherence conditions.

A family of comonads is more restrictive than an indexed comonad, because each of the data types needs to form a comonad separately. For example, our indexed option does not form a family of comonads (again, because  $\text{count}$  is not defined on  $C^{\text{D}} \alpha = 1$ ). However, given a family of comonads and indices such that  $\mathbf{r} \leq \mathbf{r} \otimes \mathbf{s}$ , we can define an indexed comonad. Briefly, to define  $\text{cobind}_{\mathbf{r},\mathbf{s}}$  of an indexed comonad, we use  $\text{cobind}_{\mathbf{r} \otimes \mathbf{s}}$  from the family, together with two lifting operations:  $\iota_{\mathbf{r} \otimes \mathbf{s},\mathbf{r}}$  and  $\iota_{\mathbf{r} \otimes \mathbf{s},\mathbf{s}}$ .

**PARAMETRIC EFFECT MONADS.** Parametric effect monads introduced by Katsumata [45] (independently to our indexed comonads) are closely related to our definition. Although presented in a more general categorical framework (and using monads), the model defines the unit operation only on the unit of a monoid and the bind operation composes effect annotations using the provided monoidal structure.

#### 4.3.6 Semantics of flat calculus

In Section 3.2, we defined the semantics of concrete (flat) context-dependent computations including implicit parameters, liveness and data-flow. Using the *flat indexed comonad* structure, we can now define a single uniform semantics that is capable of capturing all our examples, as well as other computations that can be modelled by the structure.

**CONTEXTS AND FUNCTIONS.** The modelling of contexts and functions generalizes the earlier concrete examples. We use the family of mappings  $C^{\mathbf{r}}$  as an (indexed) data-type that wraps the product of free variables of the context and the arguments of functions:

$$\begin{aligned} \llbracket x_1 : \tau_1, \dots, x_n : \tau_n @ \mathbf{r} \vdash e : \tau \rrbracket & : C^{\mathbf{r}}(\tau_1 \times \dots \times \tau_n) \rightarrow \tau \\ \llbracket \tau_1 \xrightarrow{\mathbf{r}} \tau_2 \rrbracket & = C^{\mathbf{r}}\tau_1 \rightarrow \tau_2 \end{aligned}$$

**EXPRESSIONS.** The definition of the semantics is shown in Figure 23. For readability, we write the definitions in a simple programming language notation as opposed to the point-free categorical style. However, it can be equally written using just the operations of flat indexed comonad together with  $i^{\text{th}}$  projection from a tuple represented by  $\pi_i$ , *curry* and *uncurry*, function composition, value duplication ( $\Delta : A \rightarrow A \times A$ ) and function pairing (given  $f : A \rightarrow B$  and  $g : C \rightarrow D$  then  $f \times g : A \times C \rightarrow B \times D$ ). These operations can be provided by e. g. a Cartesian-closed category.

The semantics of variable access and abstraction are the same as in the semantics of Uustalu and Vene [93], modulo the indexing. The semantics of

$$\begin{aligned}
\llbracket \Gamma @ \text{use} \vdash x_i : \tau_i \rrbracket ctx &= \pi_i (\text{counit}_{\text{use}} ctx) & (var) \\
\llbracket \Gamma @ \text{ign} \vdash c_i : \tau \rrbracket ctx &= \delta (c_i) & (const) \\
\llbracket \Gamma @ r \vdash \lambda x. e : \tau_1 \xrightarrow{s} \tau_2 \rrbracket ctx &= \lambda v. & (abs) \\
&\quad \llbracket \Gamma, x : \tau_1 @ r \wedge s \vdash e : \tau_2 \rrbracket (\text{merge}_{r,s} (ctx, v)) \\
\llbracket \Gamma @ r \oplus (s \otimes t) \vdash e_1 e_2 : \tau_2 \rrbracket ctx &= & (app) \\
&\quad \text{let } (ctx_1, ctx_2) = \text{split}_{r,s \otimes t} (\text{map}_{r \oplus (s \otimes t)} (\lambda x. (x, x)) ctx) \\
&\quad \text{in } \llbracket \Gamma @ r \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \rrbracket ctx_1 (\text{cobind}_{s,t} \llbracket \Gamma @ s \vdash e_2 : \tau_1 \rrbracket ctx_2) \\
\llbracket \Gamma @ r \vdash e : \tau \rrbracket ctx &= & (sub) \\
&\quad \llbracket \Gamma @ r' \vdash e : \tau \rrbracket (\text{lift}_{r,r'} ctx) \quad (\text{when } r' \leq r)
\end{aligned}$$

Figure 23: Categorical semantics of the flat coeffect calculus

variable access (*var*) uses  $\text{counit}_{\text{use}}$  to extract product of free-variables from the context and then projection  $\pi_i$  to obtain the variable value. Abstraction (*abs*) takes the context  $ctx$  and function argument  $v$  and merges their additional contexts using  $\text{merge}_{r,s}$ . Assuming the context  $\Gamma$  contains variables of types  $\sigma_1, \dots, \sigma_n$ , this gives us a value  $C^{r \wedge s}((\sigma_1 \times \dots \times \sigma_n) \times \tau_1)$ . Assuming that  $n$ -element tuples are associated to the left, the wrapped context is equivalent to  $\sigma_1 \times \dots \times \sigma_n \times \tau_1$ , which can then be passed to the body of the function.

The semantics of application is more complex. It first duplicates the free-variable product inside the context (using  $\text{map}_r$  and duplication). Then it splits this context using  $\text{split}_{r,s \otimes t}$ . The two contexts contain the same variables (as required by sub-expressions  $e_1$  and  $e_2$ ), but different coeffect annotations. The first context (with index  $r$ ) is used to evaluate  $e_1$ , resulting in a function  $C^t \tau_1 \rightarrow \tau_2$ . To obtain the result, we compose this with a function created by applying  $\text{cobind}_{s,t}$  on the semantics of sub-expression  $e_2$ , which is of type  $C^{s \otimes t} \sigma_1 \times \dots \times \sigma_n \rightarrow C^t \tau_1$ .

Finally, constants (*const*) are modelled by a global dictionary  $\delta$  and sub-coeffecting is interpreted by dropping additional context from the provided context  $ctx$  using  $\text{lift}_{r,r'}$  and providing it to the semantics of the assumption.

**PROPERTIES.** The categorical semantics can be used to embed context-dependent computations in functional programming languages, similarly to how monads provide a way of embedding effectful computations. More importantly, it also provides validation for the design of the type system developed in Section 4.2.4. As stated in the following theorem, the annotations in the type system match those of the semantic functions.

**Theorem 1** (Correspondence). *In all of the typing rules of the flat coeffect system, the context annotations  $r$  of typing judgements  $\Gamma @ r \vdash e : \tau$  and function types  $\tau_1 \xrightarrow{r} \tau_2$  correspond to the indices of mappings  $C^r$  in the corresponding semantic function defined by  $\llbracket \Gamma @ r \vdash e : \tau \rrbracket$ .*

*Proof.* By analysis of the semantic rules in Figure 23. □

Thanks to indexing, the statement of the theorem is significantly stronger than for a non-indexed system, because it provides the justification for our



choice of indices in the typing rules. In particular, we can see that the annotations follow from the annotations on primitive functions that define the semantics. Also, each function defining the semantics uses a distinct operation of the coefficient algebra and so the type system is the most general possible definition (within the comonadic framework we use).

Although the notion of indexed comonads presented in this section is novel and interesting in its own, the main reason for introducing it is to motivate the flat coefficient type system. This is captured by the Theorem 1. In the next section, we shift our focus from (categorical) semantics to syntactic properties of the calculus.

#### 4.4 SYNTACTIC EQUATIONAL THEORY

Each of the concrete coefficient calculi discussed in this chapter has a different notion of context, much like various effectful languages have different notions of effects (such as exceptions or mutable state). However, in all of the calculi, the context has a number of common properties that are captured by the *flat coefficient algebra*. This means that there are equational properties that hold for all of the coefficient systems. Further properties hold for systems where the context satisfies additional properties.

In this section, we look at such shared syntactic properties. This accompanies the previous section, which provided a *semantic* justification for the axioms of coefficient algebra with a *syntactic* justification. Operationally, this section can also be viewed as providing a pathway to an operational semantics for two of our systems (implicit parameters and liveness), which can be based on syntactic substitution. As we discuss later, the notion of context for data-flow is more complex.

##### 4.4.1 Syntactic properties

Before discussing the syntactic properties of general coefficient calculus formally, it should be clarified what is meant by providing “pathway to operational semantics” in this section. We do that by contrasting syntactic properties of coefficient systems with more familiar effect systems. Assuming  $e_1[x \leftarrow e_2]$  is a standard capture-avoiding syntactic substitution, the following equations define four syntactic reductions on the terms:

$$\begin{array}{lll}
 (\lambda x. e_1) e_2 & \longrightarrow_{\text{cbn}} & e_1[x \leftarrow e_2] & (\text{call-by-name}) \\
 (\lambda x. e_1) v & \longrightarrow_{\text{cbv}} & e_1[x \leftarrow v] & (\text{call-by-value}) \\
 (\lambda x. e_1) e_2 & \longrightarrow_{\text{seq}} & \text{glet } x = e_2 \text{ in } e_1 & (\text{internalized sequencing}) \\
 e & \longrightarrow_{\eta} & \lambda x. e \ x & (\eta\text{-expansion})
 \end{array}$$

The rules capture syntactic reductions that can be performed in a general calculus, without any knowledge of the specific notion of context. If the reductions preserve the type of the expression (type preservation), then operational semantics can be defined as a repeated application of the rules, until a specified normal form (i. e. a value) is reached.

The first two equations define well-known call-by-name and call-by-value reductions. The **glet** reduction (inspired by Filinski [27]) models a system where program is reduced to a sequence of primitive operations, sequentially composed using the special **glet** construct. In the rest of the section, we briefly outline the interpretation of the four rules and then we focus on call-by-value (Section 4.4.2) and call-by-name (Section 4.4.3) in more details.

The focus of our work is on the general coeffect system and so we do not discuss the operational semantics of the specific notions of context. However, some work in that area has been done by Brunel et al. [15].

**CALL-BY-NAME.** In call-by-name, the argument is syntactically substituted for all occurrences of a variable. It can be used as the basis for operational semantics of purely functional languages. However, using the rule in effectful languages breaks the *type preservation* property. For example, consider a language with effect system where functions are annotated with sets of effects such as {write}. A function  $\lambda x.y$  is effect-free:

$$y : \tau_1 \vdash \lambda x.y : \tau_1 \xrightarrow{\emptyset} \tau_2 \ \& \ \emptyset$$

Substituting an expression  $e$  with effects {write} for  $y$  changes the type of the function by adding latent effects (without changing the immediate effects):

$$\vdash \lambda x.e : \tau_1 \xrightarrow{\{\text{write}\}} \tau_2 \ \& \ \emptyset$$

Similarly to effect systems, substituting a context-dependent computation  $e$  for a variable  $y$  can add latent coeffects to the function type. However, this is not the case for *all* flat coeffect calculi. For example, call-by-name reduction preserves types and coeffects for the implicit parameters system. This means that certain coeffect systems support call-by-name evaluation strategy and could be embedded in purely functional language (such as Haskell).

**CALL-BY-VALUE.** The call-by-value evaluation strategy is often used by effectful languages. Here, an argument is first reduced to a *value* before performing the substitution. In effectful languages, value is defined syntactically. For example, in the *Effect* language [103], values are identifiers  $x$  or functions  $(\lambda x.e)$ .

The notion of *value* in coeffect systems differs from the usual syntactic understanding. A function  $(\lambda x.e)$  does not defer all context requirements of the body  $e$  and may have immediate context requirements. Thus we say that  $e$  is a value if it is a value in the usual sense *and* has not immediate context requirements. We define this formally in Section 4.4.2.

The call-by-value evaluation strategy preserves typing for a wide range of flat coeffect calculi, including all our three examples. However, it is rather weak – in order to use it, the concrete semantics needs to provide a way for reducing context-dependent term  $\Gamma @ \mathbf{r} \vdash e : \tau$  to a value, i.e. a term  $\Gamma @ \mathbf{use} \vdash e' : \tau$  with no context requirements.

**INTERNALIZED SEQUENCING.** The (*internalized sequencing*) rule captures an operational semantics where the language provides a construct representing *sequential composition* and expressions can be reduced to a normal form, consisting of a sequence of context-dependent operations. We choose to write the sequencing operation as **glet** to emphasize that this is a separate primitive and not an ordinary syntactic **let**. The normal form looks as follows:

$$\mathbf{glet} \ x_1 = e_1 \ \mathbf{in} \ \dots \ \mathbf{glet} \ x_n = e_n \ \mathbf{in} \ e$$

Here, the expressions  $e_1, \dots, e_n, e$  do not contain further nested **glet** constructs. This evaluation strategy provides context-dependent counterpart to operational view of monads developed by Filinsky [27]. We discuss how expressions reduce to the normal form in Section 4.4.4



The (*internalized sequencing*) rule is useful when defining a concrete semantics for a language that provides constructs for explicitly providing the context. For example, consider a language with implicit parameters where a parameter is defined by `letdyn ?p = e1 in e2`. Semantics for such language would provide a reduction rule:

$$\text{letdyn } ?p = e_1 \text{ in } (\text{glet } x = ?p \text{ in } e) \rightsquigarrow \text{let } x = e_1 \text{ in } e$$

The `glet` construct provides a way of sequentialising the context-requirements so that they can be discharged by matching constructs providing the required context. As discussed earlier, we focus on the general case and so we discuss when a flat coefficient calculus supports this form of evaluation (rather than looking at semantics for concrete systems).

**LOCAL SOUNDNESS AND COMPLETENESS.** Two desirable properties of calculi, coined by Pfenning and Davies [73], are *local soundness* and *local completeness*. They guarantee that the rules which introduce a function arrow (lambda abstraction) and eliminate it (application) are sufficiently strong, but not too strong.

The local soundness property is witnessed by (call-by-name)  $\beta$ -reduction, which we discussed already. The local completeness is witnessed by the  $\eta$ -expansion rule. We discuss the flat coefficient algebra conditions under which the reduction holds in Section 4.4.3.

#### 4.4.2 Call-by-value evaluation

As discussed in the previous section, call-by-value reduction can be used for most flat coefficient calculi, but it provides a very weak general model i.e. the hard work of reducing context-dependent term to a *value* has to be provided for each system. Syntactic values are defined in the usual way:

$$\begin{aligned} v \in \text{SynVal} \quad v &::= x \mid c \mid (\lambda x.e) \\ n \in \text{NonVal} \quad n &::= e_1 \ e_2 \mid \text{let } x = e_1 \text{ in } e_2 \\ e \in \text{Expr} \quad e &::= v \mid n \end{aligned}$$

The syntactic form *SynVal* captures syntactic values, but a context-dependency-free value in coefficient calculus cannot be defined purely syntactically, because a function  $(\lambda x.e)$  does not automatically defer all context requirements.

**Definition 8.** An expression  $e$  is a value, written as  $\text{val}(e)$  if it is a syntactic value, i.e.  $e \in \text{SynVal}$  and it has no context-dependencies, i.e.  $\Gamma @ \text{use} \vdash e : \tau$ .

The call-by-value substitution substitutes a value, with context requirements `use`, for a variable, whose access is also annotated with `use`. Thus, it does not affect the type and context requirements of the term:

**Lemma 2** (Call-by-value substitution). *In a flat coefficient calculus with a coefficient algebra  $(\mathcal{C}, \otimes, \oplus, \wedge, \text{use}, \text{ign}, \leq)$ , given a value  $\Gamma @ \text{use} \vdash v : \sigma$  and an expression  $\Gamma, x : \sigma @ \text{r} \vdash e : \tau$ , then substituting  $v$  for  $x$  does not change the type and context requirements, that is  $\Gamma @ \text{r} \vdash e[x \leftarrow v] : \tau$ .*

*Proof.* By induction over the type derivation, using the fact that  $x$  and  $v$  are annotated with `use` and that  $\Gamma$  is treated as a set in the flat calculus.  $\square$

The substitution lemma 2 holds for all flat coefficient systems. However, proving that call-by-value reduction preserves typing requires an additional constraint on the flat coefficient algebra, which relates the  $\wedge$  and  $\oplus$  operations. This is captured by the (*approximation*) property:

$$\mathbf{r} \wedge \mathbf{t} \leq \mathbf{r} \oplus \mathbf{t} \quad (\text{approximation})$$

Intuitively, this specifies that the  $\wedge$  operation (splitting of context requirements) under-approximates the actual context capabilities while the  $\oplus$  operation (combining of context requirements) over-approximates the actual context requirements.

The property holds for the three systems we consider – for implicit parameters, this is an equality; for liveness and data-flow (which both use a lattice), the greatest lower bound is smaller than the least upper bound.

Assuming  $\rightarrow_{\text{cbv}}$  is call-by-value reduction that reduces the term  $(\lambda x.e) v$  to a term  $e[x \leftarrow v]$ , the type preservation theorem is stated as follows:

**Theorem 3** (Type preservation for call-by-value). *In a flat coeffect system with the (approximation) property, if  $\Gamma @ \mathbf{r} \vdash e : \tau$  and  $e \rightarrow_{\text{cbv}} e'$  then  $\Gamma @ \mathbf{r} \vdash e' : \tau$ .*

*Proof.* Consider the typing derivation for the term  $(\lambda x.e) v$  before reduction:

$$\frac{\frac{\frac{\Gamma, x : \tau_1 @ \mathbf{r} \wedge \mathbf{t} \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x.e : \tau_1 \xrightarrow{\mathbf{t}} \tau_2} \quad \Gamma @ \mathbf{use} \vdash v : \tau_1}{\Gamma @ \mathbf{r} \oplus (\mathbf{use} \otimes \mathbf{t}) \vdash (\lambda x.e) v : \tau_2}}{\Gamma @ \mathbf{r} \oplus \mathbf{t} \vdash (\lambda x.e) v : \tau_2}$$

The final step simplifies the coeffect annotation using the fact that  $\mathbf{use}$  is a unit of  $\otimes$ . From Lemma 2,  $e[x \leftarrow v]$  has the same coeffect annotation as  $e$ . As  $\mathbf{r} \wedge \mathbf{t} \leq \mathbf{r} \oplus \mathbf{t}$ , we can apply sub-coeffecting:

$$(\text{sub}) \quad \frac{\Gamma @ \mathbf{r} \wedge \mathbf{t} \vdash e[x \leftarrow v] : \tau_2}{\Gamma @ \mathbf{r} \oplus \mathbf{t} \vdash e[x \leftarrow v] : \tau_2}$$

Comparing the final conclusions of the above two typing derivations shows that the reduction preserves type and coeffect annotation.  $\square$

#### 4.4.3 Call-by-name evaluation

When reducing the expression  $(\lambda x.e_1) e_2$  using the call-by-name strategy, the sub-expression  $e_2$  is substituted for all occurrences of the variable  $v$  in an expression  $e_1$ . As discussed in Section 4.4.1, the call-by-name strategy does not *in general* preserve the type of a terms in coeffect calculi, but it does preserve the typing in two interesting cases.

The typing is preserved for different reasons in two of our systems, so we briefly review the concrete examples. Then, we prove the substitution lemma for two special cases of flat coeffects (Lemma 4 and Lemma 5) and finally, we state the conditions under which typing preservation hold for flat coeffect calculi (Theorem 6).

**DATA-FLOW.** The type preservation property does not hold for data-flow. This case is similar to the example shown earlier with effectful computations. As a minimal example, consider the substitution of a context-dependent expression  $\mathbf{prev} z$  for a variable  $y$  in a function  $\lambda x.y$ :

$$\begin{array}{ll} y : \tau_1, z : \tau_1 @ 0 \vdash \lambda x.y : \tau_1 \xrightarrow{0} \tau_2 & (\text{before}) \\ z : \tau_1 @ 1 \vdash \lambda x.\mathbf{prev} z : \tau_1 \xrightarrow{1} \tau_2 & (\text{after}) \end{array}$$

After the substitution, the coeffect of the body is 1. The rule for lambda abstraction requires that  $1 = \min(r, s)$  and so the least solution is to set both  $r, s$  to 1. The substitution this affects the coeffects attached both to the function type and the overall context.

Semantically, the coeffect over-approximates the actual requirements – at run-time, the code does not actually access a previous value of the argument  $x$ . This cannot be captured by a flat coeffect system, but it can be captured using the structural system discussed in Chapter 5.

**IMPLICIT PARAMETERS.** In data-flow, there is no typing for the resulting expression that preserves the type of the function. However, this is not the case for all systems. Consider substituting an implicit parameter access  $?p$  for a free variable  $y$  under a lambda:

$$\begin{aligned} y : \tau_1 @ \emptyset \vdash \lambda x. y : \tau_1 &\xrightarrow{\emptyset} \tau_2 & (\text{before}) \\ \emptyset @ \{?p\} \vdash \lambda x. ?p : \tau_1 &\xrightarrow{\emptyset} \tau_2 & (\text{after}) \end{aligned}$$

The above shows one possible typing of the body – one that does not change the coeffects of the function type and attaches all additional coeffects (implicit parameters) to the context. In case of implicit parameters (and, more generally, systems with set-like annotations) this is always possible.

**LIVENESS.** In liveness, the type preservation also holds, but for a different reason. Consider substituting an arbitrary expression  $e$  of type  $\tau_1$  with coeffects  $r$  for a variable  $y$ :

$$\begin{aligned} y : \tau_1 @ L \vdash \lambda x. y : \tau_1 &\xrightarrow{L} \tau_2 & (\text{before}) \\ \emptyset @ L \vdash \lambda x. e : \tau_1 &\xrightarrow{L} \tau_2 & (\text{after}) \end{aligned}$$

In the original expression, both the overall context and the function type are annotated with  $L$ , because the body contains a variable access. An expression  $e$  can always be treated as being annotated with  $L$  (because  $L$  is the top element of the lattice) and so we can also treat  $e$  as being annotated with coeffects  $L$ . As a result, substitution does not change the coeffect.

**REDUCTION THEOREM.** The above examples (implicit parameters and liveness) demonstrate two particular kinds of coeffect algebra for which typing preservation holds. Proving the type preservation separately provides more insight into how the systems work. We consider the two cases separately, but find a more general formulation for both of them.

**Definition 9.** We call a flat coeffect algebra top-pointed if  $\text{use}$  is the greatest (top) coeffect scalar ( $\forall r \in \mathcal{C} . r \leq \text{use}$ ) and bottom-pointed if it is the smallest (bottom) element ( $\forall r \in \mathcal{C} . r \geq \text{use}$ ).

Liveness is an example of top-pointed coeffects as variables are annotated with  $L$  and  $D \leq L$ , while implicit parameters and data-flow are examples of bottom-pointed coeffects. For top-pointed flat coeffects, the substitution lemma holds without additional requirements:

**Lemma 4** (Top-pointed substitution). *In a top-pointed flat coeffect calculus with an algebra  $(\mathcal{C}, *, \oplus, \wedge, \text{use}, \text{ign}, \leq)$ , substituting an expression  $e_s$  with arbitrary coeffects  $s$  for a variable  $x$  in  $e_r$  does not change the coeffects of  $e_r$ :*

$$\begin{aligned} \Gamma @ s \vdash e_s : \tau_s \wedge \Gamma_1, x : \tau_s, \Gamma_2 @ r \vdash e_r : \tau_r \\ \Rightarrow \Gamma_1, \Gamma, \Gamma_2 @ r \vdash e_r[x \leftarrow e_s] : \tau_r \end{aligned}$$

*Proof.* Using sub-coffecting ( $s \leq \text{use}$ ) and a variation of Lemma 2.  $\square$

As variables are annotated with the top element  $\text{use}$ , we can substitute the term  $e_s$  for any variable and use sub-coffecting to get the original typing (because  $s \leq \text{use}$ ).

In a bottom pointed coeffect system, substituting  $e$  for  $x$  increases the context requirements. However, if the system satisfies the strong condition that  $\wedge = \otimes = \oplus$  then the context requirements arising from the substitution can be associated with the context  $\Gamma$ , leaving the context requirements of a function value unchanged. As a result, substitution does not break soundness as in effect systems. The requirement  $\wedge = \otimes = \oplus$  holds for our implicit parameters example (all three operators are a set union) and for other set-like coeffects. It allows the following substitution lemma:

**Lemma 5** (Bottom-pointed substitution). *In a bottom-pointed flat coeffect calculus with an algebra  $(\mathcal{C}, \otimes, \oplus, \wedge, \text{use}, \text{ign}, \leq)$  where  $\wedge = \otimes = \oplus$  is an idempotent and commutative operation and  $r \leq r' \Rightarrow \forall s. r \otimes s \leq r' \otimes s$  then:*

$$\begin{aligned} \Gamma @ s \vdash e_s : \tau_s \quad \wedge \quad \Gamma_1, x : \tau_s, \Gamma_2 @ r \vdash e_r : \tau_r \\ \Rightarrow \quad \Gamma_1, \Gamma, \Gamma_2 @ r \otimes s \vdash e_r[x \leftarrow e_s] : \tau_r \end{aligned}$$

*Proof.* By induction over  $\vdash$ , using the idempotent, commutative monoid structure to keep  $s$  with the free-variable context. See Appendix B.1.  $\square$

The flat system discussed here is *flexible enough* to let us always re-associate new context requirements (arising from the substitution) with the free-variable context. In contrast, the structural system discussed in Chapter 5 is *precise enough* to keep the coeffects associated with individual variables, thus preserving typing in a complementary way.

The two substitution lemmas discussed above show that the call-by-name evaluation strategy can be used for certain coeffect calculi, including liveness and implicit parameters. Assuming  $\rightarrow_{\text{cbn}}$  is the standard call-by-name reduction, the following theorem holds:

**Theorem 6** (Type preservation for call-by-name). *In a coeffect system that satisfies the conditions for Lemma 4 or Lemma 5, if  $\Gamma @ r \vdash e : \tau$  and  $e \rightarrow_{\text{cbn}} e'$  then it is also the case that  $\Gamma @ r \vdash e' : \tau$ .*

*Proof.* For top-pointed coeffect algebra (using Lemma 4), the proof is similar to the one in Theorem 3, using the facts that  $s \leq \text{use}$  and  $r \wedge t = r \oplus t$ . For bottom-pointed coeffect algebra, consider the typing derivation for the term  $(\lambda x. e_r) e_s$  before reduction:

$$\frac{\frac{\Gamma, x : \tau_s @ r \vdash e_r : \tau_r}{\Gamma @ r \vdash \lambda x. e_r : \tau_s} \xrightarrow{\text{r}} \quad \Gamma @ s \vdash e_s : \tau_s}{\Gamma @ r \oplus (s \otimes r) \vdash (\lambda x. e_r) e_s : \tau_r}$$

The derivation uses the idempotence of  $\wedge$  in the first step, followed by the (*app*) rule. The type of the term after substitution, using Lemma 5 is:

$$\frac{\Gamma, x : \tau_s @ r \vdash e_r : \tau_r \quad \Gamma @ s \vdash e_s : \tau_s}{\Gamma, x : \tau_r @ r \otimes s \vdash e_r[x \leftarrow e_s] : \tau_s}$$

From the assumptions of Lemma 5, we know that  $\otimes = \oplus$  and the operation is idempotent, so trivially:  $r \otimes s = r \oplus (s \otimes r)$   $\square$

**EXPANSION THEOREM.** The  $\eta$ -expansion (local completeness) is similar to  $\beta$ -reduction (local soundness) in that it holds for some flat coeffect systems, but not for all. Out of the examples we discuss, it holds for implicit parameters, but does not hold for liveness and data-flow.

Recall that  $\eta$ -expansion turns  $e$  into  $\lambda x. e \ x$ . In the case of liveness, the expression  $e$  may require no variables (both immediate and latent coeffects are marked as D). However, the resulting expression  $\lambda x. e \ x$  accesses a variable,

marking the context and function argument as live. In case of data-flow, the immediate coeffects are made larger by the lambda abstraction – the context requirements of the function value are imposed on the declaration site of the new lambda abstraction. We remedy this limitation in the next chapter.

However,  $\eta$ -expansion preserves the type for implicit parameters and, more generally, for any flat coeffect algebra where  $\oplus = \wedge$ . Assuming  $\rightarrow_\eta$  is the standard  $\eta$ -reduction:

**Theorem 7** (Type preservation of  $\eta$ -expansion). *In a bottom-pointed flat coeffect calculus with an algebra  $(\mathcal{C}, \otimes, \oplus, \wedge, \text{use}, \text{ign}, \leq)$  where  $\wedge = \oplus$ , if  $\Gamma @ \mathbf{r} \vdash e : \tau_1 \xrightarrow{s} \tau_2$  and  $e \rightarrow_\eta e'$  then  $\Gamma @ \mathbf{r} \vdash e' : \tau_1 \xrightarrow{s} \tau_2$ .*

*Proof.* The following derivation shows that  $\lambda x. f \ x$  has the same type as  $f$ :

$$\frac{\frac{\frac{\Gamma @ \mathbf{r} \vdash f : \tau_1 \xrightarrow{s} \tau_2 \quad x : \tau_1 @ \text{use} \vdash x : \tau_1}{\Gamma, x : \tau_1 @ \mathbf{r} \oplus (\text{use} \otimes s) \vdash f \ x : \tau_2}}{\Gamma, x : \tau_1 @ \mathbf{r} \oplus s \vdash f \ x : \tau_2}}{\Gamma, x : \tau_1 @ \mathbf{r} \wedge s \vdash f \ x : \tau_2}}{\Gamma @ \mathbf{r} \vdash \lambda x. f \ x : \tau_1 \xrightarrow{s} \tau_2}$$

The derivation starts with the expression  $e$  and derives the type for  $\lambda x. e \ x$ . The application yields context requirements  $\mathbf{r} \oplus s$ . In order to recover the original typing, this must be equal to  $\mathbf{r} \wedge s$ . Note that the derivation is showing just one possible typing – the expression  $\lambda x. e \ x$  has other types – but this is sufficient for showing type preservation.  $\square$

In summary, flat coeffect calculi do not *in general* permit call-by-name evaluation, but there are several cases where call-by-name evaluation can be used. Among the examples we discuss, these include liveness and implicit parameters. Moreover, for implicit parameters (and more generally, any set-like flat coeffect algebra), the  $\eta$ -expansion holds as well, giving us both local soundness and local completeness as coined by Pfenning and Davies [73].

#### 4.4.4 Internalized sequencing

The call-by-value and call-by-name evaluation strategies discussed in the previous section are the most important techniques for defining equational theory of flat coeffects. In this section, we briefly discuss another approach that follows the style of generic operational semantics designed by Filinski [27] for effectful computations.

The idea is to embed sequencing of context-dependent (effectful) computations as an explicit construct into the language and define a *normal form* that consists of sequentially composed context-dependent expressions. The reduction to the normal form is generic for all coeffect systems (satisfying certain conditions), while the reduction of the normal form is provided by each concrete coeffect system. The extended language with the **glet** construct and its typing is defined as follows:

$$\begin{aligned} e &::= x \mid \lambda x. e \mid e_1 \ e_2 \mid \text{let } x = e_1 \text{ in } e_2 \\ n &::= e \mid \text{glet } x = e \text{ in } n \\ \tau &::= T \mid \tau_1 \xrightarrow{r} \tau_2 \end{aligned}$$

$$(\text{glet}) \quad \frac{\Gamma @ s \vdash e_1 : \tau_1 \quad \Gamma, x : \tau_1 @ \mathbf{r} \vdash e_2 : \tau_2}{\Gamma @ \mathbf{r} \oplus (s \otimes \mathbf{r}) \vdash \text{glet } x = e_1 \text{ in } e_2 : \tau_2}$$

$$\begin{aligned}
(\lambda x. e_2) e_1 &\rightsquigarrow \text{glet } x = e_1 \text{ in } e_2 & (eval) \\
\\
(\text{glet } x = e_1 \text{ in } e_2) e_3 &\rightsquigarrow \text{glet } x = e_1 \text{ in } (e_2 e_3) & \text{if } x \notin \text{fv}(e_3) \quad (glet-app) \\
\\
\text{glet } x_2 = (\text{glet } x_1 = e_1 \text{ in } e_2) \text{ in } e_3 &\rightsquigarrow \text{glet } x_1 = e_1 \text{ in } (\text{glet } x_2 = e_2 \text{ in } e_3) & \text{if } x_1 \notin \text{fv}(e_3) \quad (glet-glet)
\end{aligned}$$

Figure 24: Reduction to normal form.

The newly introduced syntactic form  $n$  represents a normal form. The construct  $\text{glet } x = e \text{ in } n$  models an explicit sequencing of an expression  $e$ , followed by an expression  $n$ . Note that  $\text{glet}$  can only contain further  $\text{glet}$  constructs in the body, but not in the argument. The typing of the  $\text{glet}$  is the same as the typing of ordinary  $\text{let}$  construct – as discussed in Section 4.2.4, for the examples we consider, this gives a *more precise* typing than the typing of the term  $(\lambda x. n) e$ .

The reduction rules that produce a normal form are shown in Figure 24. The  $(eval)$  rule introduces  $\text{glet}$  by reducing a redex  $(\lambda x. e_2) e_1$  to an expression representing explicit sequencing of  $e_1$  and  $e_2$ . The remaining two rules specify how  $\text{glet}$  distributes across other constructs ( $\text{glet}$  and application). In  $(glet-app)$ , the  $\text{glet}$  construct appearing inside an application is lifted to the top-level; similarly  $(glet-glet)$  lifts a  $\text{glet}$  construct nested in the argument of another  $\text{glet}$ . The rules provide a way for turning an expression into a normal form.

As with *call-by-value* and *call-by-name* strategies, the *internalized sequencing* strategy can only be used with coeffect systems satisfying certain conditions. The general form of the conditions is summarized in Appendix ?. Here, we briefly consider our three examples.

**CONDITIONS.** The  $(eval)$  reduction can be safely applied for any flat coeffect calculus which satisfies the condition that the typing of the  $\text{glet}$  expression (shown above) is equivalent or more precise than typing of the expression  $(\lambda x. e_2) e_1$ . More formally:

**Definition 10.** A flat coeffect algebra  $(\mathcal{C}, \otimes, \oplus, \wedge, \text{use}, \text{ign}, \leq)$  is oriented if for all  $s, s_1, s_2, r \in \mathcal{C}$  such that  $s = s_1 \wedge s_2$  it is the case that  $s \otimes (s \oplus r) \leq s_1 \otimes (s_2 \oplus r)$ .

As discussed in Section 4.2.6, this condition is satisfied for all our three examples, strictly for liveness and data-flow and by equality for implicit parameters. For the  $(glet-app)$  and  $(glet-glet)$  rules, additional conditions arise from the typing of the original and reduced expression. The results are summarized in the following two tables.

Assuming  $\Gamma @ r_1 \vdash e_1 : \tau_1$  and  $\Gamma @ r_2 \vdash e_2 : \tau_2 \xrightarrow{s} \tau_2$  and  $\Gamma @ r_3 \vdash e_3 : \tau_3$ , the typings of the original and reduced expressions in  $(glet-app)$  rule are:

	Before	After	Satisfied
Parameters	$(r_2 \cup r_1) \cup (r_3 \cup s)$	$(r_2 \cup (r_3 \cup s)) \cup r_1$	✓
Liveness	$r_2 \sqcap (r_3 \sqcup s)$	$r_2 \sqcap (r_3 \sqcup s)$	✓
Data-flow	$\max(r_1 + r_2, r_3 + s)$	$r_1 + \max(r_2, r_3 + s)$	×

$$\boxed{\Gamma @ \mathbf{r} \vdash e : \tau}$$

$$\begin{array}{c}
\text{(typ)} \quad \frac{\Gamma @ \mathbf{r} \vdash e : \tau \quad \tau <: \tau'}{\Gamma @ \mathbf{r} \vdash e : \tau'} \\
\text{(sub)} \quad \frac{\Gamma @ \mathbf{r}' \vdash e : \tau \quad \mathbf{r}' \leq \mathbf{r}}{\Gamma @ \mathbf{r} \vdash e : \tau}
\end{array}$$

$$\boxed{\tau <: \tau'}$$

$$\begin{array}{c}
\text{(sub-trans)} \quad \frac{\tau_1 <: \tau_2 \quad \tau_2 <: \tau_3}{\tau_1 <: \tau_3} \\
\text{(sub-fun)} \quad \frac{\tau'_1 <: \tau_1 \quad \tau_2 <: \tau'_2 \quad \mathbf{r}' \geq \mathbf{r}}{\tau_1 \xrightarrow{\mathbf{r}} \tau_2 <: \tau'_1 \xrightarrow{\mathbf{r}'} \tau'_2} \\
\text{(sub-refl)} \quad \frac{}{\tau <: \tau}
\end{array}$$

Figure 25: Subtyping rules for flat coeffect calculus

Assuming  $\Gamma @ \mathbf{r}_1 \vdash e_1 : \tau_1$  and  $\Gamma @ \mathbf{r}_2 \vdash e_2 : \tau_2$  and  $\Gamma @ \mathbf{r}_3 \vdash e_3 : \tau_3$ , the typings of the original and reduced expressions in (*glet-glet*) rule are:

	Before	After	Satisfied
Parameters	$\mathbf{r}_3 + (\mathbf{r}_2 + \mathbf{r}_1)$	$(\mathbf{r}_3 + \mathbf{r}_2) + \mathbf{r}_1$	✓
Liveness	$\mathbf{r}_3$	$\mathbf{r}_3$	✓
Data-flow	$\mathbf{r}_3 \cup (\mathbf{r}_2 \cup \mathbf{r}_1)$	$(\mathbf{r}_3 \cup \mathbf{r}_2) \cup \mathbf{r}_1$	✓

This means that *internalized sequencing* provides a basis for operational semantics of liveness and implicit parameters, but (again) not for data-flow languages. For other coeffect systems, the conditions arising from the syntactic reductions have to be re-examined.

#### 4.5 SYNTACTIC PROPERTIES AND EXTENSIONS

The flat coeffect algebra introduced in Section 4.2 requires a number of laws. The laws are required for three reasons – to be able to define the categorical structure in Section 4.3, to prove equational properties in Section 4.4 and finally, to guarantee intuitive syntactic properties for constructs such as  $\lambda$ -abstraction and pairs in context-aware calculi.

In this section, we look at the last point. We discuss what syntactic equivalences are permitted by the properties of  $\wedge$  (Section 4.5.2) and we extend the calculus with pairs and units and discuss their syntactic properties (Section 4.5.3). Before doing that, the following section further develops subtyping relation for the calculus.



$$\begin{aligned}
\llbracket \Gamma @ \mathbf{r} \vdash e : \tau' \rrbracket &= \llbracket \tau <: \tau' \rrbracket \circ \llbracket \Gamma @ \mathbf{r} \vdash e : \tau \rrbracket & (typ) \\
\llbracket \tau <: \tau \rrbracket &= \text{id} & (sub-refl) \\
\llbracket \tau_1 <: \tau_3 \rrbracket &= \llbracket \tau_2 <: \tau_3 \rrbracket \circ \llbracket \tau_1 <: \tau_2 \rrbracket & (sub-trans) \\
\llbracket \tau_1 \xrightarrow{\mathbf{r}} \tau_2 <: \tau'_1 \xrightarrow{\mathbf{r}'} \tau'_2 \rrbracket &= \lambda f. & (sub-fun) \\
&\llbracket \tau_2 <: \tau'_2 \rrbracket \circ f \circ \text{map}_{\mathbf{r}} \llbracket \tau'_1 <: \tau_1 \rrbracket \circ \text{lift}_{\mathbf{r}', \mathbf{r}}
\end{aligned}$$

Figure 26: Semantics of subtyping for flat coefficients

#### 4.5.1 Subtyping for coefficients

The typing rules discussed in Section 4.2.4 include sub-coeffecting rule which makes it possible to treat an expression with smaller context requirements as an expression with greater context requirements. In the corresponding categorical semantics, this means that we can *drop* some of the provided context. However, sub-coeffecting only affects the immediate coefficients attached to the free-variable context.

In Figure 25, we add sub-typing on function types, making it possible to treat a function with smaller context requirements as a function with greater context requirements. The definition uses the standard reflexive and transitive  $<:$  operator. As the *(sub-fun)* shows, the function type is contra-variant in the input and co-variant in the output. The *(typ)* rule allows using sub-typing on expressions in the coeffect calculus.

**SEMANTICS.** We follow the same approach as in Section 4.3 and use categorical semantics to explain (and confirm) the design of the sub-typing rules. The semantics of a judgement  $\llbracket \tau <: \tau' \rrbracket$  is a function  $\llbracket \tau \rrbracket \rightarrow \llbracket \tau' \rrbracket$ . As shown in Figure 26, the semantics of the sub-typing rule *(typ)* then just composes the semantics of the original expression with the conversion produced by the semantics of the sub-typing judgement.

The rest of the Figure 26 defines the semantics of the  $<:$  operation. The reflexivity and transitivity are just the identity function and function composition, respectively. The *(sub-fun)* case is interesting – recall that the semantics of functions:

$$\llbracket \tau'_1 \xrightarrow{\mathbf{r}'} \tau'_2 \rrbracket = C^{\mathbf{r}'} \tau'_1 \rightarrow \tau'_2$$

We build the transformation using an explicit lambda abstraction that takes the original function  $f$  as an argument. To build the required function, we first drop unnecessary context using  $\text{lift}_{\mathbf{r}', \mathbf{r}}$  of type  $C^{\mathbf{r}'} \tau'_1 \rightarrow C^{\mathbf{r}} \tau'_1$ , then we use the  $\text{map}_{\mathbf{r}}$  function to transform the nested  $\tau'_1$  to  $\tau_1$ . Finally we evaluate the original function  $f$  and turn the resulting value of type  $\tau_2$  into the required result of type  $\tau'_2$ .

#### 4.5.2 Properties of lambda abstraction

In Section 4.2.1, we discussed how to reconcile two typings for lambda abstraction – for implicit parameters, the lambda function needs to split context requirements using  $\mathbf{r} \cup \mathbf{s}$ , but for data-flow and liveness it suffices to duplicate the requirement  $\mathbf{r}$  of the body. We introduced the  $\wedge$  operation as a way of providing the additional abstraction.



In this section, we first identify coeffect calculi for which the simpler (duplicating) rule is sufficient. Then we look at syntactic transformations corresponding to other common properties of the  $\wedge$  operation.

**SIMPLIFIED ABSTRACTION.** Recall that  $(\mathcal{C}, \wedge)$  is a band, that is,  $\wedge$  is idempotent and associative. The idempotence means that the context requirements of the body can be required from both the declaration site and the call site. Thus, the following (*idabs*) typing is valid (for reference, it is shown side-by-side with the ordinary lambda abstraction rule):

$$\begin{array}{c} \text{(*idabs*)} \quad \frac{\Gamma, x : \tau_1 @ \mathbf{r} \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{r}} \tau_2} \quad \text{(*abs*)} \quad \frac{\Gamma, x : \tau_1 @ \mathbf{r} \wedge \mathbf{r} \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{r}} \tau_2} \end{array}$$

To derive (*idabs*), we use idempotence on the body annotation  $\mathbf{r} = \mathbf{r} \wedge \mathbf{r}$  and then use the standard (*abs*) rule. So, (*idabs*) follows from (*abs*), but the other direction is not necessarily the case. The following condition identifies coeffect calculi where (*abs*) can be derived from (*idabs*).

**Definition 11.** A flat coeffect algebra  $(\mathcal{C}, \otimes, \oplus, \wedge, \text{use}, \text{ign}, \leq)$  is strictly oriented if for all  $s, r \in \mathcal{C}$  it is the case that  $r \wedge s \leq r$ .

**Remark 8.** For a flat coeffect calculus with a strictly oriented algebra, the standard (*abs*) rule can be derived from the (*idfun*) rule.

*Proof.* The following derives the conclusion of (*abs*) using (*idabs*), sub-coeffecting, sub-typing and the fact that the algebra is strictly oriented:

$$\begin{array}{c} \text{(*idabs*)} \quad \frac{\Gamma, x : \tau_1 @ \mathbf{r} \wedge \mathbf{s} \vdash e : \tau_2}{\Gamma @ \mathbf{r} \wedge \mathbf{s} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{r} \wedge \mathbf{s}} \tau_2} \quad \text{(*sub*)} \quad \frac{\Gamma @ \mathbf{r} \wedge \mathbf{s} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{r} \wedge \mathbf{s}} \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{r} \wedge \mathbf{s}} \tau_2} \quad (\mathbf{r} \leq \mathbf{r} \wedge \mathbf{s}) \\ \text{(*typ*)} \quad \frac{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{r} \wedge \mathbf{s}} \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{s}} \tau_2} \quad (\mathbf{r} \leq \mathbf{r} \wedge \mathbf{s}) \end{array}$$

□

The practical consequence of the Remark 8 is that, for strictly oriented coeffect calculi (such as our liveness and data-flow computations), we can use the (*idabs*) rule and get an equivalent type system. This alternative formulation removes the non-determinism of type checking that arises from the splitting of context requirements in the original (*abs*) rule. Furthermore, for data-flow and liveness the (*idabs*) rule is more precise than (*abs*).

**SYMMETRY.** The  $\wedge$  operation is idempotent and associative. In all of the three examples considered in this chapter, the operation is also *symmetric*. To make our definitions more general, we do not require this to be the case for *all* flat coeffect systems. However, systems with symmetric  $\wedge$  have the following property.

**Remark 9.** For a flat coeffect calculus such that  $\mathbf{r} \wedge \mathbf{s} = \mathbf{s} \wedge \mathbf{r}$ , assuming that  $\mathbf{r}', \mathbf{s}', \mathbf{t}'$  is a permutation of  $\mathbf{r}, \mathbf{s}, \mathbf{t}$ :

$$\frac{\Gamma, x : \tau_1, y : \tau_2 @ \mathbf{r} \wedge \mathbf{s} \wedge \mathbf{t} \vdash e : \tau_3}{\Gamma @ \mathbf{r}' \vdash \lambda x. \lambda y. e : \tau_1 \xrightarrow{\mathbf{s}'} (\tau_2 \xrightarrow{\mathbf{t}'} \tau_3)}$$

Intuitively, this means that the context requirements of a function with multiple arguments can be split arbitrarily between the declaration site and (multiple) call sites. In other words, it does not matter how the context requirements are satisfied.

$$\begin{array}{c}
\text{(pair)} \quad \frac{\Gamma @ \mathbf{r} \vdash e_1 : \tau_1 \quad \Gamma @ \mathbf{s} \vdash e_2 : \tau_2}{\mathbf{r} \oplus \mathbf{s} @ \Gamma \vdash (e_1, e_2) : \tau_1 \times \tau_2} \\
\\
\text{(proj)} \quad \frac{\Gamma @ \mathbf{r} \vdash e : \tau_1 \times \tau_2}{\Gamma @ \mathbf{r} \vdash \pi_i e : \tau_i} \\
\\
\text{(unit)} \quad \frac{}{\Gamma @ \mathbf{ign} \vdash () : \text{unit}}
\end{array}$$

Figure 27: Typing rules for pairs and units

#### 4.5.3 Language with pairs and unit

To show the key aspects of flat coeffect systems, the calculus introduced in Section 4.2 consists only of variables, abstraction, application and let binding. Here, we extend it with pairs and the unit value to sketch how it can be turned into a more complete programming language and to motivate the laws required about  $\oplus$ . The syntax of the language is extended as follows:

$$\begin{array}{lcl}
e & ::= & \dots \mid () \mid e_1, e_2 \\
\tau & ::= & \dots \mid \text{unit} \mid \tau \times \tau
\end{array}$$

The typing rules for pairs and the unit value are shown in Figure 27. The unit value (*unit*) is annotated with the **ign** coeffect (the same as other constants). Pairs, created using the  $(e_1, e_2)$  expression, are annotated with a coeffect that combines the coeffects of the two sub-expressions using the *pointwise* operator  $\oplus$ . The operator models the case when the (same) available context is split and passed to two independent sub-expressions. Finally, the (*proj*) rule is uninteresting, because  $\pi_i$  can be viewed as a pure function.

**PROPERTIES.** Pairs and the unit value in a lambda calculus typically form a monoid. Assuming  $\simeq$  is an isomorphism that performs appropriate transformation on values, without affecting other properties (here, coeffects) of the expressions. The monoid laws then correspond to the requirement that  $(e_1, (e_2, e_3)) \simeq ((e_1, e_2), e_3)$  (associativity) and the requirement that  $((), e) \simeq e \simeq (e, ())$  (unit).

Thanks to the properties of  $\oplus$ , the flat coeffect calculus obeys the monoid laws for pairs. In the following, we assume that *assoc* is a pure function transforming a pair  $(x_1, (x_2, x_3))$  to a pair  $((x_1, x_2), x_3)$ . We write  $e \equiv e'$  when for all  $\Gamma, \tau$  and  $\mathbf{r}$ , it is the case that  $\Gamma @ \mathbf{r} \vdash e : \tau$  if and only if  $\Gamma @ \mathbf{r} \vdash e' : \tau$ .

**Theorem 10.** *For a flat coeffect calculus with pairs and units, the following holds:*

$$\begin{array}{lll}
\text{assoc } (e_1, (e_2, e_3)) & \equiv & ((e_1, e_2), e_3) \quad (\text{associativity}) \\
\pi_1 (e, ()) & \equiv & e \quad (\text{right unit}) \\
\pi_2 ((), e) & \equiv & e \quad (\text{left unit})
\end{array}$$

*Proof.* Follows from the fact that  $(\mathcal{C}, \oplus, \mathbf{ign})$  is a monoid and *assoc*,  $\pi_1$  and  $\pi_2$  are pure functions (treated as constants in the language).  $\square$

The Theorem 10 motivates the requirement of the monoid structure  $(\mathcal{C}, \oplus, \mathbf{ign})$  of the flat coeffect algebra. We require only unit and associativity laws. In our three examples, the  $\oplus$  operator is also symmetric, which additionally gives us the property that  $(e_1, e_2) \simeq (e_2, e_1)$ .

## 4.6 RELATED WORK

Most of the related work leading to coeffects has already been discussed in Chapter 2 and we covered work related to individual concepts throughout the chapter. In this section, we do not repeat the discussion present elsewhere. Instead, we discuss one specific question that often arises when discussing coeffects and that is *when is a coeffect (not) an effect?*

We start with a quick overview of the ways in which effects and coeffects differ and then we briefly look at one (but illustrative) example where the two concepts overlap. We focus mainly on the equivalence between the *categorical semantics*, which reveals the nature of the computations – rather than considering just the syntactic aspects of the type system.

## 4.6.1 When is coeffect not a monad

Coeffect systems differ from effect systems in three important ways:

- Semantically, coeffects capture different notions of computation. As demonstrated in Chapter 2, coeffects track additional contextual properties required by a computation, many of which cannot be captured by a monad (e. g. liveness or data-flow).
- Syntactically, coeffect calculi use a richer algebraic structure with pointwise composition, sequential composition and context merging ( $\oplus$ ,  $\otimes$ , and  $\wedge$ ) while most effect systems only use a single operation for sequential composition (used by monadic bind).
- Syntactically, the second difference is in the lambda abstraction (*abs*). In coeffect systems, the context requirements of the body can be split between (or duplicated at) declaration site and call site, while monadic effect systems always defer all effects.

Despite the differences, our implicit parameters example can be also represented by a monad. Semantically, the *reader* monad is equivalent to the *product* comonad. Syntactically, we use the  $\cup$  operation for all three operations of the coeffect algebra. However, to enable splitting of implicit parameter requirements using the reader monad, we need to extend the monad structure and change the translation of monadic lambda abstraction.

## 4.6.2 When is coeffect a monad

Implicit parameters can be captured by a monad, but *just* a monad is not enough. Lambda abstraction in effect systems does not provide a way of splitting the context requirements between declaration site and call site (or, semantically, combining the implicit parameters available in the scope where the function is defined and those specified by the caller).

**CATEGORICAL RELATIONSHIP.** Before looking at the necessary extensions, consider the two ways of modelling implicit parameters. We assume that the function  $\mathbf{r} \rightarrow \sigma$  is a lookup function for reading implicit parameter values that is defined on a set  $\mathbf{r}$ . The two definitions are:

$$\begin{aligned} C^{\mathbf{r}}\tau &= \tau \times (\mathbf{r} \rightarrow \sigma) && (\text{product comonad}) \\ M^{\mathbf{r}}\tau &= (\mathbf{r} \rightarrow \sigma) \rightarrow \tau && (\text{reader monad}) \end{aligned}$$

The *product comonad* simply pairs the value  $\tau$  with the lookup function, while the *reader monad* is a function that, given a lookup function, produces a  $\tau$  value. As noted by Orchard [63], when used to model computation semantics, the two representations are equivalent:

**Remark 11.** *Computations modelled as  $C^r\tau_1 \rightarrow \tau_2$  using the product comonad are isomorphic to computations modelled as  $\tau_1 \rightarrow M^r\tau_2$  using the reader monad via currying/uncurrying isomorphism.*

*Proof.* The isomorphism is demonstrated by the following equation:

$$\begin{aligned} C^r\tau_1 \rightarrow \tau_2 &= (\tau_1 \times (r \rightarrow \sigma)) \rightarrow \tau_2 \\ &= \tau_1 \rightarrow ((r \rightarrow \sigma) \rightarrow \tau_2) = \tau_1 \rightarrow M^r\tau_2 \end{aligned}$$

This equivalence holds for monads and comonads (as well as *indexed* monads and comonads), but it does not extend to *flat* indexed comonads which also provide the  $\text{merge}_{r,s}$  operation to model context merging.

**DELAYING EFFECTS IN MONADS.** In the syntax of the language, the above difference is manifested by the (*abs*) rules for monadic effect systems and comonadic coeffect systems. The following listing shows the two rules side-by-side, using the effect system notation for both of them:

$$\begin{array}{c} (cabs) \quad \frac{\Gamma, x:\tau_1 \vdash e : \tau_2 \ \& \ r \cup s}{\Gamma \vdash \lambda x.e : \tau_1 \xrightarrow{s} \tau_2 \ \& \ r} \qquad (mabs) \quad \frac{\Gamma, x:\tau_1 \vdash e : \tau_2 \ \& \ r \cup s}{\Gamma \vdash \lambda x.e : \tau_1 \xrightarrow{r \cup s} \tau_2 \ \& \ \emptyset} \end{array}$$

In the comonadic (*cabs*) rule, the implicit parameters of the body are split. However, the monadic rule (*mabs*) places all requirements on the call site. This follows from the fact that monadic semantics uses the unit operation in the interpretation of lambda abstraction:

$$\llbracket \lambda x.e \rrbracket = \text{unit} (\lambda x. \llbracket e \rrbracket)$$

The type of unit is  $\alpha \rightarrow M^\alpha \emptyset$ , but in this specific case, the  $\alpha$  is instantiated to be  $\tau_1 \rightarrow M^{r \cup s} \tau_2$  and so this use of unit has a type:

$$\text{unit} : (\tau_1 \rightarrow M^{r \cup s} \tau_2) \rightarrow M^\emptyset(\tau_1 \rightarrow M^{r \cup s} \tau_2)$$

In order to split the implicit parameters of the body ( $r \cup s$  on the left-hand side) between the declaration site ( $\emptyset$  on the outer  $M$  on the right-hand side) and the call site ( $r \cup s$  on the inner  $M$  on the right-hand side), we need an operation (which we call *delay*) with the following signature:

$$\text{delay}_{r,s} : (\tau_1 \rightarrow M^{r \cup s} \tau_2) \rightarrow M^r(\tau_1 \rightarrow M^s \tau_2)$$

The operation reveals the difference between effects and coeffects – intuitively, given a function with effects  $r \cup s$ , it should execute the effects  $r$  when wrapping the function, *before* the function actually performs the effectful operation with the effects. The remaining effects  $s$  are delayed as usual, while effects  $r$  are removed from the effect annotation of the body.

Another important aspect of the signature is that the function needs to be indexed by the coeffect annotations  $r, s$ . The indices determine how the input context requirements  $r \cup s$  are split – and thus guarantee determinism of the function at run-time.

The operation cannot be implemented in a useful way for most standard monads, but the reader monad is, indeed, an exception. It is not difficult to see how it can be implemented when we expand the definitions of  $M^r\tau$ :

$$\text{delay}_{r,s} : (\tau_1 \rightarrow (r \cup s \rightarrow \sigma) \rightarrow \tau_2) \rightarrow ((r \rightarrow \sigma) \rightarrow \tau_1 \rightarrow (s \rightarrow \sigma) \rightarrow \tau_2)$$

RESTRICTING COEFFECTS IN COMONADS. As just demonstrated, we can extend monads so that the reader monad is capable of capturing the semantics of implicit parameters, including the splitting of implicit parameter requirements in lambda abstraction. Can we also go the other way round and *restrict* the comonadic semantics so that all requirements are delayed as in the (*mabs*) rule, thus modelling fully dynamically scoped parameters?

This is, indeed, possible. Recall that the semantics of lambda abstraction in the flat coeffect calculus is modelled using  $\text{merge}_{r,s}$ . The operation takes two contexts (wrapped in an indexed comonad  $C^r \alpha$ ), combines their carried values and additional contextual information (implicit parameters). To obtain the (*mabs*) rule, we can restrict the first parameter, which corresponds to the declaration site context:

$$\begin{aligned} \text{merge}_{r,s} &: C^r \alpha \times C^s \beta \rightarrow C^{r \cup s}(\alpha \times \beta) & (\text{normal}) \\ \text{merge}_{r,s} &: C^\emptyset \alpha \times C^s \beta \rightarrow C^s(\alpha \times \beta) & (\text{restricted}) \end{aligned}$$

In the (*restricted*) version of the operation, the declaration site context requires no implicit parameters and so all implicit parameters have to be satisfied by the call site. The semantics using the restricted version corresponds to the (*mabs*) rule shown above.

The idea of restricting the operations of the coeffect calculus semantics could be used more generally. We could allow any of the coeffect algebra operations  $\otimes, \wedge, \oplus$  to be *partial* and thus the restricted (fully dynamically-scoped) version of implicit parameters could be obtained just by changing the definition of  $\wedge$ . Similarly, we could obtain e.g. a fully lexically-scoped version of the system. The ability to restrict operations to partial functions has been used in the semantics of effectful computations by Tate [87].

## 4.7 CONCLUSIONS

This chapter presented the *flat coeffect calculus* – a unified system for tracking *whole-context* properties of computations, that is properties related to the execution environment or the entire context in which programs are executed. This is the first of the three *coeffect calculi* developed in this thesis.

The flat coeffect calculus is parameterized by a *flat coeffect algebra* that captures the structure of the information tracked by the type system. We instantiated the system to capture three specific systems, namely liveness, data-flow and implicit parameters. However, the system is more general and can capture numerous other applications outlined in Section 3.2.

Next, we introduced the notion of *flat indexed comonad*, which generalizes of comonads and adds additional operations needed to provide categorical semantics of the flat coeffect calculus. The indices of the flat indexed comonad operations correspond to the coeffect annotations in the type system and provide a foundation for the design of the calculus.

Finally, we discussed the equational theory for flat coeffect calculus. Although each concrete instance of flat coeffect calculus models different notion of context, there are syntactic properties that hold for all flat coeffect systems satisfying certain additional conditions. In particular, two *typing preservation* theorems prove that the operational semantics for two classes of flat coeffect calculi (including liveness and implicit parameters) can be based on the standard call-by-name reduction.

In the upcoming chapter, we move from *flat coeffect calculi*, tracking whole-context properties to *structural coeffect calculi*, tracking per-variable information, thus covering systems from the second half of Chapter 3.



In this chapter, we take the next step towards a unified theory of context-aware programming languages. As already discussed, we attempt to find abstractions that are *general enough* to capture a wide range of useful programming language features, but *specific enough* to let us identify and capture interesting properties of the context.

In Chapter 3, we discussed two notions of context. Context-aware programming languages that capture whole-context properties were generalized by the *flat coeffect calculus* in Chapter 4. Here, we consider per-variable contextual properties and we introduce the *structural coeffect calculus*.

The flat coeffect system captures a number of interesting use-cases (implicit parameters, liveness and data-flow), but provides relatively imprecise approximation and weak syntactical properties. Some of those examples can be also seen as per-variable properties. For those, structural coeffect systems give a more precise information about the context. However, we also look at other applications that arise from the work on sub-structural logics, a pathway to coeffects discussed in Section 2.3.

As we will see, the categorical semantics of the structural coeffect calculus is based on the same underlying structure as the semantics of the flat calculus. However, the syntactical properties of the calculus are quite different and so it is worth considering it in a separate chapter.

#### CHAPTER STRUCTURE AND CONTRIBUTIONS

- We present a *structural coeffect calculus* as a type system that is parameterized by a *structural coeffect algebra* (Section 5.2). We show how the system captures fine-grained liveness and data-flow information, as well as a calculus for bounded reuse (checking how many times is a variable accessed).
- We extend indexed comonads introduced in the previous chapter to *structural indexed comonads* and use them to provide the semantics of structural coeffect calculus (Section 5.3). As with the flat version, the semantics provides a motivation for the design of the calculus.
- We discuss the equational theory of the calculus. For structural coeffects, the substitution lemma holds for all instances of the calculus. Thus, we prove the type-preservation property for all structural calculi for both call-by-name and call-by-value (Section 5.4).

#### 5.1 INTRODUCTION

Compared to the previous chapter, structural coeffect calculi we consider are more homogeneous and so finding the common pattern is in some ways easier. However, the systems are more complicated as they need to keep annotations attached to individual variables and thus require explicit structural rules. Before looking at the system, we briefly consider the most important related work.

### 5.1.1 Related work

In the previous chapter, we discussed the correspondence between coeffects and effects (and between comonads and monads). As noted earlier, the  $\lambda$ -calculus is asymmetric in that an expression has multiple inputs (variables in the context), but just a single result (the resulting value) and so monads and effects have no notion directly corresponding to structural coeffect systems.

The work in this chapter is more closely related to sub-structural type systems [104]. While sub-structural systems remove some or all of *weakening*, *contraction* and *exchange* rules, our systems keep them, but use them to manipulate both the context and its annotations.

Our work follows the “language semantics” style in that we provide a structural semantics to the terms of ordinary  $\lambda$ -calculus. By contrast the most closely related work has been done in the meta-language style, which extends the terms and types of a language with constructs for explicitly manipulating the context. This includes Contextual Modal Type Theory (CMTT) [58], where variables may be of a type  $A[\Psi]$  denoting a value of type  $A$  that requires context  $\Psi$ . In CMTT,  $A[\Psi]$  is a first-class type, while structural coeffect systems do not expose coeffect annotations as stand-alone types (indexed comonads only appear in the semantics).

Our structural coeffect systems annotate the whole variable context with a *vector* of annotations. For example, a context with variables  $x$  and  $y$  annotated with  $s$  and  $t$ , respectively is written as  $x:\tau_1, y:\tau_2 @ \langle s, t \rangle$ . A benefit of this approach is that the typing judgements have the same structure as those of the flat coeffect calculus. As discussed in Chapter 6, this makes it possible to unify the two systems.

## 5.2 STRUCTURAL COEFFECT CALCULUS

In the structural coeffect calculus, a vector of variables forming the free-variable context is annotated with a vector of primitive (scalar) coeffect annotations. These annotations differ for different coeffect calculi and their properties are captured by the definition of *structural coeffect scalar* below. The scalar annotations can be e. g. integers (how many past values we need) or values of a two-point lattice specifying whether a variable is live or not.

Scalar annotations are written as  $r, s, t$  (following the style used in the previous chapter). Functions always have exactly one input variable and so they are annotated with a single coeffect scalar. Thus the expressions and types of structural coeffect calculi are the same as in the previous chapter. The only difference is that  $r, s, t$  now range over values of a *structural coeffect scalar* rather than over values of a *flat coeffect algebra* (the definitions are related, but not the same).

$$\begin{aligned} e &::= x \mid \lambda x. e \mid e_1 \ e_2 \mid \text{let } x = e_1 \text{ in } e_2 \\ \tau &::= T \mid \tau_1 \xrightarrow{r} \tau_2 \end{aligned}$$

In the previous chapter, the free variable context  $\Gamma$  was treated as a set. In the structural coeffect calculus, the order of variables matters. Thus we treat a free variable context as a vector with a uniqueness condition. We also write  $\text{len}(-)$  for the length of the vector:

$$\begin{aligned} \Gamma &= \langle x_1:\tau_1, \dots, x_n:\tau_n \rangle \quad \text{such that } \forall i, j. i \neq j \implies x_i \neq x_j \\ \text{len}(\Gamma) &= n \end{aligned}$$



For readability, we use the usual notation  $x_1 : \tau_1, \dots, x_n : \tau_n \vdash e : \tau$  for typing judgements, but the free variable context should be understood as a vector. Furthermore, the notation  $\Gamma_1, \Gamma_2$  is here seen as the tensor product. Given  $\Gamma_1 = \langle x_1 : \tau_1, \dots, x_n : \tau_n \rangle$  and  $\Gamma_2 = \langle x_{n+1} : \tau_{n+1}, \dots, x_m : \tau_m \rangle$  then  $\Gamma_1, \Gamma_2 = \Gamma_1 \times \Gamma_2 = \langle x_1 : \tau_1, \dots, x_m : \tau_m \rangle$ .

Free variable contexts are annotated with vectors of structural coeffect scalars. In what follows, we write the vectors of coeffects as  $\langle \mathbf{r}_1, \dots, \mathbf{r}_n \rangle$ . Meta-variables ranging over vectors are written as  $\mathbf{r}, \mathbf{s}, \mathbf{t}$  (using bold face and colour to distinguish them from scalar meta-variables) and the length of a coeffect vector is written as  $\text{len}(\mathbf{r})$ . The structure for working with vectors of coeffects is provided by the definition of *structural coeffect algebra* below.

### 5.2.1 Structural coeffect algebra

The *structural coeffect scalar* structure is similar to *flat coeffect algebra* with the exception that it drops the  $\wedge$  operation. It only provides a monoid  $(\mathbb{C}, \otimes, \text{use})$  modelling sequential composition of computations and a monoid  $(\mathbb{C}, \oplus, \text{ign})$  representing pointwise composition, as well as a relation  $\leq$  that defines sub-coeffecting.

**Definition 12.** A *structural coeffect scalar*  $(\mathbb{C}, \otimes, \oplus, \text{use}, \text{ign}, \leq)$  is a set  $\mathbb{C}$  together with elements  $\text{use}, \text{ign} \in \mathbb{C}$ , relation  $\leq$  and binary operations  $\otimes, \oplus$  such that  $(\mathbb{C}, \otimes, \text{use})$  and  $(\mathbb{C}, \oplus, \text{ign})$  are monoids and  $(\mathbb{C}, \leq)$  is a pre-order. That is, for all  $r, s, t \in \mathbb{C}$ :

$$\begin{aligned} r \otimes (s \otimes t) &= (r \otimes s) \otimes t & \text{use} \otimes r &= r = r \otimes \text{use} & (\text{monoid}) \\ r \oplus (s \oplus t) &= (r \oplus s) \oplus t & \text{ign} \oplus r &= r = r \oplus \text{ign} & (\text{monoid}) \\ \text{if } r \leq s \text{ and } s \leq t &\text{ then } r \leq t & t \leq t & & (\text{pre-order}) \end{aligned}$$

The following distributivity axioms hold:

$$\begin{aligned} (r \oplus s) \otimes t &= (r \otimes t) \oplus (s \otimes t) \\ t \otimes (r \oplus s) &= (t \otimes r) \oplus (t \otimes s) \end{aligned}$$

In addition, we require the following two properties of  $\text{ign}, \oplus$  and  $\leq$ :

$$\begin{aligned} r \leq r' &\Rightarrow \forall s. (r \otimes s) \leq (r' \otimes s) \\ \text{ign} &\leq (\text{ign} \otimes r) \end{aligned}$$

In the flat coeffect calculus, we used the  $\wedge$  operation to merge the annotations of contexts available from the declaration site and the call site or, in the syntactic reading, to split the context requirements. The other difference is that the  $\oplus$  operation of structural coeffect scalar is not required to be idempotent. The bounded reuse example uses  $\oplus$  which is not idempotent.

Finally, we now also require two properties that relate the ordering  $\leq$  with coeffects composed via  $\oplus$  and with  $\text{ign}$ . These were previously required for bottom-pointed substitution lemma (Lemma 5). Here, we require them for all structural coeffects as it holds for all our examples; it is needed for the structural version of the substitution lemma (Lemma 14).

In the structural coeffect calculus, we use a vector instead – combining and splitting of coeffects becomes just vector concatenation or splitting, respectively; these are provided by the tensor product. The operations on vectors are indexed by integers representing the lengths of the vectors. The additional structure required by the type system for structural coeffect calculi is given by the following definition.

**Definition 13.** A **structural coeffect algebra** is formed by a structural coeffect scalar  $(\mathbb{C}, \otimes, \oplus, \text{use}, \text{ign}, \leq)$  equipped with the following additional structures:

- Coeffect vectors  $\mathbf{r}, \mathbf{s}, \mathbf{t}$ , ranging over structural coeffect scalars indexed by vector lengths  $\mathbf{m}, \mathbf{n} \in \mathbb{N}$ .
- An operation that constructs a vector from scalars indexed by the vector length  $\langle - \rangle_{\mathbf{n}} : \mathbb{C} \times \dots \times \mathbb{C} \rightarrow \mathbb{C}^{\mathbf{n}}$  and an operation that returns the vector length such that  $\text{len}(\mathbf{r}) = \mathbf{n}$  for  $\mathbf{r} : \mathbb{C}^{\mathbf{n}}$
- A pointwise extension of the  $\otimes$  operator written as  $\mathbf{t} \otimes \mathbf{s}$  such that  $\mathbf{t} \otimes \langle \mathbf{r}_1, \dots, \mathbf{r}_n \rangle = \langle \mathbf{t} \otimes \mathbf{r}_1, \dots, \mathbf{t} \otimes \mathbf{r}_n \rangle$ .
- An indexed tensor product  $\times_{\mathbf{n}, \mathbf{m}} : \mathbb{C}^{\mathbf{n}} \times \mathbb{C}^{\mathbf{m}} \rightarrow \mathbb{C}^{\mathbf{n} + \mathbf{m}}$  that is used in both directions – for vector concatenation and for splitting – which is defined as  $\langle \mathbf{r}_1, \dots, \mathbf{r}_n \rangle \times_{\mathbf{n}, \mathbf{m}} \langle \mathbf{s}_1, \dots, \mathbf{s}_m \rangle = \langle \mathbf{r}_1, \dots, \mathbf{r}_n, \mathbf{s}_1, \dots, \mathbf{s}_m \rangle$

The fact that the tensor product  $\times_{\mathbf{n}, \mathbf{m}}$  is indexed by the lengths of the two vectors means that we can use it unambiguously for both concatenation of vectors and for splitting of vectors, provided that the lengths of the resulting vectors are known. In the following text, we usually omit the indices and write just  $\mathbf{r} \times \mathbf{s}$ , because the lengths of the coeffect vectors can be determined from the lengths of the matching free variable context vectors.

More generally, we could see the the coeffect annotations as *containers* [3]. This approach is used in Chapter 6 to unify flat and structural systems.

### 5.2.2 Structural coeffect types

The type system for the structural coeffect calculus is similar to sub-structural type systems in how it handles free variable contexts. The *syntax-driven* rules do not implicitly allow weakening, exchange or contraction – this is done by checking the types of sub-expressions in disjoint parts of the free variable context. Unlike in sub-structural logics, our system does allow weakening, exchange and contraction, but via explicit *structural* rules that perform corresponding transformation on the coeffect annoation.

**SYNTAX-DRIVEN RULES.** The syntax-driven rules of the type system are shown in Figure 28 (a). The variable access rule (*var*) annotates the corresponding variable as being accessed using *use*. Note that, as in sub-structural systems, the free variable context contains *only* the accessed variable. Other variables can be introduced using explicit weakening. Constants (*const*) are type checked in an empty variable context, which is annotated with an empty vector of coeffect annotations.

The (*abs*) rule assumes that the free variable context of the body can be split into a potentially empty *declaration site* and a singleton context containing the bound variable. The corresponding splitting is performed on the coeffect vector, uniquely associating the annotation *s* with the bound variable *x*. This means that the typing rule removes non-determinism of the type inference present in flat coeffect systems.

In (*app*), the sub-expressions  $e_1$  and  $e_2$  use free variable contexts  $\Gamma_1, \Gamma_2$  with coeffect vectors  $\mathbf{r}, \mathbf{s}$ , respectively. The function value is annotated with a coeffect scalar *t*. The coeffect annotation of the composed expression is obtained by combining the annotations associated with variables in  $\Gamma_1$  and  $\Gamma_2$ . Variables in  $\Gamma_1$  are only used to obtain the function value, resulting in coeffects  $\mathbf{r}$ . The variables in  $\Gamma_2$  are used to obtain the argument value, which is then sequentially composed with the function, resulting in  $\mathbf{t} \otimes \mathbf{s}$ .

a.) Syntax-driven typing rules:

$$\begin{aligned}
(\text{var}) \quad & \frac{}{x : \tau @ \langle \text{use} \rangle \vdash x : \tau} \\
(\text{const}) \quad & \frac{c : \tau \in \Delta}{() @ \langle \rangle \vdash c : \tau} \\
(\text{app}) \quad & \frac{\Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \quad \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma_1, \Gamma_2 @ \mathbf{r} \times (\mathbf{t} \otimes \mathbf{s}) \vdash e_1 e_2 : \tau_2} \\
(\text{abs}) \quad & \frac{\Gamma, x : \tau_1 @ \mathbf{r} \times \langle \mathbf{s} \rangle \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{s} \tau_2} \\
(\text{let}) \quad & \frac{\Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \quad \Gamma_2, x : \tau_1 @ \mathbf{s} \times \langle \mathbf{t} \rangle \vdash e_2 : \tau_2}{\Gamma_1, \Gamma_2 @ (\mathbf{t} \otimes \mathbf{r}) \times \mathbf{s} \vdash \text{let } x = e_1 \text{ in } e_2 : \tau_2}
\end{aligned}$$

b.) Structural rules for context manipulation:

$$\begin{aligned}
(\text{sub}) \quad & \frac{\Gamma_1, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}' \rangle \times \mathbf{q} \vdash e : \tau}{\Gamma_1, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} \rangle \times \mathbf{q} \vdash e : \tau} \quad (s' \leq s) \\
(\text{weak}) \quad & \frac{\Gamma @ \mathbf{r} \vdash e : \tau}{\Gamma, x : \tau_1 @ \mathbf{r} \times \langle \text{ign} \rangle \vdash e : \tau} \\
(\text{exch}) \quad & \frac{\Gamma_1, x : \tau_1, y : \tau_2, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}, \mathbf{t} \rangle \times \mathbf{q} \vdash e : \tau}{\Gamma_1, y : \tau_2, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{t}, \mathbf{s} \rangle \times \mathbf{q} \vdash e : \tau} \quad \begin{array}{l} \text{len}(\Gamma_1) = \text{len}(\mathbf{r}) \\ \text{len}(\Gamma_2) = \text{len}(\mathbf{s}) \end{array} \\
(\text{contr}) \quad & \frac{\Gamma_1, y : \tau_1, z : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}, \mathbf{t} \rangle \times \mathbf{q} \vdash e : \tau}{\Gamma_1, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} \oplus \mathbf{t} \rangle \times \mathbf{q} \vdash e[z, y \leftarrow x] : \tau} \quad \begin{array}{l} \text{len}(\Gamma_1) = \text{len}(\mathbf{r}) \\ \text{len}(\Gamma_2) = \text{len}(\mathbf{s}) \end{array}
\end{aligned}$$

Figure 28: Type system for the structural coeffect calculus

**STRUCTURAL RULES.** The remaining rules, shown in Figure 28 (b), are not syntax-directed. They allow different transformation of the free variable context. We include (*sub*) as one of the rules, allowing sub-coeffecting on coeffect scalars belonging to individual variables. The remaining rules capture *weakening*, *exchange* and *contraction* known from sub-structural systems.

Rule (*weak*) allows adding a variable to the context, extending the coeffect vector with *ign* to mark it as unused, (*exch*) provides a way to rearrange variables in the context, performing the same reordering on the coeffect vector. Finally recall that variables in the free variable context are required to be *unique*. The (*contr*) rule allows re-using a variable. We can type check sub-expressions using two separate variables and then unify them using substitution. The resulting variable is annotated with  $\oplus$  and it is the only place in the structural coeffect system where context requirements are combined (semantically, this is where the available context is shared).

### 5.2.3 Understanding structural coeffects

The type system for structural coeffects appears more complicated when compared to the flat version, but it is in many ways simpler – it removes the ambiguity arising from the use of  $\wedge$  in lambda abstraction and, as discussed in Section 5.4, has a cleaner equational theory.

**FLAT AND STRUCTURAL CONTEXT.** In flat systems, lambda abstraction splits context requirements using  $\wedge$  and application combines them using  $\oplus$ . In the structural version, both of these are replaced with  $\times$ . The  $\wedge$  operation is not needed, but  $\oplus$  is still used in the (*contr*) rule.

This suggests that  $\wedge$  and  $\oplus$  serve two roles in flat coeffacts. First, they are used as over-approximations and under-approximations of  $\times$ . This is demonstrated by the (*approximation*) requirement introduced in Section 4.4.2, which requires that  $r \wedge t \leq r \oplus t$ . Semantically, flat abstraction combines two values representing the available context, potentially discarding parts of it (under-approximation), while flat application splits the available context (a single value), potentially duplicating parts of it (over-approximation)<sup>1</sup>.

Secondly, the operator  $\oplus$  is used when the semantics passes a given context to multiple sub-expressions. In flat systems, the context is shared in (*app*) and (*pair*), because the sub-expressions may share variables. In structural systems, the sharing is isolated into an explicit contraction rule.

**LET BINDING.** The other aspect that makes structural systems simpler is that they remove the need for separate let binding. As discussed in Section 4.2.6, flat calculi include let binding that gives a *more precise* typing than combination of abstraction and application. Structural coeffact systems avoid this issue.

**Remark 12** (Let binding). *In a structural coeffact calculus, the typing of  $\text{let } x = e_1 \text{ in } e_2$  can be seen as a derived rule, i. e. its typing is equivalent to the typing of the expression  $(\lambda x. e_2) e_1$ .*

*Proof.* Consider the following typing derivation for  $(\lambda x. e_2) e_1$ . Note that in the last step, we apply (*exch*) repeatedly to swap  $\Gamma_1$  and  $\Gamma_2$ .

$$\frac{\frac{\Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \quad \frac{\Gamma_2, x : \tau_1 @ \mathbf{s} \times \langle \mathbf{t} \rangle \vdash e_2 : \tau_2}{\Gamma_2 @ \mathbf{s} \vdash \lambda x. e_2 : \tau_1 \xrightarrow{\mathbf{t}} \tau_2}}{\Gamma_2, \Gamma_1 @ \mathbf{s} \times (\mathbf{t} \otimes \mathbf{r}) \vdash (\lambda x. e_2) e_1 : \tau_2}}{\Gamma_1, \Gamma_2 @ (\mathbf{t} \otimes \mathbf{r}) \times \mathbf{s} \vdash (\lambda x. e_2) e_1 : \tau_2}$$

The assumptions and conclusions match those of the (*let*) rule.  $\square$

#### 5.2.4 Examples of structural coeffacts

The structural coeffact calculus above can be instantiated to obtain the 3 structural coeffact calculi presented in Section 3.3. Two of them – structural data-flow and structural liveness provide a more precise tracking of properties that can be tracked using flat systems. Formally, any flat coeffact algebra can be turned into a structural coeffact scalar (by dropping the  $\wedge$  operator), but this does not always give us a meaningful system – for example, implicit parameters could be attached to individual variables, but, to our knowledge, this has not been used in any practical system.

On the other hand, some of the structural systems do not have a flat equivalent, typically because there is no appropriate  $\wedge$  operator that could be added to form the flat coeffact scalar. This is the case, for example, for the bounded variable use.

**Example 13** (Structural liveness). *The structural coeffact scalar for liveness is formed by  $(\mathcal{L}, \sqcap, \sqcup, L, D, \sqsubseteq)$ , where  $\mathcal{L} = \{L, D\}$  is the same two-point lattice as in the flat version, that is  $D \sqsubseteq L$  with a join  $\sqcup$  and a meet  $\sqcap$ .*

<sup>1</sup> Because of this duality, earlier version of coeffact systems [70] used  $\wedge$  and  $\vee$ .

**Example 14** (Structural data-flow). In data-flow, context is annotated with natural numbers and the structural coeffect scalar is formed by  $(\mathbb{N}, +, \max, 0, 0, \leq)$ .

For the two examples that have both flat and structural version, obtaining the structural coeffect algebra is easy. As shown by the examples above, we simply omit the  $\wedge$  operation. The laws required by a structural coeffect algebra are the same as those required by the flat version and so the above definitions are both valid. Similar construction can be used for the *optimized data-flow* example from Section 4.2.5.

It is important to note that this gives us a systems with *different* properties. The information are now tracked per-variable rather than for entire contexts. For data-flow, we also need to adapt the typing rule for the `prev` construct. Here, we write  $+$  for a pointwise extension of the  $+$  operator, such that  $\langle r_1, \dots, r_n \rangle + k = \langle r_1 + k, \dots, r_n + k \rangle$ .

$$(prev) \frac{\Gamma @ \mathbf{r} \vdash e : \tau}{\Gamma @ \mathbf{r} + 1 \vdash prev\ e : \tau}$$

The rule appears similar to the flat one, but there is an important difference. Because of the structural nature of the type system, it only increments the required number of values for variables that are used in the expression  $e$ . Annotations of other variables can be left unchanged.

Before looking at the semantics and equational properties of structural coeffect systems, we consider bounded variable use, which is an example of structural system that does not have a flat counterpart.

**Example 15** (Bounded variable reuse). The structural coeffect algebra for tracking bounded variable use is given by  $(\mathbb{N}, *, +, 1, 0, \leq)$

Similarly to the structural calculus for data-flow, the calculus for bounded variable reuse annotates each variable with an integer. However, the integer denotes how many times is the variable *accessed* rather than how many *past values* are needed. The resulting type system is the one shown in Figure 18 in Chapter 3.

### 5.3 CATEGORICAL MOTIVATION

When we introduced structural coeffect systems in Section 3.3, we included a concrete semantics of structural liveness and bounded variable reuse. In this section, we generalize the examples using the notion of *structural indexed comonad*, which is an extension of the *indexed comonad* structure. Similarly to *flat indexed comonad*, the structural variant adds operations that are needed to embed full  $\lambda$ -calculus, this time with per-variable contexts.

As in the previous chapter, the main aim of this section is to motivate and explain the design of the structural coeffect calculus shown in Section 5.2. However, the semantics also highlights the similarities and differences between the flat and structural system. Most of the differences arise from the fact that contexts in structural coeffect systems are treated as *vectors* rather than sets.

#### 5.3.1 Semantics of vectors

Recall that in the flat coeffect calculus, the context is interpreted as a product and so a typing judgement  $x_1 : \tau_1, \dots, x_n : \tau_n @ \mathbf{r} \vdash e : \tau$  is interpreted as a morphism  $C^{\mathbf{r}}(\tau_1 \times \dots \times \tau_n) \rightarrow \tau$ . In this model, we can freely transform the value contained in the context modelled using an indexed comonad  $C^{\mathbf{r}}$ .

Previously, we defined the `map` function (in terms of `cobind` and `counit`), which transforms the value inside the context without affecting the coeffect annotation. Thus, we can use  $\text{map}_r \pi_i$  to transform a context containing product of variables  $C^r(\tau_1 \times \dots \times \tau_n)$  into a context containing a single value  $C^r\tau_i$ . This changes the carried value without affecting the coeffect  $r$ .

The ability to freely transform the variable structure is not desirable in the model of structural coeffect systems. Our aim is to guarantee (by construction) that the structure of the coeffect annotations matches the structure of variables. To achieve this, we model vectors using a structure distinct from ordinary products which we denote  $-\hat{\times}-$ . For example, the judgement  $x_1 : \tau_1, \dots, x_n : \tau_n @ \langle r_1, \dots, r_n \rangle \vdash e : \tau$  is modelled as a morphism  $C^{\langle r_1, \dots, r_n \rangle}(\tau_1 \hat{\times} \dots \hat{\times} \tau_n) \rightarrow \tau$ . We assume that the operator is equipped with necessary associativity transformations allowing us to use it freely on more than two values.

The operator is a bifunctor, but it is *not* a product in the categorical sense. In particular, there is no way to turn  $\tau_1 \hat{\times} \dots \hat{\times} \tau_n$  into  $\tau_i$  (the structure does not have projections) and so there is also no way of turning  $C^{\langle r_1, \dots, r_n \rangle}(\tau_1 \hat{\times} \dots \hat{\times} \tau_n)$  into  $C^{\langle r_1, \dots, r_n \rangle}\tau_i$ , which would break the correspondence between coeffect annotations and variable structure.

The structure created using  $-\hat{\times}-$  can be manipulated only using operations provided by the *structural indexed comonad*, which operate over variable contexts contained in an indexed comonad  $C^r$  and are designed to preserve the correspondence between vectors and annotations.

In what follows, we model (finite) vectors of length  $n$  as  $\tau_1 \hat{\times} \dots \hat{\times} \tau_n$ . As mentioned, we assume that the use of the operator can be freely re-associated. For example, when calling an operation that requires input of the form  $(\tau_1 \hat{\times} \dots \hat{\times} \tau_i) \hat{\times} (\tau_{i+1} \hat{\times} \dots \hat{\times} \tau_n)$ , we use an argument  $(\tau_1 \hat{\times} \dots \hat{\times} \tau_n)$  and assume that the appropriate transformation is inserted.

### 5.3.2 Indexed comonads, revisited

The semantics of structural coeffect calculus reuses the definition of *indexed comonad* almost without change. The additional structure that is required for context manipulation (merging and splitting) is different and is here provided by the *structural indexed comonad* structure that we introduce in this section.

Recall the definition from Section 4.3.3, which defines an indexed comonad over a monoid  $(\mathcal{C}, \otimes, \text{use})$  as a triple  $(C^r, \text{counit}_{\text{use}}, \text{cobind}_{r,s})$ . The triple consists of a family of object mappings  $C^r$ , and two mappings that involve context-dependent morphisms of the form  $C^r\tau_1 \rightarrow \tau_2$ .

In the structural coeffect calculus, we work with morphisms of the form  $C^r\tau_1 \rightarrow \tau_2$  representing function values (appearing in the language), but also of the form  $C^{\langle r_1, \dots, r_n \rangle}(\tau_1 \hat{\times} \dots \hat{\times} \tau_n) \rightarrow \tau$ , modelling expressions in a context. To capture this, we need to revisit the definition and use *coeffect vectors* in some of the operations.

**Definition 14.** Given a monoid  $(\mathcal{C}, \otimes, \text{use})$  with a pointwise extension of the  $\otimes$  operator to a vector (written as  $\mathbf{t} \otimes \mathbf{s}$ ) and an operation lifting scalars to vectors  $\langle - \rangle$ , an indexed comonad over a category  $\mathcal{C}$  is a triple  $(C^r, \text{counit}_{\text{use}}, \text{cobind}_{s,r})$ :

- $C^r$  for all  $r \in \bigcup_{m \in \mathbb{N}} \mathcal{C}^m$  is a family of object mappings
- $\text{counit}_{\text{use}}$  is a mapping  $C^{\langle \text{use} \rangle} \alpha \rightarrow \alpha$
- $\text{cobind}_{s,r}$  is a mapping  $(C^r \alpha \rightarrow \beta) \rightarrow (C^{s \otimes r} \alpha \rightarrow C^{\langle s \rangle} \beta)$



The object mapping  $C^{\mathbf{r}}$  is now indexed by a vector rather than by a scalar  $C^r$  as in the previous chapter. This new definition supersedes the old one, because a flat coeffect annotation can be seen as singleton vectors.

The operation  $\text{counit}_{\text{use}}$  operates on a vector of length one. This means that it will always return a single value rather than a vector created using  $-\hat{\times}-$ . The  $\text{cobind}_{s,r}$  operation is, perhaps surprisingly, indexed by a coeffect vector and a coeffect scalar. This asymmetry is explained by the fact that the input function  $(C^{\mathbf{r}}\alpha \rightarrow \beta)$  takes a vector of variables, but always produces just a single value. Thus the resulting function also takes a vector of variables, but always returns a context with a vector containing just one value. In other words,  $\alpha$  may contain  $\hat{\times}$ , but  $\beta$  may not, because the coeffect calculus has no way of constructing values containing  $\hat{\times}$ .

### 5.3.3 Structural indexed comonads

The flat indexed comonad structure extends indexed comonads with operations  $\text{merge}_{r,s}$  and  $\text{split}_{r,s}$  that combine or split the additional (flat) context and are annotated with the flat coeffect operations  $\wedge$  and  $\oplus$ , respectively.

In the structural version, the corresponding operations operate convert between a (wrapped) vector of values represented using  $\hat{\times}$  and ordinary pairs of contexts containing parts of the vector. The vectors of coeffect annotations are split or merged using  $\times$  of the structural coeffect algebra, in a way that mirrors the wrapped vectors (variable structure).

The following definition includes  $\text{lift}_{r',r}$ , which is similar as before and models sub-coffecting and also  $\text{dup}_{r,s}$  which models duplication of a variable in a context needed for the semantics of contraction:

**Definition 15.** Given a structural coeffect algebra formed by  $(\mathcal{C}, \otimes, \oplus, \text{use}, \text{ign}, \leq)$  with operations  $\langle - \rangle$  and  $\otimes$ , a structural indexed comonad is an indexed comonad over the monoid  $(\mathcal{C}, \otimes, \text{use})$  equipped with families of operations  $\text{merge}_{r,s}$ ,  $\text{split}_{r,s}$ ,  $\text{dup}_{r,s}$  and  $\text{lift}_{r',r}$  where:

- $\text{merge}_{r,s}$  is a family of mappings  $C^{\mathbf{r}}\alpha \times C^{\mathbf{s}}\beta \rightarrow C^{\mathbf{r} \times \mathbf{s}}(\alpha \hat{\times} \beta)$
- $\text{split}_{r,s}$  is a family of mappings  $C^{\mathbf{r} \times \mathbf{s}}(\alpha \hat{\times} \beta) \rightarrow C^{\mathbf{r}}\alpha \times C^{\mathbf{s}}\beta$
- $\text{dup}_{r,s}$  is a family of mappings  $C^{\langle \mathbf{r} \oplus \mathbf{s} \rangle} \alpha \rightarrow C^{\langle \mathbf{r}, \mathbf{s} \rangle}(\alpha \hat{\times} \alpha)$
- $\text{lift}_{r',r}$  is a family of mappings  $C^{\langle \mathbf{r}' \rangle} \alpha \rightarrow C^{\langle \mathbf{r} \rangle} \alpha$  for all  $\mathbf{r}', \mathbf{r}$  such that  $\mathbf{r} \leq \mathbf{r}'$

Here, the following equalities must hold:

$$\text{merge}_{r,s} \circ \text{split}_{r,s} \equiv \text{id} \quad \text{id} \equiv \text{split}_{r,s} \circ \text{merge}_{r,s}$$

These operations differ from those of the flat indexed comonad in that the merge and split operations are required to be inverse functions and to preserve the additional information about the context. This was not required for the flat system where the operations could under-approximate or over-approximate. Note that the operations use  $\hat{\times}$  to combine or split the contained values. This means that they operate on free-variable vectors rather than on ordinary products.

The  $\text{dup}$  mapping is a new operation that was not required for a flat calculus. It takes a variable context with a single variable annotated with  $\mathbf{r} \oplus \mathbf{s}$ , duplicates the value of the variable  $\alpha$  and splits the additional context between the two new variables. In a flat calculus, this operation was expressed using ordinary tuple construction, which is not possible here – the returned context needs to contain a two-element vector  $\alpha \hat{\times} \alpha$ .

$$\begin{aligned}
& \llbracket x : \tau @ \langle \text{use} \rangle \vdash x : \tau \rrbracket ctx = \text{counit}_{\text{use}} ctx & (var) \\
& \llbracket \Gamma @ \text{ign} \vdash c_i : \tau \rrbracket ctx = \delta(c_i) & (const) \\
& \llbracket \Gamma_1, \Gamma_2 @ \mathbf{r} \times (\mathbf{t} \otimes \mathbf{s}) \vdash e_1 e_2 : \tau_2 \rrbracket ctx = & (app) \\
& \quad \text{let } (ctx_1, ctx_2) = \text{split}_{\mathbf{r}, \mathbf{t} \otimes \mathbf{s}} ctx \\
& \quad \text{in } \llbracket \Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{\mathbf{t}} \tau_2 \rrbracket ctx_1 (\text{cobind}_{\mathbf{t}, \mathbf{s}} \llbracket \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1 \rrbracket ctx_2) \\
& \llbracket \Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{s}} \tau_2 \rrbracket ctx = \lambda v. & (abs) \\
& \quad \llbracket \Gamma, x : \tau_1 @ \mathbf{r} \times \langle \mathbf{s} \rangle \vdash e : \tau_2 \rrbracket (\text{merge}_{\mathbf{r}, \langle \mathbf{s} \rangle} (ctx, v)) \\
& \llbracket \Gamma, x : \tau_1 @ \mathbf{r} \times \langle \text{ign} \rangle \vdash e : \tau \rrbracket ctx = & (weak) \\
& \quad \text{let } (ctx_1, \_) = \text{split}_{\mathbf{r}, \langle \text{ign} \rangle} ctx \text{ in } \llbracket \Gamma @ \mathbf{r} \vdash e : \tau \rrbracket ctx_1 \\
& \llbracket \Gamma_1, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} \rangle \times \mathbf{q} \vdash e : \tau \rrbracket ctx = & (sub) \\
& \quad \llbracket \Gamma_1, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}' \rangle \times \mathbf{q} \vdash e : \tau \rrbracket (\text{nest}_{\mathbf{r}, \langle \mathbf{s} \rangle, \langle \mathbf{s}' \rangle, \mathbf{q}} \text{lift}_{\mathbf{s}, \mathbf{s}'} ctx) \\
& \llbracket \Gamma_1, y : \tau_2, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{t}, \mathbf{s} \rangle \times \mathbf{q} \vdash e : \tau \rrbracket ctx = & (exch) \\
& \quad \llbracket \Gamma_1, x : \tau_1, y : \tau_2, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}, \mathbf{t} \rangle \times \mathbf{q} \vdash e : \tau \rrbracket \\
& \quad (\text{nest}_{\mathbf{r}, \langle \mathbf{t}, \mathbf{s} \rangle, \langle \mathbf{s}, \mathbf{t} \rangle, \mathbf{q}} \text{swap}_{\mathbf{t}, \mathbf{s}} ctx) \\
& \llbracket \Gamma_1, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} \oplus \mathbf{t} \rangle \times \mathbf{q} \vdash e[z, y \leftarrow x] : \tau \rrbracket ctx = & (contr) \\
& \quad \llbracket \Gamma_1, y : \tau_1, z : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}, \mathbf{t} \rangle \times \mathbf{q} \vdash e : \tau \rrbracket \\
& \quad (\text{nest}_{\mathbf{r}, \langle \mathbf{s} \oplus \mathbf{t} \rangle, \langle \mathbf{s}, \mathbf{t} \rangle, \mathbf{q}} \text{dup}_{\mathbf{s}, \mathbf{t}} ctx)
\end{aligned}$$

The definition uses the following auxiliary definitions:

$$\begin{aligned}
& \text{swap}_{\mathbf{t}, \mathbf{s}} : C^{\langle \mathbf{t}, \mathbf{s} \rangle}(\alpha \hat{\times} \beta) \rightarrow C^{\langle \mathbf{s}, \mathbf{t} \rangle}(\beta \hat{\times} \alpha) \\
& \text{swap}_{\mathbf{t}, \mathbf{s}} ctx = \\
& \quad \text{let } (ctx_1, ctx_2) = \text{split}_{\langle \mathbf{t} \rangle, \langle \mathbf{s} \rangle} ctx \\
& \quad \text{in } \text{merge}_{\langle \mathbf{s} \rangle, \langle \mathbf{t} \rangle} (ctx_2, ctx_1) \\
& \text{nest}_{\mathbf{r}, \mathbf{s}, \mathbf{s}', \mathbf{t}} : (C^{\mathbf{s}} \beta \rightarrow C^{\mathbf{s}'} \beta') \rightarrow C^{\mathbf{r} \times \mathbf{s} \times \mathbf{t}}(\alpha \hat{\times} \beta \hat{\times} \gamma) \rightarrow C^{\mathbf{r} \times \mathbf{s}' \times \mathbf{t}}(\alpha \hat{\times} \beta' \hat{\times} \gamma) \\
& \text{nest}_{\mathbf{r}, \mathbf{s}, \mathbf{s}', \mathbf{t}} f ctx = \\
& \quad \text{let } (ctx_1, ctx') = \text{split}_{\mathbf{r}, \mathbf{s} \times \mathbf{t}} ctx \\
& \quad \text{let } (ctx_2, ctx_3) = \text{split}_{\mathbf{s}, \mathbf{t}} ctx' \\
& \quad \text{in } \text{merge}_{\mathbf{r}, \mathbf{s}' \times \mathbf{t}} (ctx_1, \text{merge}_{\mathbf{s}', \mathbf{t}} (f ctx_2, ctx_3))
\end{aligned}$$

Figure 29: Categorical semantics of the structural coeffect calculus

Finally, the lift mapping is almost the same as in the flat version. It operates on a singleton vector, which is equivalent to operating on a scalar as before. The operation could be extended to a vector in a pointwise way, but we keep it simple and perform sub-coeffecting on individual variables.

### 5.3.4 Semantics of structural calculus

The concrete semantics for liveness and bounded variable use shown in Sections 3.3.1 and 3.3.2 suggests that semantics of structural coeffect calculi tend to be more complex than semantics of flat coeffect calculi. The complexity comes from the fact that we need a more expressive representation of the variable context – e.g. a vector of optional values (i.e. wrapped in the Maybe data type). Additionally, the structural system needs to pass separate variable contexts to the sub-expressions.



The latter aspect is fully captured by the semantics shown in this section. The earlier point is left to the concrete notion of structural coeffect. Our model still gives us the flexibility of defining the concrete representation of variable vectors. We explore a number of examples in Section 5.3.5 and start by looking at a unified categorical semantics defined in terms of *structural indexed comonads*.

**CONTEXTS AND FUNCTIONS.** In the structural coeffect calculus, expressions in context are interpreted as functions taking a vector (represented using  $-\hat{\times}-$ ) wrapped in a structure indexed with a vector of annotations such as  $C^{\mathbf{r}}$ . Functions take only a single variable as an input and so the structure is annotated with a scalar, such as  $C^{\mathbf{r}}$ , which we treat as being equivalent to a singleton vector annotation  $C^{\langle \mathbf{r} \rangle}$ :

$$\begin{aligned} \llbracket x_1 : \tau_1, \dots, x_n : \tau_n @ \langle \mathbf{r}_1, \dots, \mathbf{r}_n \rangle \vdash e : \tau \rrbracket & : C^{\langle \mathbf{r}_1, \dots, \mathbf{r}_n \rangle}(\tau_1 \hat{\times} \dots \hat{\times} \tau_n) \rightarrow \tau \\ \llbracket \tau_1 \xrightarrow{\mathbf{r}} \tau_2 \rrbracket & = C^{\langle \mathbf{r} \rangle} \tau_1 \rightarrow \tau_2 \end{aligned}$$

Note that the instances of flat indexed comonad ignored the fact that the variable context wrapped in the data structure is a product. This is not generally the case for the structural indexed comonads – the definitions shown in Section 5.3.5 are given specifically for  $C^{\langle \mathbf{r}_1, \dots, \mathbf{r}_n \rangle}(\tau_1 \hat{\times} \dots \hat{\times} \tau_n)$  rather than more generally for  $C^{\mathbf{r}}\alpha$ . The need to examine the structure of the variable context is another reason for using  $-\hat{\times}-$  when interpreting expressions in contexts.

**EXPRESSIONS.** A semantics of structural coeffect calculi is shown in Figure 29. As in the previous chapter, the semantics is written in a programming language style using constructs such as let-binding rather than using a categorical (point-free) notation. As before, the semantics can equally be written using primitives such as currying, uncurrying, function pairing etc.

The following summarizes how the standard syntax-driven rules work, highlighting the differences from the flat version:

- When accessing a variable (*var*), the context now contains *only* the accessed variable and so the semantics is just  $\text{count}_{\text{use}}$  without a projection. Constants (*const*) are interpreted using a global dictionary  $\delta$  as earlier.
- The semantics of flat function application first duplicated the context so that the same variables can be passed to both sub-expressions. This is no longer needed – the (*app*) rule splits the variables *including* the additional context into two parts. Passing the first context to the semantics of  $e_1$  gives us a function  $C^{\langle \mathbf{t} \rangle} \tau_1 \rightarrow \tau_2$ .

A value  $C^{\langle \mathbf{t} \rangle} \tau_1$  required in order to call the function is obtained by applying  $\text{cobind}_{\mathbf{t}, \mathbf{s}}$  to the semantics of  $e_2$ . The result  $C^{\mathbf{t} \otimes \mathbf{s}}(\dots \hat{\times} \dots \hat{\times} \dots) \rightarrow C^{\langle \mathbf{t} \rangle} \tau_1$  is then called with the latter part of the split input context.

- The semantics of function abstraction (*abs*) is syntactically the same as in the flat version – the only difference is that we now merge a free-variable context with a singleton vector, both at the level of variable assignments and at the level of coeffect annotations.

The semantics for the non-syntax-driven rules (weakening, exchange, contraction) performs transformations on the free-variable context. Weakening (*weak*) splits the context and ignores the part corresponding to the removed

variable. If we were modelling the semantics in a language with a linear type system, this would require an additional operation for ignoring an unused context annotated with  $\text{ign}$ .

The remaining rules perform a transformation anywhere inside the free-variable vector. To simplify writing the semantics, we define a helper  $\text{nest}_{\mathbf{r}, \mathbf{s}, \mathbf{s}', \mathbf{t}}$  that splits the variable vector into three parts, transforms the middle part and then merges them, using the newly transformed middle part.

The transformations on the middle part are quite simple. The  $(\text{sub})$  rule uses  $\text{lift}_{\mathbf{s}, \mathbf{s}'}$  to discard some of the available additional context; the  $(\text{exch})$  rule swaps two single-variable contexts and the  $(\text{contr})$  rule uses the  $\text{dup}_{\mathbf{s}, \mathbf{t}}$  operation to duplicate a variable while splitting its additional context.

**PROPERTIES.** As in the flat calculus, the main reason for defining the categorical semantics in this chapter is to provide validation for the design of the calculus. The following correspondence theorem states that the annotations in the typing rules of the structural coeffect calculus correspond to the indices of the semantics. Thus, the calculus captures a context-dependent property if it can be modelled by a *structural indexed comonad*. As we show in the next section, this is the case for all three discussed examples (liveness, data-flow, bounded variable reuse).

**Theorem 13** (Correspondence). *In all of the typing rules of the structural coeffect system, the context annotations  $\mathbf{r}$  and  $\mathbf{s}$  of typing judgements  $\Gamma @ \mathbf{r} \vdash e : \tau$  and function types  $\tau_1 \xrightarrow{\mathbf{s}} \tau_2$  correspond to the indices of mappings  $C^{\mathbf{r}}$  and  $C^{\langle \mathbf{s} \rangle}$  in the corresponding semantic function defined by  $\llbracket \Gamma @ \mathbf{r} \vdash e : \tau \rrbracket$ .*

*Proof.* By analysis of the semantic rules in Figure 29.  $\square$

As in the flat calculus, the primitive operations of the structural indexed comonad are each annotated with a unique operations provided by the coeffect annotations. This means that the semantics fully determines the structure of the typing rules of the structural coeffect calculus.

### 5.3.5 Examples of structural indexed comonads

The categorical semantics for structural coeffect calculus is easily instantiated to give semantics for a concrete calculus. In this section, we revisit the three examples discussed throughout this chapter – structural liveness, data-flow and bounded variable reuse. Some aspects of the first two examples will be similar to flat versions discussed in Section 4.3 – they are based on the same data structures (option and a list, respectively), but the data structures are composed differently. Generally speaking, rather than having a data structure over a product of variables, we now have a vector of variables over a specific data structure.

The abstract semantics does not specify how vectors of variables should be represented, so this can vary in concrete instantiations. In all our examples, we represent a vector of variables as a product written using  $\times$ . To distinguish between products representing vectors and ordinary products (e.g. a product of contexts returned by  $\text{split}$ ), we write vectors using  $\langle a, \dots, b \rangle$  rather than the parentheses, used for ordinary tuples.

**DATA-FLOW.** It is interesting to note that the semantics of data-flow and bounded variable reuse (discussed next) both keep a product of multiple values for each variable, so they are both built around an *indexed list* data structure. However, their  $\text{cobind}$  and  $\text{dup}$  operations work differently. We start by

looking at the structure modelling data-flow computations. For readability, variables in bold face (such as  $\mathbf{a}_i$ ) range over vectors while ordinary notation (such as  $a_i$ ) is used for individual values.

**Example 16** (Indexed list for data-flow). *The indexed list model of data-flow computations is defined over a structural coefficient algebra  $(\mathbb{N}, +, \max, 0, 0, \leq)$ . The data type  $C^{(n_1, \dots, n_k)}$  is indexed by required number of past variables for each individual variable. It is defined over a vector of variables  $\alpha_1 \hat{\times} \dots \hat{\times} \alpha_k$  and it keeps a product containing a current value followed by  $n_i$  past values:*

$$C^{(n_1, \dots, n_k)}(\alpha_1 \hat{\times} \dots \hat{\times} \alpha_k) = \underbrace{(\alpha_1 \times \dots \times \alpha_1)}_{(n_1+1)\text{-times}} \times \dots \times \underbrace{(\alpha_k \times \dots \times \alpha_k)}_{(n_k+1)\text{-times}}$$

The mappings that define the structural indexed comonad include the split and merge operations that are shared by the other two examples (discussed below):

$$\begin{aligned} \text{merge}_{\langle m_1, \dots, m_k \rangle, \langle n_1, \dots, n_l \rangle}(\langle \mathbf{a}_1, \dots, \mathbf{a}_k \rangle, \langle \mathbf{b}_1, \dots, \mathbf{b}_l \rangle) &= \\ \langle \mathbf{a}_1, \dots, \mathbf{a}_k, \mathbf{b}_1, \dots, \mathbf{b}_l \rangle & \\ \text{split}_{\langle m_1, \dots, m_k \rangle, \langle n_1, \dots, n_l \rangle}(\langle \mathbf{a}_1, \dots, \mathbf{a}_k, \mathbf{b}_1, \dots, \mathbf{b}_l \rangle) &= \\ \langle \langle \mathbf{a}_1, \dots, \mathbf{a}_k \rangle, \langle \mathbf{b}_1, \dots, \mathbf{b}_l \rangle \rangle & \end{aligned}$$

The remaining mappings that are required by structural indexed comonad and capture the essence of data-flow computations are defined as:

$$\begin{aligned} \text{counit}_0 \langle \langle a_0 \rangle \rangle &= a_0 \\ \text{cobind}_{m, \langle n_1, \dots, n_k \rangle} f \langle \langle a_{1,0}, \dots, a_{1,m+n_1} \rangle, \dots, \langle a_{k,0}, \dots, a_{k,m+n_k} \rangle \rangle &= \\ \langle \langle f \langle \langle a_{1,0}, \dots, a_{1,n_1} \rangle, \dots, \langle a_{k,0}, \dots, a_{k,n_k} \rangle \rangle, \dots, & \\ f \langle \langle a_{1,m}, \dots, a_{1,m+n_1} \rangle, \dots, \langle a_{k,m}, \dots, a_{k,m+n_k} \rangle \rangle \rangle & \\ \text{dup}_{m,n} \langle \langle a_1, \dots, a_{\max(m,n)} \rangle \rangle &= \langle \langle a_1, \dots, a_m \rangle, \langle a_1, \dots, a_n \rangle \rangle \\ \text{lift}_{k',k} \langle \langle a_0, \dots, a_{k'} \rangle \rangle &= \langle \langle a_0, \dots, a_k \rangle \rangle \quad (\text{when } k \leq k') \end{aligned}$$

The definition of the indexed list data structure relies on the fact that the number of annotations corresponds to the number of variables combined using  $\hat{\times}$ . It then creates a vector of lists containing  $n_i + 1$  values for the  $i$ -th variable (the annotation represents the number of required *past* values so one more value is required).

The split and merge operations are defined separately, because they are not specific to the example. They operate on the top-level vectors of variables (without looking at the representation of the variable). This means that we can re-use the same definitions for the following two examples (with the only difference that  $\mathbf{a}_i, \mathbf{b}_i$  will there represent options rather than lists).

The mappings that explain how data-flow computations work are *cobind* (representing sequential composition) and *dup* (representing context sharing or parallel composition). In *cobind*, we get  $k$  vectors corresponding to  $k$  variables, each with  $m + n_i$  values. The operation calls  $f$   $m$ -times to obtain  $m$  past values required as the result of type  $C^{(m)}\beta$ .

The *dup* <sub>$m,n$</sub>  operation needs to produce a two-variable context containing  $m$  and  $n$  values, respectively, of the input variable. The input provides  $\max(m, n)$  values, so the definition is simply a matter of restriction. Finally, *counit* extracts the value of its single variable and *lift* drops additional past values that are not required.

**BOUNDED REUSE.** As mentioned earlier, the semantics of calculus for bounded reuse is also based on the indexed list structure. Rather than representing possibly different past values that can be shared (see the definition of  $\text{dup}$ ), the list now represents multiple copies of the same value as each value can only be accessed once. This semantics follows that of Girard [35].

**Example 17** (Indexed list for bounded reuse). *The indexed list model of bounded variable reuse is defined over a structural coeffect algebra  $(\mathbb{N}, *, +, 1, 0, \leq)$ . The data type  $C^{(n_1, \dots, n_k)}$  is a vector containing  $n_i$  values of  $i$ -th variable:*

$$C^{(n_1, \dots, n_k)}(\alpha_1 \hat{\times} \dots \hat{\times} \alpha_k) = \underbrace{(\alpha_1 \times \dots \times \alpha_1)}_{n_1\text{-times}} \times \dots \times \underbrace{(\alpha_k \times \dots \times \alpha_k)}_{n_k\text{-times}}$$

The merge and split operations are defined as in Example 16. The operations that capture the behaviour of bounded reuse are:

$$\text{counit}_1 \langle\langle a_0 \rangle\rangle = a_0$$

$$\text{lift}_{k',k} \langle\langle a_0, \dots, a_{k'} \rangle\rangle = \langle\langle a_0, \dots, a_k \rangle\rangle \quad (\text{when } k \leq k')$$

$$\text{dup}_{m,n} \langle\langle a_1, \dots, a_{m+n} \rangle\rangle = \langle\langle a_1, \dots, a_m \rangle, \langle a_{m+1}, \dots, a_{m+n} \rangle\rangle$$

$$\begin{aligned} \text{cobind}_{m, \langle n_1, \dots, n_k \rangle} f \langle\langle a_{1,0}, \dots, a_{1,m*n_1} \rangle, \dots, \langle a_{k,0}, \dots, a_{k,m*n_k} \rangle\rangle = \\ \langle f \langle\langle a_{1,0}, \dots, a_{1,n_1-1} \rangle, \dots, \langle a_{k,0}, \dots, a_{k,n_k-1} \rangle\rangle, \dots, \\ f \langle\langle a_{1,(m-1)*n_1}, \dots, a_{1,(m-1)*n_1} \rangle, \dots, \langle a_{k,m*n_k-1}, \dots, a_{k,m*n_k-1} \rangle\rangle \rangle \end{aligned}$$

The counit and lift operations are defined as previously – variable access extracts the only value of the only variable and sub-coeffecting allows discarding multiple copies of a value that are not needed.

In the bounded variable reuse system, variable sharing is annotated with the  $+$  operator (in contrast with  $\text{max}$  used in data-flow). The  $\text{dup}_{m,n}$  operation thus splits the  $m+n$  available values between two vectors of length  $m$  and  $n$ , without *sharing* a value. The  $\text{cobind}$  operation works similarly – it splits  $m * n_i$  available values of each variable into  $m$  vectors containing  $n_i$  copies and then calls the  $f$  function  $m$ -times to obtain  $m$  resulting values without sharing any input value.

**LIVENESS.** In both data-flow and bounded reuse, the data type is defined as a vector of values obtained by applying the indexed list type constructor to types of individual variables. We can generalize this pattern. Given a parameterized (indexed) type constructor  $D^{\mathbf{l}}\alpha$ , we define  $C^{(l_1, \dots, l_n)}$  in terms of a vector of  $D^{l_i}$  types. For liveness, the definition lets us reuse some of the mappings used when defining the semantics of flat liveness. However, we cannot fully define the semantics of the structural version in terms of the flat version – the  $\text{cobind}$  operation is different and we need an appropriate  $\text{dup}$  operation.

**Example 18** (Structural indexed option). *Given a structural coeffect algebra formed by  $(\{L, D\}, \sqcap, \sqcup, L, D, \sqsubseteq)$  and the indexed option data type  $D^{\mathbf{l}}$ , such that  $D^D\alpha = 1$  and  $D^L\alpha = \alpha$ , the data type for structural indexed option comonad is:*

$$C^{(n_1, \dots, n_k)}(\alpha_1 \hat{\times} \dots \hat{\times} \alpha_k) = D^{n_1}\alpha_1 \times \dots \times D^{n_k}\alpha_k$$

The merge and split operations are defined as earlier. The remaining operations model variable liveness as follows:

$$\begin{aligned}
 \text{cobind}_{L, \langle l_1, \dots, l_n \rangle} f \langle a_1, \dots, a_n \rangle &= \langle f \langle a_1, \dots, a_n \rangle \rangle \\
 \text{cobind}_{D, \langle D, \dots, D \rangle} f \langle () , \dots, () \rangle &= \langle D \rangle \\
 \text{dup}_{D, D} \langle () \rangle &= \langle () , () \rangle & \text{counit}_L \langle a \rangle &= a \\
 \text{dup}_{L, D} \langle a \rangle &= \langle a , () \rangle & \text{lift}_{L, L} \langle a \rangle &= \langle a \rangle \\
 \text{dup}_{D, L} \langle a \rangle &= \langle () , a \rangle & \text{lift}_{L, D} \langle a \rangle &= \langle () \rangle \\
 \text{dup}_{L, L} \langle a \rangle &= \langle a , a \rangle & \text{lift}_{D, D} \langle () \rangle &= \langle () \rangle
 \end{aligned}$$

When the expected result of the cobind operation is dead (second case), the operation can ignore all inputs and directly return the unit value  $()$ . Otherwise, it passes the vector of input variables to  $f$  as-is – no matter whether the individual values are live or dead. The  $L$  annotation is a unit with respect to  $\cap$  and so the annotations expected by  $f$  are the same as those required by the result of cobind.

The dup operation resembles with the flat version of split – this is expected as duplication in the flat calculus is performed by first duplicating the variable context (using map) and then applying split. Here, the duplication returns a pair. Depending on the required coeffect annotations, this may copy (duplicate) the value, or it may produce an empty context.

Finally, counit extracts a value which is always present as guaranteed by the type  $C^{(L)} \alpha \rightarrow \alpha$ . The lifting operation models sub-coeffecting which can turn a context with a value into a dead context (second case); otherwise it behaves as identity.

#### 5.4 SYNTACTIC EQUATIONAL THEORY

Similarly to the flat version, each concrete instance of the structural coeffect calculus has a different notion of context. As before, the properties of the flat coeffect algebra guarantee that certain equational properties hold in all instances of the structural coeffect calculus.

In this section, we look at these common properties that we get “for free” from the structural coeffect calculus. We start the discussion by briefly considering the key aspects that make the equational theory of flat and structural coeffects different.

##### 5.4.1 From flat coeffects to structural coeffects

When discussing syntactic reductions for the flat calculus in Section 4.4, we noted that no single technique of specifying syntactic reduction works universally for all flat coeffect calculi. We considered multiple alternatives (call-by-name, call-by-value, internalized substitution) that can be used as the basis for operational semantics for different calculi that satisfy certain additional properties.

The structural coeffect calculus has more preferable equational properties. In particular, we can prove both  $\beta$ -reduction and  $\eta$ -expansion using just the properties of structural coeffect algebra. For this reason, we focus on these two reductions in this section. Using the terminology of Pfenning and Davies [73], the structural coeffect calculus satisfies both the *local soundness* and the *local completeness* properties, witnessed by  $\beta$ -reduction and  $\eta$ -expansion, respectively.

**SUBSTITUTION FOR FLAT COEFFECTS.** The less obvious (*top-pointed*) variant of the substitution lemma for flat coeffects (Lemma 5) required all operations of the flat coeffect algebra to coincide. This enables substitution to preserve the type of expressions, because all additional requirements arising as the result of the substitution can be associated with the declaration context. For example, consider the following example where implicit parameter  $?offset$  is substituted for the variable  $y$ :

$$\begin{array}{lcl} y:\text{int} @ \emptyset \vdash \lambda x. y + ?total & : \text{int} \xrightarrow{\{?total\}} \text{int} & (\text{before}) \\ () @ \{?offset\} \vdash \lambda x. ?offset + ?total & : \text{int} \xrightarrow{\{?total\}} \text{int} & (\text{after}) \end{array}$$

The typing judgement obtained in (*after*) preserves the type of the expression (function value) from the original typing (*before*). This is possible thanks to the non-determinism involved in the typing rule for lambda abstraction – as all operators of the flat coeffect algebra used here are  $\cup$ , we can place the additional requirement on the outer context. Note that this is not the *only* possible typing, but it is *permissible* typing.

Here, the flat coeffect calculus gives us typing with limited *precision*, but enough *flexibility* to prove the substitution lemma.

**SUBSTITUTION FOR STRUCTURAL COEFFECTS.** In contrast, the substitution lemma (Lemma 14, shown later on page 111) for structural coeffects can be proven because structural coeffect systems provide enough *precision* to identify exactly with which variable should a context requirement be associated.

The following example shows a situation similar to the previous one. Here, we use structural data-flow calculus (writing  $\text{prev } e$  to obtain previous value of the expression  $e$ ) and we substitute  $w + z$  for  $y$ :

$$\begin{array}{lcl} y:\text{int} @ \langle 2 \rangle \vdash \lambda x. \text{prev } (x + \text{prev } y) & : \text{int} \xrightarrow{1} \text{int} & (\text{before}) \\ w:\text{int}, z:\text{int} @ 2 * \langle 1, 1 \rangle \vdash \lambda x. \text{prev } (x + \text{prev } (w + z)) & : \text{int} \xrightarrow{1} \text{int} & (\text{after}) \\ w:\text{int}, z:\text{int} @ \langle 2, 2 \rangle \vdash \lambda x. \text{prev } (x + \text{prev } (w + z)) & : \text{int} \xrightarrow{1} \text{int} & (\text{final}) \end{array}$$

The type of the function does not change, because the structural type system associates the annotation  $1$  with the bound variable  $x$  and the substitution does not affect how the variable  $x$  is used.

The other aspect demonstrated in the example is how the coeffect of the substituted variable affects the free-variable context of the substituted expression. Here, the original variable  $y$  is annotated with  $2$  and we substitute it for an expression  $w + z$  with free variables  $w, z$  annotated with  $\langle 1, 1 \rangle$ . The substitution applies the operation  $\otimes$  (modelling sequential composition) to the annotation of the new context – in the above example  $2 * \langle 1, 1 \rangle = \langle 2, 2 \rangle$ .

#### 5.4.2 Holes and substitution lemma

As demonstrated in the previous section, reduction (and substitution) in the structural coeffect calculus may need to replace a *single* variable with a *vector* of variables. More importantly, because the system uses explicit contraction, we may also need to substitute for multiple variables in the variable context at the same time.

Consider the expression  $\lambda x. x + x$ . It is type-checked by type-checking  $x_1 + x_2$ , contracting  $x_1$  and  $x_2$  and then applying lambda abstraction. During the reduction of  $(\lambda x. x + x) (y + z)$  we need to substitute  $y_1 + z_1$  for  $x_1$  and  $y_2 + z_2$  for  $x_2$ . This is similar to substitution lemma in other structural

variants of  $\lambda$ -calculus, such as the bunched typing system [60]. To express the substitution lemma later in this section, we first define the notion of a *context with holes*:

**Definition 16** (Context with holes). *A context with holes is a context such as  $x_1 : \tau_1, \dots, x_k : \tau_k @ \langle r_1, \dots, r_k \rangle$ , where some of the variable typings  $x_i : \tau_i$  and corresponding coeffacts  $r_i$  are replaced by holes written as  $-$ .*

$$\Delta[-@-]_n = \Delta[\underbrace{-@- \mid \dots \mid -@-}_{n\text{-times}}]$$

$$\Delta[-@-]_n := -, \Gamma @ \langle - \rangle \times s \quad \text{where } \Gamma @ s \in \Delta[-@-]_{n-1}$$

$$\Delta[-@-]_n := x : \tau, \Gamma @ \langle r \rangle \times s \quad \text{where } \Gamma @ s \in \Delta[-@-]_n$$

$$\Delta[-@-]_0 := () @ \langle \rangle$$

A context with  $n$  holes may either start with a hole, followed by a context with  $n - 1$  holes, or it may start with a variable followed by a context with  $n$  holes. Note that the definition ensures that the locations of variable holes correspond to the locations of coeffact annotation holes. Given a context with holes, we can fill the holes with other contexts using the *hole filling* operation and obtain an ordinary coeffact-annotated context.

**Definition 17** (Hole filling). *Given a context with  $n$  holes  $\Delta @ s \in \Delta[-@-]_n$ , the hole filling operation written as  $\Delta @ s[\Gamma_1 @ r_1 \mid \dots \mid \Gamma_n @ r_n]$ , which replaces the holes by the specified variables and corresponding coeffact annotations, is defined as:*

$$-, \Delta @ \langle - \rangle \times s [\Gamma_1 @ r_1 \mid \Gamma_2 @ r_2 \mid \dots] = \Gamma_1, \Gamma_2 @ r_1 \times r_2$$

$$\text{where } \Gamma_2 @ r_2 = \Delta @ s[\Gamma_2 @ r_2 \mid \dots]$$

$$x_1 : \tau, \Delta @ \langle r_1 \rangle \times s [\Gamma_1 @ r_1 \mid \Gamma_2 @ r_2 \mid \dots] = x_1 : \tau, \Gamma_2 @ \langle r_1 \rangle \times r_2$$

$$\text{where } \Gamma_2 @ r_2 = \Delta @ s[\Gamma_1 @ r_1 \mid \Gamma_2 @ r_2 \mid \dots]$$

$$() @ \langle \rangle [] = () @ \langle \rangle$$

When we substitute an expression with coeffacts  $t$  (associated with variables  $\Gamma$ ) for a variable that has coeffacts  $s$ , the resulting coeffacts of  $\Gamma$  need to combine  $t$  and  $s$ . Unlike in the flat coeffact systems, the structural substitution does not require all coeffact algebra operations to coincide and so the combination is more interesting than in the bottom-pointed substitution for flat coeffacts, where it used the only available operator (Lemma 5).

Substitution can be seen as a form of sequencing, i.e. the term  $e_1[x \leftarrow e_2]$  should have the same type as the term **let**  $x = e_2$  **in**  $e_1$ . In other words, we first need to obtain the value of the expression (requiring  $t$ ) and then use it in context with requirements  $s$ . Thus the free variables of the expression *after* substitution are annotated with  $s \circledast t$ , using the (scalar-vector extension) of the sequential composition operator  $\circledast$ .

An operational reading of the resulting coeffact annotation is that, to evaluate each of the variables annotated with coeffacts in the vector  $t$ , we first need to evaluate the substituted expression with coeffacts  $s$ , followed by the rest of the expression with coeffacts specified by  $t$ .



**Lemma 14** (Multi-nary substitution). *Given an expression with multiple holes filled by variables  $x_i : \tau_i$  with coeffects  $s_k$ :*

$$\Gamma @ \mathbf{r} [x_1 : \tau_1 @ \langle \mathbf{s}_1 \rangle \mid \dots \mid x_k : \tau_k @ \langle \mathbf{s}_k \rangle] \vdash e_r : \tau_r$$

*and a expressions  $e_i$  with free-variable contexts  $\Gamma_i$  annotated with  $\mathbf{t}_i$ :*

$$\Gamma_1 @ \mathbf{t}_1 \vdash e_1 : \tau_1 \quad \dots \quad \Gamma_k @ \mathbf{t}_k \vdash e_k : \tau_k$$

*then substituting the expressions  $e_i$  for variables  $x_i$  results in an expression with a context where the original holes are filled by contexts  $\Gamma_i$  with coeffects  $s_i @ \mathbf{t}_i$ :*

$$\Gamma @ \mathbf{r} [\Gamma_1 @ \mathbf{s}_1 @ \mathbf{t}_1 \mid \dots \mid \Gamma_k @ \mathbf{s}_k @ \mathbf{t}_k] \vdash e_r[x_1 \leftarrow e_1] \dots [x_k \leftarrow e_k] : \tau_r$$

*Proof.* By induction over  $\vdash$ , using the multi-nary aspect of the substitution in the proof of the contraction case (see Appendix B.2).  $\square$

#### 5.4.3 Reduction and expansion

In the Chapter 4, we discussed call-by-value separately from call-by-name, because the proof of call-by-value substitution has fewer prerequisites. In this section, we consider full  $\beta$ -reduction, which encompasses both call-by-value and call-by-name. We also show that  $\eta$ -expansion preserves the types. Both of the properties hold for a system with any structural coeffect algebra.

**REDUCTION THEOREM.** In a full  $\beta$ -reduction, written as  $\rightarrow_\beta$ , we can replace the redex  $(\lambda x. e_2) e_1$  by the expression  $e_r[x \leftarrow e_s]$  anywhere inside a term. The subject reduction theorem guarantees that this does not change the type of the term.

**Theorem 15** (Type preservation). *In a structural coeffect system with a structural coeffect algebra formed by  $(\mathcal{C}, *, \oplus, \text{use}, \text{ign}, \leq)$  and operations  $\langle - \rangle$  and  $\otimes$ , if  $\Gamma @ \mathbf{r} \vdash e : \tau$  and  $e \rightarrow_\beta e'$  using the full  $\beta$ -reduction then  $\Gamma @ \mathbf{r} \vdash e' : \tau$ .*

*Proof.* Consider the typing derivation for the redex  $(\lambda x. e_r) e_s$  before the reduction (note that this is similar to the typing of let binding discussed in Remark 12, but we do not swap the two parts of the free-variable context):

$$\frac{\frac{\Gamma_r, x : \tau_s @ \mathbf{r} \times \langle \mathbf{t} \rangle \vdash e_r : \tau_r}{\Gamma_s @ \mathbf{s} \vdash e_s : \tau_s} \quad \Gamma_r @ \mathbf{r} \vdash \lambda x. e_r : \tau_s \xrightarrow{\mathbf{t}} \tau_r}{\Gamma_r, \Gamma_s @ \mathbf{r} \times (\mathbf{t} @ \mathbf{s}) \vdash (\lambda x. e_r) e_s : \tau_r}$$

For the substitution lemma, we first rewrite the typing judgement for  $e_r$ , i. e.  $\Gamma_r, x : \tau_s @ \mathbf{r} \times \langle \mathbf{t} \rangle \vdash e_r : \tau_r$  as a context with a single hole filled by the  $x$  variable:  $\Gamma_r, - @ \mathbf{r} \times - [x : \tau_s @ \langle \mathbf{t} \rangle] \vdash e_r : \tau_r$ . Now we can perform the substitution using Lemma 14:

$$\frac{\frac{\Gamma_r, - @ \mathbf{r} \times - [x : \tau_s @ \langle \mathbf{t} \rangle] \vdash e_r : \tau_r \quad \Gamma_s @ \mathbf{s} \vdash e_s : \tau_s}{\Gamma_r, - @ \mathbf{r} \times - [\Gamma_s @ \mathbf{t} @ \mathbf{s}] \vdash e_r[x \leftarrow e_s] : \tau_r}}{\Gamma_r, \Gamma_s @ \mathbf{r} \times (\mathbf{t} @ \mathbf{s}) \vdash e_r[x \leftarrow e_s] : \tau_r}$$

The last step applies the hole filling operation, showing that substitution preserves the type of the term.  $\square$

Because of the vector (free monoid) structure of coeffect annotations  $\mathbf{r}$ ,  $\mathbf{s}$ , and  $\langle \mathbf{t} \rangle$ , these are uniquely associated with  $\Gamma_r$ ,  $\Gamma_s$ , and  $x$  respectively. Therefore, substituting  $e_s$  (which has coeffects  $\mathbf{s}$ ) for  $x$  introduces the context requirements specified by  $\mathbf{s}$  which are composed with the requirements  $\mathbf{t}$  associated with  $x$ , i. e. the variable being substituted.

**EXPANSION THEOREM.** Structural coeffect systems also exhibit  $\eta$ -equality, therefore satisfying both *local soundness* and *local completeness* as required by Pfenning and Davies [73]. Informally, this means that abstraction does not introduce too much, and application does not eliminate too much.

**Theorem 16** ( $\eta$ -expansion). *In a structural coeffect system with a structural coeffect algebra formed by  $(\mathbb{C}, \otimes, \oplus, \text{use}, \text{ign}, \leq)$  and operations  $\langle - \rangle$  and  $\otimes$ , if  $\Gamma @ \mathbf{r} \vdash e : \tau$  and  $e \rightarrow_\eta e'$  using the full  $\eta$ -reduction then  $\Gamma @ \mathbf{r} \vdash e' : \tau$ .*

*Proof.* The following derivation shows that  $\lambda x.f \ x$  has the same type and coeffects as the original expression  $f$ :

$$\frac{\frac{\Gamma @ \mathbf{r} \vdash f : \tau_1 \xrightarrow{s} \tau_2 \quad x : \tau_1 @ \langle \text{use} \rangle \vdash x : \tau_1}{\Gamma, x : \tau_1 @ \mathbf{r} \times (s \otimes \langle \text{use} \rangle) \vdash f \ x : \tau_2}}{\Gamma, x : \tau_1 @ \mathbf{r} \times \langle s \rangle \vdash f \ x : \tau_2} \quad \Gamma @ \mathbf{r} \vdash \lambda x.f \ x : \tau_1 \xrightarrow{s} \tau_2$$

The second step uses the fact that  $s \otimes \langle \text{use} \rangle = \langle s \otimes \text{use} \rangle = \langle s \rangle$  arising from the monoid  $(\mathbb{C}, \otimes, \text{use})$  of the scalar coeffect structure.  $\square$

The  $\eta$ -expansion property discussed in this section highlights another difference between coeffects and effects. The  $\eta$ -equality property does not hold for many notions of effect. For example, in a language with output effects,  $e = (\text{print "hi"}; (\lambda x.x))$  has different effects to its  $\eta$ -converted form  $\lambda x.ex$  because the immediate effects of  $e$  are hidden by the purity of  $\lambda$ -abstraction. In the coeffect calculus, the *(abs)* rule allows immediate contextual requirements of  $e$  to “float outside” of the enclosing  $\lambda$ . Furthermore, the free monoid nature of  $\times$  in structural coeffect systems allows the exact immediate requirements of  $\lambda x.ex$  to match those of  $e$ .

## 5.5 CONCLUSIONS

This chapter completes the key development of this thesis – the presentation of the coeffect framework, consisting of two calculi capturing properties of context-aware computations introduced in Chapter 3. In the previous chapter, we focused on the whole-context properties of computations and we developed *flat coeffect calculus* to capture them. This chapter develops *structural coeffect calculus*, which captures *per-variable* contextual properties. The system provides a precise analysis of liveness and data-flow and allows other interesting uses such as tracking of variable accesses based on bounded linear logic.

Following the structure of the previous chapter, the structural coeffect calculus is parameterized by a *structural coeffect algebra*. The two definitions are similar – both require operations  $\otimes$  and  $\oplus$  that model sequential and pointwise composition, respectively. For flat coeffects, we required  $\wedge$  to model context merging. For structural coeffects, we instead use a vector (free monoid) with the  $\times$  operation – which serves a similar purpose as  $\wedge$ .

In order to keep track of separate annotations for each variable, we use a system with explicit structural rules (contraction, weakening and exchange) that manipulate the structure of variables and the structure of annotations at the same time.

The chapter presents two technical results. The first is the semantics of structural coeffect calculus in terms of *structural indexed comonads*. The semantics serves two purposes – it unifies our three examples (bounded reuse, data-flow and liveness) and it demonstrates suitability of the type system.

The second technical result is equational theory for the structural coeffect calculus. In particular, we show that  $\beta$ -reduction and  $\eta$ -expansion have the typing preservation property for any instance of the structural coeffect calculus. These two strong properties are desirable for programming languages, but are often not satisfied (e. g. by languages with effects).

Two questions remain open for the upcoming chapter. Firstly, is there a way to present flat and structural coeffect calculi in a uniform way as a single system? Secondly, what is the best pathway towards a practical implementation of programming languages based on coeffect systems?

The main goal of this thesis is to provide a *unified* calculus for tracking context dependence. We have not achieved this goal yet. In Chapter 3, we identified two kinds of contextual properties that we further covered separately. Flat coeffects in Chapter 4 track whole-context properties, while structural coeffects, covered in Chapter 5, track per-variable properties. In this chapter, we unify the two notions.

Although the results presented in this thesis are mainly of a theoretical nature, we indeed believe that coeffects should be integrated in main-stream programming languages. This is addressed in the second part of the chapter, where we outline possible approach for a practical implementation of coeffect systems. Finally, we conclude the chapter with a brief discussion of an alternative approach to coeffect system, based on the meta-language style.

#### CHAPTER STRUCTURE AND CONTRIBUTIONS

- We start by unifying the two kinds of coeffects discussed so far. In Section 6.1, we introduce the *unified coeffect calculus*, which generalizes flat and structural systems and can be instantiated to track both flat and structural properties.
- We outline one possible approach for practical implementations of coeffects (Section 6.2). The technique is based on embedding comonadic computations via a lightweight syntactic extension and using rich type systems for capturing the structure of coeffect algebra.
- Finally, we discuss an alternative approach to defining coeffect systems based on the meta-language style, which highlights the relationship between our work and related work arising from modal logics such as CMTT [58] (Section 6.3).

The unified coeffect calculus presented in the first part of the chapter is a novel result that completes the theory developed in this thesis. The next two sections present important future and related work, respectively.

#### 6.1 THE UNIFIED COEFFECT CALCULUS

The type systems of the flat coeffect calculus (Figure 22) and the structural coeffect calculus (Figure 28) differ in a number of ways. Understanding the differences is the key to reconciling the two systems:

- The structural coeffect calculus contains explicit rules for context manipulation (weakening, contraction, exchange). In the flat coeffect calculus, these rules are not defined explicitly, but are admissible.
- In the structural coeffect calculus, the variable context is treated as a vector and is annotated with a vector of (scalar) coeffects. In the flat coeffect calculus, the variable context is a finite partial function (from names to types) and is annotated with a single (scalar) coeffect.

- In the flat coeffect calculus, we distinguish between *splitting* context requirements and *merging* of context requirements ( $\oplus$  and  $\wedge$ , respectively). In the structural coeffect calculus, the operations (which model appending and splitting vectors) are invertible and so the structural coeffect algebra uses the same tensor product  $\times$ .

In the unified calculus presented in this section, we address the three differences as follows. We include explicit rules for context manipulation in the calculus; in systems that arise from flat calculi, the structural rules do not change coeffects and can thus be applied freely. We generalize the structure of coeffect annotations using the notion of a “container” which can be specialized to obtain a single annotation or a vector of annotations. This could potentially be used to capture other structures such as trees in bunched typing [60]. Finally, we distinguish between splitting and merging of context requirements (using the notation  $\times$  and  $\bowtie$ , respectively). For structural coeffect calculi, the two operators coincide, but for flat coeffect calculi, they provide the needed flexibility.

#### 6.1.1 Shapes and containers

Our generalization of coeffect structure using a *coeffect container* is based on containers of Abbott et al. [3]. Interestingly, containers have later been linked to comonads by Ahman et al. [4]. Intuitively a container describes data types such as lists, trees or streams. A container is formed by shapes (e. g. lengths of lists)<sup>1</sup>. For every shape, we can obtain a set of positions in the container (e. g. offsets in a list of a specified length). More formally:

**Definition 18.** A container consists of a pair  $(S, P)$ , usually written as  $S \triangleleft P$ . Here,  $S$  is a set of shapes and  $P$  is a shape-indexed family  $P : S \rightarrow \text{Set of positions}$ .

Well-known examples of containers include lists, non-empty lists, (unbounded) streams, trees and the singleton data type (which contains exactly one element). Containers relevant to our work are lists and singleton data types:

- The container representing lists is given by  $S \triangleleft P$  where shapes are integers  $S = \text{Nat}$  (lengths of a list). The set of positions for a given length  $n$  is the set of indices  $P(n) = \{1 \dots n\}$ .
- The container representing the singleton data type is given by  $S \triangleleft P$  where shapes are given by a singleton set  $S = \{*\}$  and the set of positions for the shape  $*$  contains exactly one position. To follow the intuition based on offset or index, we write the single position as 0, that is  $P(*) = \{0\}$ .

In the unified coeffect calculus, the structure of coeffect annotations is defined by a container with additional operations (discussed later) that links it with the free-variable context  $\Gamma$ .

#### 6.1.2 Structure of coeffects

In the structural coeffect calculus, the coeffect annotation was formed by a vector of coeffect scalars. The annotations in the unified coeffect calculus are similar, but a *vector* is replaced with a more general *container*. The primitive

<sup>1</sup> Shapes can also be intuitively thought of as sizes, but this is not fully precise as some containers may have multiple shapes of the same length, e. g. differently ballanced trees.

coeffect annotations in the unified calculus are formed of *coeffect scalars*. The coeffect scalars are equipped with the same structure as in structural coeffect calculus (Definition 12). In this section, we refer to it as *unified coeffect scalar* (and we repeat the definition below). Then we define *unified coeffect containers* which determines how coeffect scalar values are attached to the free-variable context. Finally, we define the *unified coeffect algebra* which consists of shape-indexed coeffect scalar values.

As in the structural coeffect calculus, the contexts in the unified calculus are annotated with shape-indexed coeffects, written as  $\Gamma @ \mathbf{r} \vdash e : \tau$ ; functions take just a single input parameter and so are annotated with scalar coeffect values  $\sigma \xrightarrow{\mathbf{r}} \tau$ .

**COEFFECT SCALARS.** The following definition of the coeffect scalar structure repeats the Definition 12 from the previous chapter.

**Definition 19.** A *unified coeffect scalar*  $(\mathcal{C}, \otimes, \oplus, \text{use}, \text{ign}, \leq)$  is a set  $\mathcal{C}$  together with elements  $\text{use}, \text{ign} \in \mathcal{C}$ , relation  $\leq$  and binary operations  $\otimes, \oplus$  such that  $(\mathcal{C}, \otimes, \text{use})$  and  $(\mathcal{C}, \oplus, \text{ign})$  are monoids and  $(\mathcal{C}, \leq)$  is a pre-order. That is, for all  $r, s, t \in \mathcal{C}$ :

$$\begin{aligned} r \otimes (s \otimes t) &= (r \otimes s) \otimes t & \text{use} \otimes r &= r = r \otimes \text{use} & (\text{monoid}) \\ r \oplus (s \oplus t) &= (r \oplus s) \oplus t & \text{ign} \oplus r &= r = r \oplus \text{ign} & (\text{monoid}) \\ \text{if } r \leq s \text{ and } s \leq t & \text{ then } r \leq t & t \leq t & & (\text{pre-order}) \end{aligned}$$

The following distributivity axioms hold:

$$\begin{aligned} (r \oplus s) \otimes t &= (r \otimes t) \oplus (s \otimes t) \\ t \otimes (r \oplus s) &= (t \otimes r) \oplus (t \otimes s) \end{aligned}$$

In addition, we require the following two properties of  $\text{ign}, \oplus$  and  $\leq$ :

$$\begin{aligned} r \leq r' &\Rightarrow \forall s. (r \otimes s) \leq (r' \otimes s) \\ \text{ign} &\leq (\text{ign} \otimes r) \end{aligned}$$

As previously, the monoid  $(\mathcal{C}, \otimes, \text{use})$  models sequential composition; the laws guarantee an underlying category structure;  $\text{use}$  and  $\text{ign}$  represent an accessed and unused variable, respectively.

The  $\oplus$  operation models combining of context requirements arising from multiple parts of a program. The meaning depends on the coeffect container. The operation can either combine requirements of individual variables (structural coeffects) or requirements attached to the whole context of multiple sub-expressions (flat coeffects).

**COEFFECT CONTAINERS.** The coeffect container is a container that determines how are scalar coeffects attached to free-variable contexts. In addition to a container  $S \triangleleft P$  formed by shapes and shape-indexed positions, the coeffect container provides a mapping that returns the shape corresponding to a free-variable context.

The mapping between the shape of the variable context and the shape of the coeffect annotation is not necessarily bijective. For example, coeffect annotations in flat systems have just a single shape  $S = \{*\}$ .

In the coeffect judgment  $\Gamma @ \mathbf{r} \vdash e : \tau$ , the coeffect annotation  $\mathbf{r}$  is drawn from the set of coeffect scalars  $\mathcal{C}$  indexed by the shape corresponding to  $\Gamma$ . We write  $s = \text{len}(\Gamma)$  for the shape corresponding to  $\Gamma$ . The operation  $P(s)$  returns a set of positions and so we can write  $\mathbf{r} \in P(s) \rightarrow \mathcal{C}$  as a mapping from positions (defined by the shape) to scalar coeffects. We usually write this as

the exponent  $\mathbf{r} \in \mathbb{C}^{\mathbf{P}(\mathbf{s})}$ . The coeffect container is also equipped with an operation that appends shapes (when we concatenate variable contexts) and two special shapes in  $\mathbf{S}$  representing the empty context and the singleton context.

**Definition 20.** A *coeffect container* structure  $(\mathbf{S} \triangleleft \mathbf{P}, \diamond, \hat{0}, \hat{1}, \text{len}(-))$  comprises a container  $\mathbf{S} \triangleleft \mathbf{P}$  with a binary operation  $\diamond$  on  $\mathbf{S}$  for appending shapes, a mapping from free-variable contexts to shapes  $\text{len}(\Gamma) \in \mathbf{S}$ , and elements  $\hat{0}, \hat{1} \in \mathbf{S}$  such that  $(\mathbf{S}, \diamond, \hat{0})$  is a monoid.

The elements  $\hat{0}$  and  $\hat{1}$  represent the shapes of the empty and the singleton free-variable context, respectively. The  $\diamond$  operation corresponds to concatenation of free-variable contexts. Given  $\Gamma_1$  and  $\Gamma_2$  such that  $\mathbf{s}_1 = \text{len}(\Gamma_1)$ ,  $\mathbf{s}_2 = \text{len}(\Gamma_2)$ , we require that  $\mathbf{s}_1 \diamond \mathbf{s}_2 = \text{len}(\Gamma_1, \Gamma_2)$ .

As discussed earlier, we use two kinds of coeffect containers that describe the structure of vectors (for structural coeffects) and the shape of trivial singleton container (for flat coeffects):

**Example 19.** Structural coeffect container is defined as  $(\mathbf{S} \triangleleft \mathbf{P}, |-|, +, 0, 1)$  where  $\mathbf{S} = \mathbb{N}$  and  $\mathbf{P}(\mathbf{n}) = \{1 \dots \mathbf{n}\}$ . The shape mapping  $|\Gamma|$  returns the number of variables in  $\Gamma$ . Empty and singleton contexts are annotated with 0 and 1, respectively, and shapes of combined contexts are added so that  $|\Gamma_1, \Gamma_2| = |\Gamma_1| + |\Gamma_2|$ .

Therefore, a coeffect annotation is a vector  $\mathbf{r} \in \mathbb{C}^{\mathbf{P}(\mathbf{n})}$  and assigns a coeffect scalar  $\mathbf{r}(i) \in \mathbb{C}$  for each position (corresponding to a variable  $x_i$  in the context).

**Example 20.** Flat coeffect container is defined as  $(\mathbf{S} \triangleleft \mathbf{P}, |-|, \diamond, \star, \star)$ . The container is defined as a singleton data type  $\mathbf{S} = \{\star\}$  and  $\mathbf{P}(\star) = \{0\}$  with a constant function  $|\Gamma| = \star$  and a trivial operation  $\star \diamond \star = \star$ .

That is, there is a single shape  $\star$  with a single position and all free-variable contexts have the same singleton shape. Therefore, a coeffect annotation is drawn from  $\mathbb{C}^{\{\star\}}$  which is isomorphic to  $\mathbb{C}$  and so a coeffect scalar  $\mathbf{r} \in \mathbb{C}$  is associated with every free-variable context.

**Example 21.** Similarly to the previous example, we can also define a coeffect container with no positions, i.e.  $(\mathbf{S} \triangleleft \mathbf{P}, |-|, \diamond, \star, \star)$  where  $\mathbf{S} = \{\star\}$ ,  $\mathbf{P}(\star) = \emptyset$ ,  $|\Gamma| = \star$  and  $\star \diamond \star = \star$ . This reduces our system to the simply-typed  $\lambda$ -calculus with no context annotations, because  $\mathbf{P}(\star) = \emptyset$  and so coeffect annotations would be from the empty set  $\mathbb{C}^\emptyset$ .

**UNIFIED COEFFECT ALGEBRA.** The unified coeffect calculus annotates typing judgments with shape-indexed coeffect annotations. The *unified coeffect algebra* combines a coeffect scalar and a coeffect container to define shape-indexed coeffects and operations for manipulating these.

The definition here reconciles the third point discussed in Section 6.1 – the fact that flat coeffects use separate operations for splitting and merging ( $\oplus$  and  $\wedge$ ) while structural coeffects use the tensor  $\otimes$ . In the unified calculus, we use two operators that can, however, coincide.

**Definition 21.** Given a *unified coeffect scalar*  $(\mathbb{C}, \otimes, \oplus, \text{use}, \text{ign}, \leq)$  and a *coeffect container*  $(\mathbf{S} \triangleleft \mathbf{P}, \text{len}(-), \diamond, \hat{0}, \hat{1})$  a *unified coeffect algebra* extends the two structures with  $(\bowtie, \times, \perp)$  where  $\perp \in \mathbb{C}^{\mathbf{P}(\hat{0})}$  is a coeffect annotation for the empty context and  $\bowtie, \times$  are families of operations that combine coeffect annotations indexed by shapes. That is  $\forall \mathbf{n}, \mathbf{m} \in \mathbf{S}$ :

$$\bowtie_{\mathbf{m}, \mathbf{n}}, \times_{\mathbf{m}, \mathbf{n}} : \mathbb{C}^{\mathbf{P}(\mathbf{m})} \times \mathbb{C}^{\mathbf{P}(\mathbf{n})} \rightarrow \mathbb{C}^{\mathbf{P}(\mathbf{m} \diamond \mathbf{n})}$$



A coeffect algebra induces the following three additional operations:

$$\langle - \rangle : \mathcal{C} \rightarrow \mathcal{C}^{\mathbf{P}(\hat{\mathbf{i}})}$$

$$\langle \mathbf{x} \rangle = \lambda_.x$$

$$\otimes_{\mathbf{m}} : \mathcal{C} \times \mathcal{C}^{\mathbf{P}(\mathbf{m})} \rightarrow \mathcal{C}^{\mathbf{P}(\mathbf{m})}$$

$$\mathbf{r} \otimes \mathbf{s} = \lambda s. \mathbf{r} \otimes \mathbf{s}(s)$$

$$\text{len}(-) : \mathcal{C}^{\mathbf{P}(\mathbf{m})} \rightarrow \mathbf{m}$$

$$\text{len}(\mathbf{r}) = \mathbf{m}$$

The  $\langle - \rangle$  operation lifts a scalar coeffect to a shape-indexed coeffect that is indexed by the shape of a singleton context. The  $\otimes_{\mathbf{m}}$  operation is a left multiplication of a vector by a scalar. As we always use bold face for vectors and ordinary face for scalars (as well as a distinct colour), using the same symbol is not ambiguous. We also tend to omit the subscript  $\mathbf{m}$  and write  $\otimes$ .

Finally, we define  $\text{len}(-)$  as an operation that returns the shape of a given shape-indexed coeffect. The only purpose is to simplify notation, as we tend to avoid subscripts, but often need to specify that shapes of variable context and coeffect match, e. g.  $\text{len}(\mathbf{r}) = \text{len}(\Gamma)$ .

**SPLITTING AND MERGING COEFFECTS.** The operators  $\times$  and  $\bowtie$  combine shape-indexed coeffects associated with two contexts. For example, assume we have  $\Gamma_1$  and  $\Gamma_2$  with coeffects  $\mathbf{r} \in \mathcal{C}^{\mathbf{P}(\mathbf{m})}$  and  $\mathbf{s} \in \mathcal{C}^{\mathbf{P}(\mathbf{n})}$ . In the structural system, the context shapes  $\mathbf{m}, \mathbf{n}$  denote the number of variables in the two contexts. The combined context  $\Gamma_1, \Gamma_2$  has a shape  $\mathbf{m} \diamond \mathbf{n}$  and the combined coeffects  $\mathbf{r} \times \mathbf{s}, \mathbf{r} \bowtie \mathbf{s} \in \mathcal{C}^{\mathbf{P}(\mathbf{m} \diamond \mathbf{n})}$  are indexed by that shape.

For structural coeffect systems such as bounded reuse, both  $\times$  and  $\bowtie$  are just the tensor product  $\times$  of vectors. For flat coeffect systems, the operations can be defined independently, letting  $\bowtie = \wedge$  and  $\times = \oplus$ .

The difference between  $\bowtie$  and  $\times$  is clarified by the semantics (Section 6.1.6), where  $\mathbf{r} \bowtie \mathbf{s}$  is an annotation of the *codomain* of a morphism that merges the capabilities provided by two contexts (in the syntactic reading, splits the context requirements), while  $\mathbf{r} \times \mathbf{s}$  is an annotation of the *domain* of a morphism that splits the capabilities of a single context into two parts (in the syntactic reading, merges their context requirements). Syntactically, this means that we always use  $\bowtie$  in the rule *assumptions* and  $\times$  in *conclusions*.

### 6.1.3 Unified coeffect type system

The unified coeffect system in Figure 30 resembles the structural type system shown in Figure 28. Rather than explaining the rules one-by-one, we focus on how the unified system differs from the structural system.

The type system for the unified coeffect calculus is parameterized by a coeffect scalar  $(\mathcal{C}, \otimes, \oplus, \text{use}, \text{ign}, \leq)$  together with a coeffect algebra  $(\bowtie, \times, \perp)$  and the derived constructs  $\langle - \rangle$ ,  $\text{len}(-)$  and  $\otimes$ . As in the structural system, free-variable contexts  $\Gamma$  are treated as vectors modulo duplicate use of variables – the associativity is built-in. The order of variables matters, but can be changed using the exchange rule. The context annotations  $\mathbf{r}, \mathbf{s}, \mathbf{t}$  are shape-indexed coeffects (rather than simple vectors as before). As before, functions are annotated with coeffects scalars.

a.) Syntax-driven typing rules:

$$\begin{aligned}
(var) \quad & \frac{}{x : \tau @ \langle \text{use} \rangle \vdash x : \tau} \\
(const) \quad & \frac{c : \tau \in \Delta}{() @ \perp \vdash c : \tau} \\
(app) \quad & \frac{\Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \quad \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma_1, \Gamma_2 @ \mathbf{r} \times (\mathbf{t} \otimes \mathbf{s}) \vdash e_1 e_2 : \tau_2} \\
(abs) \quad & \frac{\Gamma, x : \tau_1 @ \mathbf{r} \times \langle \mathbf{s} \rangle \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{s} \tau_2} \\
(let) \quad & \frac{\Gamma_1 @ \mathbf{r} \vdash e_1 : \tau_1 \quad \Gamma_2, x : \tau_1 @ \mathbf{s} \times \langle \mathbf{t} \rangle \vdash e_2 : \tau_2}{\Gamma_1, \Gamma_2 @ (\mathbf{t} \otimes \mathbf{r}) \times \mathbf{s} \vdash \text{let } x = e_1 \text{ in } e_2 : \tau_2}
\end{aligned}$$

b.) Structural rules for context manipulation:

$$\begin{aligned}
(sub) \quad & \frac{\Gamma_1, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}' \rangle \times \mathbf{q} \vdash e : \tau}{\Gamma_1, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} \rangle \times \mathbf{q} \vdash e : \tau} \quad (s' \leq s) \\
(weak) \quad & \frac{\Gamma @ \mathbf{r} \vdash e : \tau}{\Gamma, x : \tau_1 @ \mathbf{r} \times \langle \text{ign} \rangle \vdash e : \tau} \\
(exch) \quad & \frac{\Gamma_1, x : \tau_1, y : \tau_2, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} \rangle \times \langle \mathbf{t} \rangle \times \mathbf{q} \vdash e : \tau}{\Gamma_1, y : \tau_2, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{t} \rangle \times \langle \mathbf{s} \rangle \times \mathbf{q} \vdash e : \tau} \quad \begin{array}{l} \text{len}(\Gamma_1) = \text{len}(\mathbf{r}) \\ \text{len}(\Gamma_2) = \text{len}(\mathbf{s}) \end{array} \\
(contr) \quad & \frac{\Gamma_1, y : \tau_1, z : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} \rangle \times \langle \mathbf{t} \rangle \times \mathbf{q} \vdash e : \tau}{\Gamma_1, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} \oplus \mathbf{t} \rangle \times \mathbf{q} \vdash e[z, y \leftarrow x] : \tau} \quad \begin{array}{l} \text{len}(\Gamma_1) = \text{len}(\mathbf{r}) \\ \text{len}(\Gamma_2) = \text{len}(\mathbf{s}) \end{array}
\end{aligned}$$

Figure 30: Type system for the unified coeffect calculus

**SYNTAX-DRIVEN RULES.** The  $(var)$  rule is syntactically the same as in the structural system, but it should be read differently. The  $\langle - \rangle$  operation does not create a *vector*, but a *shape-indexed coeffect* that returns *use* for all positions in the singleton shape  $\hat{1}$ . The  $(const)$  rule annotates empty context with a special annotation of the shape  $\hat{0}$ .

In a structural system, the two annotations correspond to a singleton and empty vector, respectively. However, for a singleton shape with one position, the annotations are equivalent to annotating variables with a scalar *use* and constants with a scalar *ign*.

In the  $(app)$ ,  $(abs)$  and  $(let)$  rules, the only change from the structural system is that the vector concatenation  $\times$  is now replaced with context splitting/merging of the unified coeffect algebra. As already mentioned, we use splitting of context requirements  $\times$  in rule *assumptions* and merging of context requirements  $\times$  in rule *conclusions*.

Note that we use the terms *merging* and *splitting* in the syntactic (top-down) sense. As discussed in Chapter 4, Section 4.2.2, we can also read the rules in semantic (bottom-up) sense, in which case assumptions merge available contextual information and conclusions split available contextual information.

**STRUCTURAL RULES.** The merging/splitting operations in the structural rules are changed in the same way as in the rules above. It is also worth noting that structural coeffect system used vectors of multiple elements. For example,  $\langle s, t \rangle$  and  $\langle t, s \rangle$  in the (*exch*) rule denoted two-element vectors. In the unified system, this is replaced with merging/splitting of two lifted scalars:  $\langle s \rangle \bowtie \langle t \rangle$  and  $\langle t \rangle \bowtie \langle s \rangle$ .

In structural systems, the two notations mean the same thing – we are simply concatenating or splitting two singleton vectors. However, this generalization allows us to capture flat coeffect systems as well. The lifting operation in flat systems simply returns the lifted scalar and operators  $\bowtie$  and  $\bowtie$  correspond to operations on coeffect scalars. As discussed in Section 6.1.5, thanks to the properties of coeffect scalars, contraction, weakening and exchange that do not affect the coeffect annotation are admissible for all flat systems embedded in the unified coeffect calculus.

#### 6.1.4 Structural coeffects

The unified coeffect system uses a general notion of context shape, but it has been designed with structural and flat systems in mind. In this and the next section, we show how it captures the two coeffect systems from Chapter 4 and Chapter 5.

The unified calculus closely resembles the structural system and so using it to model structural systems is easy – given a coeffect scalar, we use the coeffect container that describes a *vector* of annotations (Example 19) and define a coeffect algebra formed by a vector (free monoid) of scalars.

**Definition 22.** Given a coeffect scalar  $(\mathcal{C}, \otimes, \oplus, \text{use}, \text{ign}, \leq)$  a structural coeffect system is defined by:

- A coeffect container  $(S \triangleleft P, \vdash, +, 0, 1)$  where  $S = \mathbb{N}$  and  $P(n) = \{1 \dots n\}$  and  $|x_1 : \tau_1, \dots, x_n : \tau_n| = n$ .
- A coeffect algebra  $(\times, \times, \epsilon)$  where  $\times$  and  $\epsilon$  are shape-indexed versions of the binary operation and the unit of a free monoid over  $\mathcal{C}$ . That is  $\epsilon : \mathcal{C}^{P(0)}$  is an empty vector and  $\times : \mathcal{C}^{P(n)} \times \mathcal{C}^{P(m)} \rightarrow \mathcal{C}^{P(n+m)}$  appends vectors.

The definition is valid since the shape operations form a monoid  $(\mathbb{N}, +, 0)$  and  $\text{len}(-)$  (calculating the length of a list) is a monoid homomorphism from the free monoid to the monoid of shapes.

**PROPERTIES.** An important property of the unified system is that, when used in a structural way as discussed above, it gives calculi with the same properties as the structural system described in Chapter 5. This can be easily seen by comparing the Figure 30 with the Figure 28 and using the free monoid interpretation of the unified coeffect algebra.

**Remark 17.** The system described in Definition 22 is equivalent to the structural coeffect system described in Figure 28. That is, a typing derivation using a structural coeffect embedded in the unified system is valid if and only if the corresponding derivation is valid in the structural system.

Using the above definition, our unified coeffect system can capture per-variable coeffect properties discussed in Section 3.3. This includes the system for bounded reuse (which is only used in the structural form) and precise tracking of per-variable dataflow and liveness.

## 6.1.5 Flat coeffects

The same unified coeffect system can be used to capture systems that track whole-context (flat) coeffects such as implicit parameters. This is achieved using a singleton-shaped container for coeffect annotations. The resulting system has explicit structural rules (and is syntactically different from the standard flat coeffect system), but we show that they are equivalent.

Flat coeffect systems are characterised by a singleton set of shapes (Example 20). In this setting, the context annotations  $\mathcal{C}^{\mathbf{P}(\star)} = \{0\} \rightarrow \mathcal{C}$  are equivalent to coeffect scalars  $\mathcal{C}$ . In addition to the coeffect scalar structure, we need to define  $\times$  and  $\bowtie$ . Our examples of flat coeffects use  $\oplus$  (merging of scalar coeffects) for  $\times$  (merging of shaped coeffect annotations). The  $\bowtie$  operation (corresponding to  $\wedge$  in flat coeffect calculus) needs to be provided explicitly.

**Definition 23.** Given a coeffect scalar  $(\mathcal{C}, \otimes, \oplus, \text{use}, \text{ign}, \leq)$  and a binary operation  $\wedge : \mathcal{C} \times \mathcal{C} \rightarrow \mathcal{C}$  such that  $(\mathbf{r} \wedge \mathbf{s}) \leq (\mathbf{r} \oplus \mathbf{s})$ , the unified coeffect algebra modelling a flat coeffect systems consists of:

- A flat coeffect container  $(S \triangleleft P, | - |, \diamond, \star, \star)$  as defined in Example 20.
- A flat coeffect algebra  $(\wedge, \oplus, \text{ign})$ , i. e.  $\times = \oplus$  and  $\perp = \text{ign}$  with an additional binary operation  $\bowtie = \wedge$ .

Intuitively, the requirement  $(\mathbf{r} \wedge \mathbf{s}) \leq (\mathbf{r} \oplus \mathbf{s})$ , which could be also written as  $(\mathbf{r} \times \mathbf{s}) \leq (\mathbf{r} \times \mathbf{s})$ , denotes that splitting context requirements and then re-combining them preserves all the requirements from the assumptions. The system may be imprecise and conclusions can overapproximate assumptions, but it cannot lose requirements. This is fundamental for showing that exchange and contraction are admissible in the unified system.

**PROPERTIES.** To show that the typing of flat properties in the unified system is equivalent to the typing in the flat system, we show that a valid typing judgement in the first system is a valid typing judgement in the second system and vice versa.

In one direction, we show that the unified system (capturing flat properties) permits weakening, contraction and exchange rules that do not change the coeffect annotations. This guarantees that a valid judgement in flat system is also valid in the unified system.

**Lemma 18.** A unified coeffect calculus capturing flat properties admits weakening that does not change the coeffect annotation.

*Proof.* The rule is admissible using the following derivation:

$$\frac{\frac{\Gamma @ \mathbf{r} \vdash e : \tau}{\Gamma, x : \tau_1 @ \mathbf{r} \times \langle \text{ign} \rangle \vdash e : \tau}}{\Gamma, x : \tau_1 @ \mathbf{r} \oplus \text{ign} \vdash e : \tau}}{\Gamma, x : \tau_1 @ \mathbf{r} \vdash e : \tau}$$

We write  $\mathbf{r}$  rather than  $\mathbf{r}$ , because we are tracking flat properties. The first step is an application of the (*weak*) rule. Next, we use the fact that  $\times = \oplus$  and  $\langle \text{ign} \rangle = \text{ign}$ , which is the unit of the monoid  $(\mathcal{C}, \oplus, \text{ign})$ .  $\square$

**Lemma 19.** A unified coeffect calculus capturing flat properties admits exchange that does not change the coeffect annotation.

*Proof.* We use the idempotence of  $\wedge$  and  $\oplus$  together with the (*exch*) rule:

$$\frac{\frac{\frac{\Gamma_1, x:\tau_1, y:\tau_2, \Gamma_2 @ \mathbf{r} \vdash e : \tau}{\Gamma_1, x:\tau_1, y:\tau_2, \Gamma_2 @ \mathbf{r} \wedge \mathbf{r} \wedge \mathbf{r} \vdash e : \tau}}{\Gamma_1, x:\tau_1, y:\tau_2, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{r} \rangle \times \langle \mathbf{r} \rangle \times \mathbf{r} \vdash e : \tau}}{\frac{\Gamma_1, y:\tau_2, x:\tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{r} \rangle \times \langle \mathbf{r} \rangle \times \mathbf{r} \vdash e : \tau}{\Gamma_1, y:\tau_2, x:\tau_1, \Gamma_2 @ \mathbf{r} \oplus \mathbf{r} \oplus \mathbf{r} \oplus \mathbf{r} \vdash e : \tau}}{\Gamma_1, y:\tau_2, x:\tau_1, \Gamma_2 @ \mathbf{r} \vdash e : \tau}$$

Using idempotence, we first duplicate the annotation  $\mathbf{r}$  to get a coeffect in the form  $\mathbf{r} \times \langle \mathbf{r} \rangle \times \langle \mathbf{r} \rangle \times \mathbf{r}$  as required by the assumption of the (*exch*) rule. Note that  $\langle - \rangle$  can be added freely as  $\langle \mathbf{r} \rangle$  is equivalent to  $\mathbf{r}$ . After applying (*exch*), we use idempotence of  $\times$ .  $\square$

**Lemma 20.** *A unified coeffect calculus capturing flat properties admits contraction that does not change the coeffect annotation.*

*Proof.* Similarly to exchange, the proof uses idempotence of  $\wedge$  and  $\oplus$ :

$$\frac{\frac{\frac{\Gamma_1, y:\tau_1, z:\tau_1, \Gamma_2 @ \mathbf{r} \wedge \mathbf{r} \wedge \mathbf{r} \wedge \mathbf{r} \vdash e : \tau}{\Gamma_1, y:\tau_1, z:\tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{r} \rangle \times \langle \mathbf{r} \rangle \times \mathbf{r} \vdash e : \tau}}{\Gamma_1, x:\tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{r} \oplus \mathbf{r} \rangle \times \mathbf{r} \vdash e[z, y \leftarrow x] : \tau}}{\Gamma_1, x:\tau_1, \Gamma_2 @ \mathbf{r} \oplus \mathbf{r} \oplus \mathbf{r} \oplus \mathbf{r} \vdash e[z, y \leftarrow x] : \tau} \quad \square$$

In the last two cases, we need to turn the coeffect into a form that is required by the exchange and contraction rules. Aside from idempotence, we could use the unit property and obtain e. g.  $\text{ign} \wedge \mathbf{r} \wedge \text{ign} \wedge \text{ign}$ . However, this approach does not work because  $\oplus$  and  $\wedge$  may have different unit elements (in fact, we do not even require the existence of a unit for  $\wedge$ ).

In the other direction, we need to show that any valid judgement in the unified system (tracking flat properties) is also valid in the flat system. The syntax-directed rules are the same in both systems, but we need to show that any use of (explicit) weakening, contraction and exchange can be derived in the flat system.

**Lemma 21.** *Weakening, as defined in a unified coeffect calculus capturing flat properties, is admissible in the flat coeffect calculus.*

*Proof.* Similar to the proof in Lemma 18. The property follows from the fact that  $\perp = \langle \text{ign} \rangle$  is the unit of  $\oplus$ .  $\square$

**Lemma 22.** *Contraction, as defined in a unified coeffect calculus capturing flat properties, is admissible in the flat coeffect calculus.*

*Proof.* The application of (*contr*) rule has the following form:

$$\frac{\frac{\frac{\Gamma_1, y:\tau_1, z:\tau_1, \Gamma_2 @ \mathbf{r} \wedge \mathbf{s} \wedge \mathbf{t} \wedge \mathbf{q} \vdash e : \tau}{\Gamma_1, y:\tau_1, z:\tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} \rangle \times \langle \mathbf{t} \rangle \times \mathbf{q} \vdash e : \tau}}{\Gamma_1, x:\tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} \oplus \mathbf{t} \rangle \times \mathbf{q} \vdash e[z, y \leftarrow x] : \tau}}{\Gamma_1, x:\tau_1, \Gamma_2 @ \mathbf{r} \oplus \mathbf{s} \oplus \mathbf{t} \oplus \mathbf{q} \vdash e[z, y \leftarrow x] : \tau}$$

From Definition 23, we know that  $(\mathbf{r} \wedge \mathbf{s} \wedge \mathbf{t} \wedge \mathbf{q}) \leq (\mathbf{r} \oplus \mathbf{s} \oplus \mathbf{t} \oplus \mathbf{q})$ . Thus, the judgement can be derived using the (*sub*) rule of flat coeffect calculus.  $\square$

**Lemma 23.** *Exchange, as defined in a unified coeffect calculus capturing flat properties, is admissible in the flat coeffect calculus.*

*Proof.* The application of (*exch*) rule has the following form:

$$\frac{\frac{\frac{\Gamma_1, x:\tau_1, y:\tau_2, \Gamma_2 @ r \wedge s \wedge t \wedge q \vdash e : \tau}{\Gamma_1, x:\tau_1, y:\tau_2, \Gamma_2 @ r \times \langle s \rangle \times \langle t \rangle \times q \vdash e : \tau}}{\Gamma_1, y:\tau_2, x:\tau_1, \Gamma_2 @ r \times \langle t \rangle \times \langle s \rangle \times q \vdash e : \tau}}{\Gamma_1, y:\tau_2, x:\tau_1, \Gamma_2 @ r \oplus t \oplus s \oplus q \vdash e : \tau}$$

From idempotence of  $\wedge$ , we get that  $r \wedge s \wedge t \wedge q = r \wedge t \wedge s \wedge q$ . Thus the judgement can be derived using (*sub*) as in contraction.  $\square$

A consequence of the equivalence discussed above is that the unified coeffect system can capture all properties that can be captured by the flat coeffect system – including implicit parameters, rebindable resources, Haskell type classes (discussed by Orchard [62]), data-flow and variable liveness.

### 6.1.6 Semantics of unified calculus

The semantics of the unified calculus can be obtained by generalizing the semantics of the structural calculus in the same way as we generalized the type system. However, it is worth exploring the type signatures of the operations of the underlying (comonad-based) structure.

The semantics of the unified coeffect calculus follows the same pattern as the semantics of the structural calculus (Section 5.3) with a number of differences. The index  $r$  of the object mapping  $C^r$  is an element of the structure  $\mathcal{C}^{P(n)}$  for a specified container  $S \triangleleft P$  and a shape  $n$  determined as  $n = \text{len}(\Gamma)$ . To match the type system discussed in Section 6.1.3, we need to modify the domains and codomains of the merge, split and dup operations as follows.

$$\begin{aligned} \text{merge}_{r,s} &: C^r \alpha \times C^s \beta \rightarrow C^{r \times s}(\alpha \hat{\times} \beta) \\ \text{split}_{r,s} &: C^{r \times s}(\alpha \hat{\times} \beta) \rightarrow C^r \alpha \times C^s \beta \\ \text{dup}_{r,s} &: C^{(r \oplus s)} \alpha \rightarrow C^{(r)} \times C^{(s)}(\alpha \hat{\times} \alpha) \end{aligned}$$

The key difference is that the vector concatenation  $\times$  is replaced with  $\times$  in the *domain* of the operations (in split) and with  $\times$  in the *codomain* (in merge, but also in dup which models contraction). This means that the splitting and merging the context can change the coeffect annotation:

$$\text{merge}_{r,s} \circ \text{split}_{r,s} : C^{r \times s}(\alpha \hat{\times} \beta) \rightarrow C^{r \times s}(\alpha \hat{\times} \beta)$$

For flat coeffect calculi where  $(r \times s) \leq (r \times s)$ , splitting and then merging a context may lose some of the available contextual capabilities (and increase context requirements). This illuminates why our proofs of Lemma 22 and Lemma 23 relied on sub-coeffecting in the flat system.

**RELATED AND FUTURE WORK.** The semantics presented here serves mainly to inform the design of the type system. For this reason, we keep the semantics simple and close to the concrete notions used in the type system such as flat, structural and unified coeffect algebras. In particular, the indexed comonad structure consists of object mapping  $C^r$  that is indexed by a corresponding algebra:

- In the flat coeffect calculus,  $C^r$  is indexed by scalar coeffects  $r \in \mathcal{C}$
- In the structural coeffect calculus, the object mapping  $C^r$  is indexed by vectors of scalars  $r \in \bigcup_{m \in \mathbb{N}} \mathcal{C}^m$
- In unified coeffect calculus with a coeffect container  $S \triangleleft P$ , the mapping  $C^r$  is indexed by shape-indexed scalars  $r \in \bigcup_{s \in S} \mathcal{C}^{P(m)}$

In this treatment, the indices are formed by objects that are modelled outside of category theory. A fully categorical semantics can be found in a joint paper with Orchard [71]. There, the index (coeffect scalars and shape-indexed scalars) are modelled as categories too. The definition is written using the following notation:

- $[\mathbb{C}, \mathbb{D}]$  is a category of functors between categories  $\mathbb{C}$  and  $\mathbb{D}$
- $B^A$  is an exponential object  $A \Rightarrow B$  in a Cartesian-closed category

In the categorical semantics of coeffect calculi, coeffect scalars are modelled as a category of scalars  $\mathbb{I}$  such that  $\text{obj}(\mathbb{I}) = \mathcal{C}$ . The structure of the context (indexed comonad) over the input then becomes a functor indexed by the free-variable context shape  $\text{len}(\Gamma)$  and coeffect annotations  $\mathbf{r}$ , i. e.  $C_{\mathbf{r}}^{\text{len}(\Gamma)}$ :

$$\llbracket \Gamma @ \mathbf{R} \vdash e : \tau \rrbracket : C_{\mathbf{r}}^{\text{len}(\Gamma)} \llbracket \Gamma \rrbracket \rightarrow \llbracket \tau \rrbracket$$

The structure  $C$  can be thought of as a dependent product of functors  $C^{\mathbf{n}}$  over possible shapes  $\mathbf{n} \in S$  where:

$$C^{\mathbf{n}} : \mathbb{I}^{P(\mathbf{n})} \rightarrow [C^{P(\mathbf{n})}, C]$$

For a fixed context shape  $\mathbf{n}$  the functor  $C^{\mathbf{n}} : \mathbb{I}^{P(\mathbf{n})} \rightarrow [C^{P(\mathbf{n})}, C]$  maps a coeffect indexed by positions  $P(\mathbf{n})$  to a functor from a context  $C^{\mathbf{n}}$  to an object in  $C$ . That is, given a coeffect annotation (matching the shape of the context), we get a functor  $\in [C^{\mathbf{n}}, C]$ . From a programming perspective, this functor defines a data structure that models the additional context provided to the program. The shape of this data structure depends on the coeffect annotation  $\mathbb{I}^{\mathbf{n}}$ . For example, in bounded reuse, the annotation defines the number of values needed for each variable and the functor will be formed by lists of length matching the required number.

In the model above, shapes and positions are still treated as ordinary sets. Indeed, in the joint paper [71], we treat shapes as just sets of sets. In this chapter, we instead define *coeffect container* as an extension of containers, which have a well-defined categorical structure [2]. A future work is adapting the semantics of the unified coeffect calculus so that it is based on the semantics of containers. However, note that context is *not* a (simple) container, because containers are used as indices (coeffect annotations) of the object that wrap the free-variable context.

## 6.2 TOWARDS PRACTICAL COEFFECTS

As discussed earlier, the main focus of this thesis is the development of the much needed *theory of context-aware programming languages* and so discussing the details of a practical implementation of coeffect tracking is beyond the scope of the thesis. However, this section briefly outlines one possible pathway towards this goal.

Many of the examples of contextual computation that we discussed earlier have been implemented as a single-purpose programming language feature (e. g. implicit parameters [51] or distributed computations [56, 21]). However, the main contribution of this thesis is that it captures *multiple* different notions of context-aware computations using just a *single* common abstraction. For this reason, we advocate that future practical implementations of coeffects should not be single-purpose language features, but rather reusable abstractions that can be instantiated with a concrete *coeffect algebra* specified by the user.



In order to do this, programming languages need to provide two features; one that allows embedding of context-aware computations themselves in programs akin to the “do” notation in Haskell (Section 6.2.1) and one that allows tracking of the contextual information in the type system (Section 6.2.2). To make constraint resolution in type inference easier, we also discuss how the coeffect algebra can be simplified (Section ??)

### 6.2.1 Embedding contextual computations

The embedding of *contextual* computations into programming languages can follow the successful model of *effectful* computations. In purely functional programming languages such as Haskell, effectful computations are embedded by *implementing* their model within the language and inserting the necessary (monadic) plumbing. This is made easier by the “do” notation or monad comprehensions [43, 34], which insert the monadic operations automatically.

The recently proposed “codo” notation [65] provides similar automation for context-aware computations based on comonads. The notation follows the semantics of our flat coeffect calculus (Chapter 4). Extending the “codo” notation to support calculi based on the structural coeffect calculus (Chapter 5) is an interesting future work – this would require explicitly manipulating individual context variables and application of structural rules, which is not needed in flat coeffects.

In ML-like languages, effects (and many coeffects) are built-in into the language semantics, but they can still benefit from a special notation for explicitly marking effectful (coeffectful) blocks of code. In F#, this is done using *computation expressions* [72] that differ from the “do” notation in two ways. First, they support wider range of constructs, making it possible to wrap existing F# code within a block without other changes. Second, they support other abstractions including monads, applicative functors and monad transformers. It would be interesting to see if computation expressions can be extended to support programming with computations based on flat/structural indexed comonads.

More lightweight syntax for effectful computation can be obtained using techniques that automatically insert the necessary monadic plumbing (without special syntax). This has been done for effectful computations [81] and a similar approach would be worth exploring for coeffects.

### 6.2.2 Coeffect annotations as types

The other aspect of practical implementation of coeffects is that of tracking the context requirements (coeffect annotations) in the type system. To achieve this (without resorting to a single-purpose language feature) the type system needs to be able to capture various kinds of coeffect algebras. The structures required in this thesis include sets (with union or intersection), natural numbers (with addition, maximum, minimum and multiplication), two-point lattice (for liveness) and free monoids (vectors of annotations).

The work [66] on embedding effect systems in Haskell demonstrated that the recent additions to the Haskell type system provide enough power to implement structures such as sets at the type level. Using these to embed coeffect systems in Haskell is one fruitful future direction for applied coeffects.



In dependently-typed programming languages such as Agda or Idris [12, 13], the embedding of coeffects can be implemented more directly (as terms implementing sets or lattices can be lifted to the type level). However, we believe that coeffect tracking does not require full dependent types and can be made accessible in more mainstream languages. Dependent ML [105] provides an interesting example of a language with some dependent typing (e.g. computations with integers) which is still close to its non-dependently-typed predecessor ML.

Another approach for embedding computations into the type system has been pioneered by F# *type providers* [83]. Technically, type providers are compiler extensions that generate types based on external data sources or other information in order to provide easy access to data or services. A similar approach could be used for embedding *algebras* such as coeffect algebras into the type system. An *algebra provider* would then be a library that specifies the objects of the algebra, its equational laws and generalization rules for type inference. This could provide an easy-to-use way of embedding coeffect tracking in pragmatic languages such as F#. It is worth noting that the mechanism could also subsume F# units of measure [47]; these could be alternatively provided via one such *algebra provider*.

### 6.2.3 Alternative formulation using coeffect lattice

Compared to typical effect systems based on sets, coeffect systems have richer structure consisting of four operations ( $\otimes, \wedge, \oplus$  and a relation  $\leq$ ). This allows capturing interesting properties, but it makes working with coeffect annotations difficult – it complicates tasks such as type generalization and constraint solving in type inference. Here, we look at one possible simplification of the structure.

- As mentioned in Section 4.2.3, when the monoid  $(\mathcal{C}, \oplus, \text{ign})$  is a semi-lattice, the  $\leq$  relation can often be expressed in terms of the  $\oplus$  operation ( $r \leq s \iff r \oplus s = s$ ). This does not give a definition equivalent to our, but it is a simplification that is consistent with all examples in this thesis.
- Another approach that works for most (but not all) systems we discussed is to require the structure  $(\mathcal{C}, \oplus, \wedge)$  to form a lattice. This is consistent with all our structural coeffect systems as well as flat systems for liveness and data-flow, but not with the system for tracking of implicit parameters (or sets of rebindable resources).

The system for tracking implicit parameters does not fit the lattice-based model, because it uses the same operation  $\cup$  for both  $\oplus$  and  $\wedge$ .

This also means that the system for implicit parameters cannot be used with a system where lambda abstraction duplicates the context requirements (Section 4.5.2).

**LATTICE-BASED COEFFECTS.** Putting implicit parameters aside, there are many other coeffect systems for which we could use the following definition of coeffect scalars (for simplicity, we only consider flat systems):

**Definition 24.** A *flat coeffect lattice*  $(\mathcal{C}, \otimes, \leq, \text{use}, \text{ign})$  is a set  $\mathcal{C}$  together with elements  $\text{use}, \text{ign} \in \mathcal{C}$  and an ordering  $\leq$  such that  $(\mathcal{C}, \otimes, \text{use})$  is a monoid and  $(\mathcal{C}, \leq)$  is a lattice with the least element  $\text{ign}$ .

a.) Lattice-based flat systems (selected rules):

$$\begin{aligned}
 (abs) \quad & \frac{\Gamma, x : \tau_1 @ \mathbf{t} \vdash e : \tau_2}{\Gamma @ \mathbf{r} \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{s}} \tau_2} \quad (\mathbf{t} \leq \mathbf{r}, \mathbf{t} \leq \mathbf{s}) \\
 (app) \quad & \frac{\Gamma @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{\mathbf{t}} \tau_2 \quad \Gamma @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma @ \mathbf{u} \vdash e_1 e_2 : \tau_2} \quad (\mathbf{r} \leq \mathbf{u}, (\mathbf{s} \otimes \mathbf{t}) \leq \mathbf{u})
 \end{aligned}$$

b.) Lattice-based structural systems (selected rules):

$$(contr) \quad \frac{\Gamma_1, y : \tau_1, z : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}, \mathbf{t} \rangle \times \mathbf{q} \vdash e : \tau}{\Gamma_1, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{u} \rangle \times \mathbf{q} \vdash e[z, y \leftarrow x] : \tau} \quad (\mathbf{s} \leq \mathbf{u}, \mathbf{t} \leq \mathbf{u})$$

Figure 31: Flat and structural coeffects using lattice-based formulation

Assuming  $\oplus$  and  $\wedge$  are the least upper bound and greatest lower bound of the lattice, we require the following distributivity axioms:

$$\begin{aligned}
 (\mathbf{r} \oplus \mathbf{s}) \otimes \mathbf{t} &= (\mathbf{r} \otimes \mathbf{t}) \oplus (\mathbf{s} \otimes \mathbf{t}) \\
 \mathbf{t} \otimes (\mathbf{r} \oplus \mathbf{s}) &= (\mathbf{t} \otimes \mathbf{r}) \oplus (\mathbf{t} \otimes \mathbf{s})
 \end{aligned}$$

The definition provides all the operations of a flat coeffect algebra, but it defines a more rigid structure. In particular,  $\wedge$  and  $\oplus$  are both defined in terms of  $\leq$ . This simplifies the system as the operations of the underlying indexed comonad structure can be expressed using just  $\otimes$  and  $\leq$ :

$$\begin{aligned}
 \text{counit}_{\text{use}} &: C^{\text{use}} \alpha \rightarrow \alpha \\
 \text{cobind}_{\mathbf{r}, \mathbf{s}} &: (C^{\mathbf{r}} \alpha \rightarrow \beta) \rightarrow (C^{\mathbf{r} \otimes \mathbf{s}} \alpha \rightarrow C^{\mathbf{s}} \beta) \\
 \text{merge}_{\mathbf{r}, \mathbf{s}, \mathbf{t}} &: C^{\mathbf{r}} \alpha \times C^{\mathbf{s}} \beta \rightarrow C^{\mathbf{t}}(\alpha \times \beta) \quad (\mathbf{t} \leq \mathbf{r}, \mathbf{t} \leq \mathbf{s}) \\
 \text{split}_{\mathbf{r}, \mathbf{s}, \mathbf{t}} &: C^{\mathbf{t}}(\alpha \times \beta) \rightarrow C^{\mathbf{r}} \alpha \times C^{\mathbf{s}} \beta \quad (\mathbf{r} \leq \mathbf{t}, \mathbf{s} \leq \mathbf{t}) \\
 \text{lift}_{\mathbf{r}, \mathbf{s}} &: C^{\mathbf{r}} \alpha \rightarrow C^{\mathbf{s}} \alpha \quad (\mathbf{s} \leq \mathbf{r})
 \end{aligned}$$

Note that we still need *two distinct* operators in the system. Analysis of our examples shows that the operator used for sequential composition  $\otimes$  is not necessarily related in any way to the ordering on the lattice provided by  $\leq$ .

However, the operations merge, split and lift can now all use ordering. For example, the merge operation previously returned a result with greatest lower bound  $\mathbf{r} \wedge \mathbf{s}$ . Now, we require any element that is smaller than  $\mathbf{r}$  and  $\mathbf{s}$ . This adds an implicit sub-coeffecting to the operation – for any  $\mathbf{t}$  such that  $\mathbf{t} \leq \mathbf{r}$  and  $\mathbf{t} \leq \mathbf{s}$ , it is also the case that  $\mathbf{t} \leq (\mathbf{r} \wedge \mathbf{s})$ . This means that we can compose the original merge operation which returns greatest lower bound with lift to get the (less precise) operation required here.

**LATTICE-BASED TYPE SYSTEM.** Using lattice-based coeffect structure leads to the typing rules shown in Figure 31. In the flat coeffect system, this affects both application and abstraction – in abstraction, the context-requirements of the body should be smaller than the context available on the call site and declaration site. In application, we require more than what is required by the sub-expressions.

Structural coeffect systems only require least upper bound  $\oplus$  in the *(contr)* rule. This can be rewritten similarly to the flat *(app)* rule – when contracting two variables, the context requirements of the resulting variable are greater than the requirements associated with each of the original variables.

In summary, the lattice-based coeffect system has some advantages over the design used in this thesis. In particular, it uses better-known structures and is simpler. However, we do not use it as it rules out one of the important motivating examples (implicit parameters) and, even with the simplification, solving constraints with  $\otimes$  and  $\leq$  is still an open question.

### 6.3 COEFFECT META-LANGUAGE

In Section 2.2.1, we discussed two ways of using monads in programming language semantics as introduced by Moggi [55]. The first approach is to use monads in the *semantics* of an effectful language. The second approach is to extend the language with (additional) monadic constructs that can then be used for writing effectful monads explicitly.

In this thesis, we focused on the first approach. In both flat and structural coeffect calculi, the term language is that of a simply-typed  $\lambda$ -calculus, and we used (flat or structural) indexed comonads to give the semantics for the language and to derive type system for it.

In this section, we briefly discuss the other technique. That is, we embed indexed comonads into a  $\lambda$ -calculus as additional constructs. To do that, we introduce the type constructor  $C^r\tau$  which represents a value  $\tau$  wrapped in additional context (semantically, this corresponds to an indexed comonad) and we add language constructs that correspond to the operations of indexed comonads.

This section provides a brief sketch of the coeffect meta-language in order to highlight the relationship between coeffects and closely related work on contextual modal type theory (CMTT) [58]. Developing the system further is an interesting future research direction.

#### 6.3.1 *Coeffects and contextual modal type theory*

As discussed in Section 2.2.3, context-aware computations are related to modal logics – comonads have been used to model the  $\Box$  modality and as a basis for meta-languages that include  $\Box$  as a type constructor [11, 73, 11, 58]. Nanevski et al. [58] extend an S4 term language to a contextual modal type theory (CMTT). From the perspective of this thesis, CMTT can be viewed as a *meta-language* version of our coeffect calculus.

**CONTEXT IN CMTT AND COEFFECTS.** Aside from the fact that coeffect calculi use comonads for *semantics* and CMTT embeds comonads (the  $\Box$  modality) into the meta-language, there are two important differences.

Firstly, the *context* in CMTT is a set of variables required by a computation, which makes CMTT useful for meta-programming and staged computations. In coeffect calculi, the context requirements are formed by an abstract coeffect algebra, which is more general and can capture variable contexts, but also integers, two-point lattices, etc.

Secondly, CMTT uses different intuitive understanding of the comonad (type constructor) and the associated operations. In categorical semantics of coeffect calculi, the  $C^r\tau$  constructor refers to a value of type  $\tau$  *together* with additional context specified by the coeffect annotation  $r$  (e.g. a list of past values or additional resources).

In contrast in CMTT<sup>2</sup> the type  $C^\Psi\tau$  models a value that *requires* the context  $\Psi$  in order to produce value  $\tau$ . This also changes the interpretation of the two operations of a comonad:

$$\begin{aligned} \text{counit} &: C^{\text{use}}\alpha \rightarrow \alpha \\ \text{cobind} &: (C^{\mathbf{r}}\alpha \rightarrow \beta) \rightarrow C^{\mathbf{r} \otimes \mathbf{s}}\alpha \rightarrow C^{\mathbf{s}}\beta \end{aligned}$$

Two readings of the signatures are possible, but they give quite different meanings to the operations:

- *Coeffect interpretation.* The counit operation extracts a value and does not require any additional context; the cobind operation requires context  $\mathbf{r} \otimes \mathbf{s}$ , uses the part specified by  $\mathbf{r}$  to evaluate the function, ending with a value  $\beta$  together with remaining context  $\mathbf{s}$ .
- *CMTT interpretation.* The counit operation evaluates a computation that requires no additional context to obtain a  $\alpha$  value; given a function that turns a computation requiring context  $\mathbf{r}$  into a value  $\beta$ , the cobind operation can turn a computation that requires context  $\mathbf{r} \otimes \mathbf{s}$  into a computation that requires just  $\mathbf{s}$  and contains  $\beta$  (a part of the context requirements is eliminated by the function).

Although the different reading does not affect formal properties of the systems, it is important to understand the difference when discussing the two systems as they provide a different intuition.

The sketch of a coeffect meta-language in the following section attempts to bridge the gap between coeffects and CMTT. Just like CMTT, it embeds comonads as language constructs, but it annotates them with a (flat) coeffect algebra, thus it generalizes CMTT which tracks only sets of variables. Future work on the coeffect meta-language would thus be an interesting development for both coeffect systems and CMTT.

### 6.3.2 Coeffect meta-language

The coeffect meta-language could be designed using both flat and structural indexed comonads. For simplicity, this section only discusses the flat variant. The syntax of types and terms of the language includes the type constructor  $C^{\mathbf{r}}\tau$  and four additional language constructs:

$$\begin{aligned} \tau &::= \alpha \mid \tau_1 \rightarrow \tau_2 \mid C^{\mathbf{r}}\tau \\ e &::= v \mid \lambda x. e \mid e_1 \ e_2 \mid !e \\ &\quad \mid \text{let box } x = e_1 \text{ in } e_2 \\ &\quad \mid \text{split } e_1 \text{ into } x, y \text{ in } e_2 \\ &\quad \mid \text{merge } e_1, e_2 \text{ into } x \text{ in } e_2 \end{aligned}$$

The  $!e$  and **let box** constructs correspond to the counit and cobind operation of the comonad. To define meta-language for flat indexed comonads, we also include **split** and **merge** that embed the corresponding operations.

**TYPES FOR COEFFECT META-LANGUAGE.** The type system for the language is shown in Figure 32. The first part shows the usual typing rules for simply-typed  $\lambda$ -calculus. For simplicity, we omit typing rules for pairs, but those need to be present as the **merge** operation works on tuples.

The second part of the typing rules is more interesting. The (*counit*) operation extracts a value from a comonadic context and corresponds to variable

<sup>2</sup> To avoid using different notations, we write  $C^\Psi\tau$  instead of the original  $[\Psi]\tau$

a.) Typing rules for the simply typed  $\lambda$ -calculus

$$\begin{aligned}
 (var) \quad & \frac{x : \tau \in \Gamma}{\Gamma \vdash x : \tau} \\
 (app) \quad & \frac{\Gamma \vdash e_1 : \tau_2 \rightarrow \tau_1 \quad \Gamma \vdash e_2 : \tau_2}{\Gamma \vdash e_1 e_2 : \tau_1} \\
 (abs) \quad & \frac{\Gamma, x : \tau_1 \vdash e : \tau_2}{\Gamma \vdash \lambda x. e : \tau_1 \rightarrow \tau_2}
 \end{aligned}$$

b.) Additional typing rules arising from *flat indexed comonads*

$$\begin{aligned}
 (cobind) \quad & \frac{\Gamma \vdash e_1 : C^{r \oplus s} \tau_1 \quad \Gamma, x : C^r \tau_1 \vdash e_2 : \tau_2}{\Gamma \vdash \text{let box } x = e_1 \text{ in } e_2 : C^s \tau_2} \\
 (counit) \quad & \frac{\Gamma \vdash e : C^{use} \tau}{\Gamma \vdash !e : \tau} \\
 (split) \quad & \frac{\Gamma \vdash e_1 : C^{r \oplus s} \tau_1 \quad \Gamma, x : C^r \tau_1, y : C^s \tau_1 \vdash e_2 : \tau_2}{\Gamma \vdash \text{split } e_1 \text{ into } x, y \text{ in } e_2 : \tau_2} \\
 (merge) \quad & \frac{\Gamma \vdash e_1 : C^r \tau_1 \quad \Gamma \vdash e_2 : C^s \tau_2 \quad \Gamma, x : C^{r \wedge s}(\tau_1 \times \tau_2) \vdash e_3 : \tau_3}{\Gamma \vdash \text{merge } e_1, e_2 \text{ into } x \text{ in } e_3 : \tau_3}
 \end{aligned}$$

Figure 32: Type system for the (flat) coeffect meta-language

access in coeffect calculi. The `let box` construct (*cobind*) takes an input  $e_1$  with context  $r \oplus s$  and a computation that turns a variable  $x$  with a context  $r$  into a value  $\tau_2$ . The result is a computation that produces a  $\tau_2$  value with the remaining context specified by  $s$ . Note that the expressions  $e_2$  and  $e_1$  correspond to the first and second arguments of the *cobind* operation. The keyword `let box` is chosen following CMTT<sup>3</sup>.

The `split` and `merge` constructs follow a similar pattern. They both apply some transformation on one or two values in a context and then add the new value as a fresh variable to the variable context. We omit sub-coeffecting, but it could be easily added following the method used elsewhere.

### 6.3.3 Embedding flat coeffect calculus

The *meta-language* approach of embedding comonads in a language is more general than the *semantics* approach. This thesis focuses on a narrower use that better guides the design of a type system for context-aware programming languages.

However, it is worth demonstrating that the (flat) coeffect calculus can be embedded in the meta-language described in the previous section. This may be desirable, e. g. when using the meta-language for reasoning about context-aware computations. We briefly consider the embedding as it illuminates the relationship between coeffects and CMTT (unfortunately it is not possible to embed coeffect calculi in CMTT because of the more general annotations structure).

<sup>3</sup> The rule is similar to the `letbox` rule for ICML [58, p. 14], although it differs because of our generalization of comonads where `bind` composes coeffect annotations rather than requiring the same annotation everywhere.

$$\begin{aligned}
& \llbracket \Gamma @ \text{use} \vdash x_i : \tau_i \rrbracket_v = \pi_i(!v) \\
& \llbracket \Gamma @ r \oplus (s \otimes t) \vdash e_1 \ e_2 : \tau_2 \rrbracket_v = \\
& \quad \text{split } v \text{ into } v_s, v_{rt} \text{ in} \\
& \quad \llbracket \Gamma @ s \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \rrbracket_{v_s} (\text{let box } v_r = v_{rt} \text{ in } \llbracket \Gamma @ r \vdash e_2 : \tau_1 \rrbracket_{v_r}) \\
& \llbracket \Gamma @ r \vdash \lambda x. e : \tau_1 \xrightarrow{s} \tau_2 \rrbracket_v = \lambda x. \\
& \quad \text{merge } v, x \text{ into } v_{rs} \text{ in } \llbracket (\Gamma, x : \tau_1) @ r \wedge s \vdash e : \tau_2 \rrbracket_{v_{rs}}
\end{aligned}$$


---

Figure 33: Embedding flat coeffect calculus in coeffect meta-language

Given a typing judgement  $\Gamma @ r \vdash e : \tau$  in the flat coeffect calculus, we define  $\llbracket \Gamma @ r \vdash e : \tau \rrbracket_v$  as its embedding in the coeffect meta-language. Note that the translation is indexed by  $v$ , which is a name of variable used to represent the entire variable context of the source language. The translation is defined in Figure 33. The embedding resembles the semantics discussed in Section 4.3. This is not surprising as the meta-language directly mirrors the operations of an indexed comonad.

#### 6.4 CONCLUSIONS

In this chapter, we completed our quest of presenting a *unified* system for tracking contextual properties and then we discussed two important topics of future and related work that require in-depth discussion.

The key technical contribution of this chapter is the *unified coeffect system* (Section 6.1), which unifies the flat and structural system presented in the previous two chapters. To achieve this, we introduced *coeffect container*, which determines how are coeffect annotations attached to variable contexts. We then discussed two instances of the structure that capture flat and structural properties.

In the rest of the chapter, we discussed two topics of future and related work. First (Section 6.2), we considered pathways to practical implementations of coeffect systems, including a discussion of a novel lattice-based coeffect system, which is simpler, but cannot capture all our motivating examples. Finally, we discussed how our work relates to meta-language based on comonads (Section 6.3). We present a *coeffect meta-language* that follows similar style to CMTT, but is based on indexed comonads. Next, we discussed how to embed flat coeffect systems in this coeffect meta-language, which elucidates the relationship between our work and CMTT.

## CONCLUSIONS

---

Some of the most fundamental academic work is not the one solving hard research problems, but the one that changes how we understand the world. Some philosophers argue that *language* is the key for understanding how we think, while in science the dominant thinking is determined by *paradigms* [49] or *research programmes* [50]. In a way, programming languages play a similar role for computer science and software development.

This thesis aims to change how developers and programming language designers think about *context* or *execution environments* of programs. Such context or execution environment has numerous forms – mobile applications access network, GPS location or user’s personal data; journalists obtain information published on the web or through open government data initiatives. This thesis aims to change our understanding of *context* so that the above examples are viewed uniformly through a single programming language abstraction, which we call *coeffects*, rather than as disjoint cases. This is done through the following three contributions.

### 7.1 UNIFIED PRESENTATION.

In  $\lambda$ -calculus, the term *context* is usually used for the free-variable context. However, there are many other programming language features that are related to context or program’s execution environment. In Chapter 3, we revisit many of such features – including resource tracking in distributed computing, cross-platform development, data-flow programming and liveness analysis, but also Haskell’s type classes and implicit parameters.

The main contribution of the chapter is that it presents the disjoint language features in a unified way. We show type systems and semantics for many of the languages, illuminating the fact that they are closely related.

Considering applications is one way of approaching the theory of coeffects introduced in this thesis. Other pathways to coeffects are discussed in Chapter 2, which looks at theoretical developments leading to coeffects, including the work on effect systems, comonadic semantics and linear logics.

### 7.2 FLAT COEFFECT CALCULUS.

The applications discussed in Chapter 3 fall into two categories. In the first category (Section 3.2), the additional contextual information are *whole-context* properties. They either capture properties of the execution environment or affect the whole free-variable context.

In Chapter 4, we develop a *flat coeffect calculus* which gives us an unified way of tracking *whole-context* properties. The calculus is parameterized by a *flat coeffect algebra* that captures the algebraic properties that contextual information satisfy. Concrete instances of flat coeffects include Haskell’s implicit parameters, whole-context liveness and whole-context data-flow.

Our focus is on the syntactic properties of the calculus – we give a type system (Section 4.2) and discuss equational theory of the calculus (Section 4.4). In the flat coeffect calculus,  $\beta$ -reduction and  $\eta$ -expansion do not generally preserve type of expressions, but we identify two conditions when



this is the case – this gives us a basis for operational semantics of our calculi for liveness and implicit parameters. The design of the type system is further validated by categorical semantics (Section 4.3) which models flat coefficient calculus in terms of *flat indexed comonad* – a structure based on categorical dual of monads.

### 7.3 STRUCTURAL COEFFECT CALCULUS.

In the second category of context-aware systems discussed in Section 3.3, contextual properties are *per-variable*. The systems discussed here resemble sub-structural logics, but rather than *restricting* variable use, they *track* additional information about how variables are used.

We unify systems with *per-variable* contextual properties in Chapter 5, which describes our *structural coefficient calculus* (Section 5.2). Similarly to the flat variant, the calculus is parameterized by a *structural coefficient algebra*. Concrete instances of the calculus track bounded variable use (i.e. how many times is variable accessed), data-flow properties (how many past values are needed) and liveness (i.e. can variable be accessed).

The structural coefficient calculus has desirable equational properties (Section 5.4). In particular, type preservation for  $\beta$ -reduction and  $\eta$ -expansion holds for all structural coefficient calculi. This follows from the fact that structural coefficient associates contextual requirements with individual variables and preserves the connection by including explicit structural rules (weakening, exchange and contraction).

The Chapters 4 and 5 present the main novel technical contributions of this thesis. In Chapter 6, we develop the two calculi further by introducing a unified presentation of the two (by parameterizing the calculus by the *shape* of context) and we present an alternative presentation that makes type inference easier.

### 7.4 SUMMARY

## BIBLIOGRAPHY

---

- [1] M. Abadi, A. Banerjee, N. Heintze, and J. G. Riecke. A core calculus of dependency. In *Proceedings of POPL*, 1999.
- [2] M. Abbott, T. Altenkirch, and N. Ghani. Categories of containers. In *Foundations of Software Science and Computation Structures*, pages 23–38. Springer, 2003.
- [3] M. Abbott, T. Altenkirch, and N. Ghani. Containers: constructing strictly positive types. *Theoretical Computer Science*, 342(1):3–27, 2005.
- [4] D. Ahman, J. Chapman, and T. Uustalu. When is a container a comonad? In *Proceedings of the 15th international conference on Foundations of Software Science and Computational Structures*, FOSSACS’12, pages 74–88, Berlin, Heidelberg, 2012. Springer-Verlag.
- [5] A. W. Appel. *Modern compiler implementation in ML*. Cambridge University Press, 1998.
- [6] R. Atkey. Parameterised notions of computation. *J. Funct. Program.*, 19, 2009.
- [7] J. E. Bardram. The java context awareness framework (jcaf)—a service infrastructure and programming framework for context-aware applications. In *Pervasive Computing*, pages 98–115. Springer, 2005.
- [8] A. Benveniste, P. Caspi, S. A. Edwards, N. Halbwachs, P. Le Guernic, and R. De Simone. The synchronous languages 12 years later. *Proceedings of the IEEE*, 91(1):64–83, 2003.
- [9] G. Biegel and V. Cahill. A framework for developing mobile, context-aware applications. In *Pervasive Computing and Communications, 2004. PerCom 2004. Proceedings of the Second IEEE Annual Conference on*, pages 361–365. IEEE, 2004.
- [10] G. Bierman, M. Hicks, P. Sewell, G. Stoye, and K. Wansbrough. Dynamic rebinding for marshalling and update, with destruct-time  $\lambda$ . In *Proceedings of the eighth ACM SIGPLAN international conference on Functional programming*, ICFP ’03, pages 99–110, New York, NY, USA, 2003. ACM.
- [11] G. M. Bierman and V. C. V. de Paiva. On an intuitionistic modal logic. *Studia Logica*, 65:2000, 2001.
- [12] A. Bove, P. Dybjer, and U. Norell. A brief overview of agda—a functional language with dependent types. In *Theorem Proving in Higher Order Logics*, pages 73–78. Springer, 2009.
- [13] E. Brady. Idris, a general-purpose dependently typed programming language: Design and implementation. *Journal of Functional Programming*, 23(05):552–593, 2013.
- [14] S. Brookes and S. Geva. Computational comonads and intensional semantics. *Applications of Categories in Computer Science*. London Mathematical Society Lecture Note Series, Cambridge University Press, 1992.

- [15] A. Brunel, M. Gaboardi, D. Mazza, and S. Zdancewic. A core quantitative coeffect calculus. In *ESOP*, pages 351–370, 2014.
- [16] J. Cheney, A. Ahmed, and U. A. Acar. Provenance as dependency analysis. In *Proceedings of the 11th international conference on Database programming languages*, DBPL’07, pages 138–152, Berlin, Heidelberg, 2007. Springer-Verlag.
- [17] J. Cheney, S. Chong, N. Foster, M. Seltzer, and S. Vansummeren. Provenance: a future history. In *Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications*, pages 957–964. ACM, 2009.
- [18] J. Cheney, S. Lindley, and P. Wadler. A practical theory of language-integrated query. In *Proceedings of ICFP*, ICFP ’13, pages 403–416, 2013.
- [19] J. Clarke. *SQL Injection Attacks and Defense*. Syngress, 2009.
- [20] J.-L. Cola and M. Pouzet. Type-based initialization analysis of a synchronous dataflow language. *Int. J. Softw. Tools Technol. Transf.*, 6(3):245–255, Aug. 2004.
- [21] E. Cooper, S. Lindley, P. Wadler, and J. Yallop. Links: Web programming without tiers. FMCO ’00, 2006.
- [22] P. Costanza and R. Hirschfeld. Language constructs for context-oriented programming: an overview of context. In *Proceedings of the 2005 symposium on Dynamic languages*, DLS ’05, pages 1–10, New York, NY, USA, 2005. ACM.
- [23] K. Crary, D. Walker, and G. Morrisett. Typed memory management in a calculus of capabilities. In *Proceedings of the 26th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 262–275. ACM, 1999.
- [24] R. Davies and F. Pfenning. A modal analysis of staged computation. *J. ACM*, 48(3):555–604, May 2001.
- [25] Developers (Android). Creating multiple APKs for different API levels. <http://developer.android.com/training/multiple-aps/api.html>, 2013.
- [26] W. Du and L. Wang. Context-aware application programming for mobile devices. In *Proceedings of the 2008 C3S2E conference*, C3S2E ’08, pages 215–227, New York, NY, USA, 2008. ACM.
- [27] A. Filinski. Monads in action. In *Proceedings of POPL*, 2010.
- [28] A. Filinski. Towards a comprehensive theory of monadic effects. In *Proceeding of the 16th ACM SIGPLAN international conference on Functional programming*, ICFP ’11, pages 1–1, 2011.
- [29] C. Flanagan and M. Abadi. Types for Safe Locking. *ESOP ’99*, 1999.
- [30] C. Flanagan and S. Qadeer. A type and effect system for atomicity. In *Proceedings of Conference on Programming Language Design and Implementation*, PLDI ’03.

- [31] O. Frieder and M. E. Segal. On dynamically updating a computer program: From concept to prototype. *Journal of Systems and Software*, 14(2):111–128, 1991.
- [32] M. Gabbay and A. Nanevski. Denotation of syntax and metaprogramming in contextual modal type theory (cmtt). *CoRR*, abs/1202.0904, 2012.
- [33] D. K. Gifford and J. M. Lucassen. Integrating functional and imperative programming. In *Proceedings of Conference on LISP and func. prog.*, LFP '86, 1986.
- [34] G. Giorgidze, T. Grust, N. Schweinsberg, and J. Weijers. Bringing back monad comprehensions. *ACM SIGPLAN Notices*, 46(12):13–22, 2012.
- [35] J.-Y. Girard, A. Scedrov, and P. J. Scott. Bounded linear logic: a modular approach to polynomial-time computability. *Theoretical computer science*, 97(1):1–66, 1992.
- [36] Google. What is API level. Retrieved from <http://developer.android.com/guide/topics/manifest/uses-sdk-element.html#ApiLevels>.
- [37] N. Halbwachs, P. Caspi, P. Raymond, and D. Pilaud. The synchronous data flow programming language lustre. *Proceedings of the IEEE*, 79(9):1305–1320, 1991.
- [38] W. Halfond, A. Orso, and P. Manolios. Wasp: Protecting web applications using positive tainting and syntax-aware evaluation. *IEEE Trans. Softw. Eng.*, 34(1):65–81, Jan. 2008.
- [39] W. G. Halfond, A. Orso, and P. Manolios. Using positive tainting and syntax-aware evaluation to counter sql injection attacks. In *Proceedings of the 14th ACM SIGSOFT international symposium on Foundations of software engineering*, pages 175–185. ACM, 2006.
- [40] T. Harris, S. Marlow, S. Peyton-Jones, and M. Herlihy. Composable memory transactions. In *Proceedings of the tenth ACM SIGPLAN symposium on Principles and practice of parallel programming*, pages 48–60. ACM, 2005.
- [41] M. Hicks, J. T. Moore, and S. Nettles. *Dynamic software updating*, volume 36. ACM, 2001.
- [42] R. Hirschfeld, P. Costanza, and O. Nierstrasz. Context-oriented programming. *Journal of Object Technology*, 7(3), 2008.
- [43] S. L. P. Jones. *Haskell 98 language and libraries: the revised report*. Cambridge University Press, 2003.
- [44] P. Jouvelot and D. K. Gifford. Communication Effects for Message-Based Concurrency. Technical report, Massachusetts Institute of Technology, 1989.
- [45] S.-y. Katsumata. Parametric effect monads and semantics of effect systems. In *Proceedings of the 41st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '14, pages 633–645, New York, NY, USA, 2014. ACM.

- [46] A. Kennedy. Types for units-of-measure: Theory and practice. In *Central European Functional Programming School*, pages 268–305. Springer, 2010.
- [47] A. J. Kennedy. Relational parametricity and units of measure. In *Proceedings of the 24th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 442–455. ACM, 1997.
- [48] R. B. Kieburtz. Codata and Comonads in Haskell, 1999.
- [49] T. S. Kuhn. *The structure of scientific revolutions*. University of Chicago Press, 1970.
- [50] I. Lakatos. *Methodology of Scientific Research Programmes: Philosophical Papers: v. 1*. Cambridge University Press.
- [51] J. R. Lewis, M. B. Shields, E. Meijert, and J. Launchbury. Implicit parameters: dynamic scoping with static types. In *Proceedings of POPL*, POPL '00, 2000.
- [52] F. Loitsch and M. Serrano. Hop client-side compilation. *Trends in Functional Programming, TFP*, pages 141–158, 2007.
- [53] J. M. Lucassen and D. K. Gifford. Polymorphic effect systems. In *Proceedings of the 15th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, POPL '88, pages 47–57, New York, NY, USA, 1988. ACM.
- [54] E. Meijer, B. Beckman, and G. Bierman. Linq: reconciling object, relations and xml in the .net framework. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, SIGMOD '06, pages 706–706, New York, NY, USA, 2006. ACM.
- [55] E. Moggi. Notions of computation and monads. *Inf. Comput.*, 93:55–92, July 1991.
- [56] T. Murphy, VII., K. Crary, and R. Harper. Type-safe distributed programming with ML5. *TGC'07*, pages 108–123, 2008.
- [57] T. Murphy VII, K. Crary, R. Harper, and F. Pfenning. A symmetric modal lambda calculus for distributed computing. *LICS '04*, pages 286–295, 2004.
- [58] A. Nanevski, F. Pfenning, and B. Pientka. Contextual modal type theory. *ACM Trans. Comput. Logic*, 9(3):23:1–23:49, June 2008.
- [59] D. L. Niki Vazou. Remarrying effects and monads. *Proceedings of MSFP (to appear)*, 2014.
- [60] P. O'Hearn. On bunched typing. *J. Funct. Program.*, 13(4):747–796, July 2003.
- [61] P. W. O'Hearn, J. C. Reynolds, and H. Yang. Local reasoning about programs that alter data structures. In *Proceedings of the 15th International Workshop on Computer Science Logic*, CSL '01, pages 1–19, London, UK, UK, 2001. Springer-Verlag.
- [62] D. Orchard. Programming contextual computations.

- [63] D. Orchard. Should I use a Monad or a Comonad? Unpublished draft, 2012.
- [64] D. Orchard and A. Mycroft. Efficient and correct stencil computation via pattern matching and static typing. In *Proceedings of DSL 2011*, arXiv preprint arXiv:1109.0777, 2011.
- [65] D. Orchard and A. Mycroft. A notation for comonads. In *Implementation and Application of Functional Languages*, pages 1–17. Springer, 2013.
- [66] D. Orchard and T. Petricek. Embedding effect systems in haskell. In *Proceedings of the 2014 ACM SIGPLAN Symposium on Haskell*, Haskell '14, pages 13–24, 2014.
- [67] T. Petricek. Client-side scripting using meta-programming.
- [68] T. Petricek. Evaluations strategies for monadic computations. In *Proceedings of Mathematically Structured Functional Programming*, MSFP 2012.
- [69] T. Petricek. Understanding the world with f#. Available at <http://channel9.msdn.com/posts/Understanding-the-World-with-F>.
- [70] T. Petricek, D. Orchard, and A. Mycroft. Coeffects: unified static analysis of context-dependence. In *Proceedings of International Conference on Automata, Languages, and Programming - Volume Part II*, ICALP 2013.
- [71] T. Petricek, D. Orchard, and A. Mycroft. Coeffects: A calculus of context-dependent computation. In *Proceedings of the 19th ACM SIGPLAN International Conference on Functional Programming*, ICFP '14, pages 123–135, 2014.
- [72] T. Petricek and D. Syme. The f# computation expression zoo. In *Proceedings of Practical Aspects of Declarative Languages*, PADL 2014.
- [73] F. Pfenning and R. Davies. A judgmental reconstruction of modal logic. *Mathematical Structures in Comp. Sci.*, 11(4):511–540, Aug. 2001.
- [74] A. Russo, K. Claessen, and J. Hughes. A library for light-weight information-flow security in haskell. In *Proceedings of the first ACM SIGPLAN symposium on Haskell*, Haskell '08, pages 13–24, 2008.
- [75] A. Sabelfeld and A. C. Myers. Language-based information-flow security. *IEEE J.Sel. A. Commun.*, 21(1):5–19, Sept. 2006.
- [76] T. Sans and I. Cervesato. QWeSST for Type-Safe Web Programming. In *Third International Workshop on Logics, Agents, and Mobility*, LAM'10, 2010.
- [77] M. Serrano. Hop, a fast server for the diffuse web. In *Coordination Models and Languages*, pages 1–26. Springer, 2009.
- [78] P. Sewell, J. J. Leifer, K. Wansbrough, F. Z. Nardelli, M. Allen-Williams, P. Habouzit, and V. Vafeiadis. Acute: High-level programming language design for distributed computation. *J. Funct. Program.*, 17(4-5):547–612, July 2007.
- [79] V. Simonet. Flow caml in a nutshell. In *Proceedings of the first APPSEM-II workshop*, pages 152–165, 2003.

- [80] G. Stoyale, M. Hicks, G. Bierman, P. Sewell, and I. Neamtiu. Mutatis mutandis: safe and predictable dynamic software updating. In *ACM SIGPLAN Notices*, volume 40, pages 183–194. ACM, 2005.
- [81] N. Swamy, N. Guts, D. Leijen, and M. Hicks. Lightweight monadic programming in ml. In *Proceedings of the 16th ACM SIGPLAN international conference on Functional programming, ICFP '11*, pages 15–27, New York, NY, USA, 2011. ACM.
- [82] D. Syme. Leveraging .NET meta-programming components from F#: integrated queries and interoperable heterogeneous execution. In *Proceedings of the 2006 workshop on ML*, pages 43–54. ACM, 2006.
- [83] D. Syme, K. Battocchi, K. Takeda, D. Malayeri, and T. Petricek. Themes in information-rich functional programming for internet-scale data sources. In *Proceedings of the 2013 Workshop on Data Driven Functional Programming, DDFP '13*, pages 1–4, 2013.
- [84] D. Syme, A. Granicz, and A. Cisternino. Building mobile web applications. In *Expert F# 3.0*, pages 391–426. Springer, 2012.
- [85] D. Syme, T. Petricek, and D. Lomov. The f# asynchronous programming model. In *Practical Aspects of Declarative Languages*, pages 175–189. Springer, 2011.
- [86] J. Talpin and P. Jouvelot. The type and effect discipline. In *Logic in Computer Science, 1992. LICS'92.*, pages 162–173, 1994.
- [87] R. Tate. The sequential semantics of producer effect systems. In *Proceedings of the 40th annual ACM SIGPLAN-SIGACT symposium on Principles of programming languages, POPL '13*, pages 15–26, New York, NY, USA, 2013. ACM.
- [88] The F# Software Foundation. F#. See <http://fsharp.org>, 2014.
- [89] P. Thiemann. A unified framework for binding-time analysis. In *TAPSOFT'97: Theory and Practice of Software Development*, pages 742–756. Springer, 1997.
- [90] F. Tip. A survey of program slicing techniques. *Journal of programming languages*, 3(3):121–189, 1995.
- [91] M. Tofte and J.-P. Talpin. Region-based memory management. *Information and Computation*, 132(2):109–176, 1997.
- [92] T. Uustalu and V. Vene. The essence of dataflow programming. In *Proceedings of the Third Asian conference on Programming Languages and Systems, APLAS'05*, pages 2–18, Berlin, Heidelberg, 2005. Springer-Verlag.
- [93] T. Uustalu and V. Vene. Comonadic Notions of Computation. *Electron. Notes Theor. Comput. Sci.*, 203:263–284, June 2008.
- [94] T. Uustalu and V. Vene. The Essence of Dataflow Programming. *Lecture Notes in Computer Science*, 4164:135–167, Nov 2006.
- [95] P. Vogt, F. Nentwich, N. Jovanovic, E. Kirda, C. Kruegel, and G. Vigna. Cross site scripting prevention with dynamic data tainting and static analysis. In *Proceeding of the Network and Distributed System Security Symposium (NDSS)*, volume 42, 2007.



- [96] D. Volpano, C. Irvine, and G. Smith. A sound type system for secure flow analysis. *J. Comput. Secur.*, 4:167–187, January 1996.
- [97] J. Vouillon and V. Balat. From bytecode to javascript: the js\_of\_ocaml compiler. *Software: Practice and Experience*, 2013.
- [98] B. Wadge. Monads and intensionality. In *International Symposium on Lucid and Intensional Programming*, volume 95, 1995.
- [99] W. W. Wadge and E. A. Ashcroft. *LUCID, the dataflow programming language*. Academic Press Professional, Inc., San Diego, CA, USA, 1985.
- [100] P. Wadler. Strictness analysis aids time analysis. In *Proceedings of the 15th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, pages 119–132. ACM, 1988.
- [101] P. Wadler. Linear types can change the world! In *Programming Concepts and Methods*. North, 1990.
- [102] P. Wadler and S. Blott. How to make ad-hoc polymorphism less ad hoc. In *Proceedings of the 16th ACM SIGPLAN-SIGACT symposium on Principles of programming languages, POPL '89*, pages 60–76, New York, NY, USA, 1989. ACM.
- [103] P. Wadler and P. Thiemann. The marriage of effects and monads. *ACM Trans. Comput. Logic*, 4:1–32, January 2003.
- [104] D. Walker. *Substructural Type Systems*, pages 3–43. MIT Press.
- [105] H. Xi. Dependent ml an approach to practical programming with dependent types. *Journal of Functional Programming*, 17(02):215–286, 2007.



## APPENDIX A

This appendix gives full typing derivations for the examples presented in Chapter 3, which demonstrate simple coeffect calculi for implicit parameters, liveness and data-flow.

## A.1 COEFFECT TYPING FOR IMPLICIT PARAMETERS

The example on page 31 considers the typing of a function that adds the values of two implicit parameters:  $\lambda x. ?a + ?b$ . Note that addition is a function (and so we write  $(+) ?a ?b$ ) that requires no implicit parameters. Thus our typing derivation starts with:

$$\Gamma_0 = (+) : \text{int} \xrightarrow{\emptyset} \text{int} \xrightarrow{\emptyset} \text{int}$$

The typing for the sub-expression  $(+) ?a$  is shown in Figure 34 (a). Note that the resulting function does not require any implicit parameters, so there is no non-determinism so far. It is worth noting that this would change if we used  $\eta$ -expansion and wrote  $(\lambda y. (+) ?a)$ . We refer to the above as  $\Delta$ . The rest of the expression is shown in Figure 34 (b).

Here, the last rule applies (*app*) and so we non-deterministically split the set of required resources. This means that we need  $r_1, r_2$  such that  $r_1 \cup r_2 = \{?a : \text{int}, ?b : \text{int}\}$ . The Figure 34 (c) summarizes the 9 options.

## A.2 COEFFECT TYPING FOR LIVENESS

The Section 3.2.3 discusses a coeffect system where the coeffect annotations capture whether a variable may be accessed (marked as live using **L**) or whether it is definitely not used (marked as dead using **D**). The following three examples are considered:

$$(\lambda x. 42) \ y \quad (1)$$

$$\text{twoTimes } 42 \quad (2)$$

$$(\lambda x. x) \ 42 \quad (3)$$

The typing derivation for the expression (1) is shown in Figure 35 (a). The most interesting aspect about the previous example is the use of the (*app*) rule, which marks the resulting context as dead, even though a variable is accessed in the second part of the expression (this part never needs to be evaluated).

Assuming  $\Gamma_0 = \text{twoTimes} : \text{int} \xrightarrow{\text{L}} \text{int}$ , the typing derivation for (2) is shown in Figure 35 (b). Here, the variable context is marked as live. This is not because the argument of *twoTimes* is marked as live, but because the function itself is a variable that (always) needs to be obtained from the variable context. Finally, the derivation for (3) is shown in Figure 35 (c).

## A.3 COEFFECT TYPING FOR DATA-FLOW

The Section 3.2.4 presents a coeffect type system that tracks the maximal number of past values required by a data-flow computation. The discussion includes the following example:

```
( if (prev tick) = 0
  then (λx → prev x)
  else (λx → x) )    (prev counter)
```

In order to give typing for the above example, we first need to extend the language with conditionals. The typing rule for the **if** construct is:

$$(if) \frac{\Gamma @ \mathbf{m} \vdash e : \text{bool} \quad \Gamma @ \mathbf{n} \vdash e_1 : \tau \quad \Gamma @ \mathbf{n} \vdash e_2 : \tau}{\Gamma @ \max(\mathbf{m}, \mathbf{n}) \vdash \text{if } e \text{ then } e_1 \text{ else } e_2 : \tau_1}$$

Given a condition that requires  $\mathbf{m}$  past and two (alternative) bodies that each require  $\mathbf{n}$  past values, the composed expression requires at most the maximum of the two, i.e. the same context is passed to both the condition and the body using point-wise composition. As we will see, the fact that both branches require the same number of past values means that we need to use the sub-coeffecting rule.

Assuming  $\Gamma_0 = \text{tick} : \text{int}, \text{counter} : \text{int}$ , the Figure 36 shows the typing derivation for the example. In the first part, we derive the types for the sub-expressions of the application and the conditional. Note that to obtain compatible function types, we use sub-coeffecting *before* abstraction in the typing of  $(\lambda x.x)$  – the type  $\text{int} \xrightarrow{1} \text{int}$  is a valid overapproximation. Finally, the typing of the application yields the requirement of two past values, calculated as  $\max(1, 1 + 1)$ .

a.) Typing for the sub-expression  $(+) ?a$

$$\begin{array}{c}
 (var) \quad \frac{}{\Gamma_0, x : \text{int} @ \emptyset \vdash (+) : \text{int} \xrightarrow{\emptyset} \text{int} \xrightarrow{\emptyset} \text{int}} \\
 (app) \quad \frac{\Gamma_0, x : \text{int} @ \{?a : \text{int}\} \vdash ?a : \text{int} \quad (param)}{\Gamma_0, x : \text{int} @ \{?a : \text{int}\} \vdash (+) ?a : \text{int} \xrightarrow{\emptyset} \text{int}}
 \end{array}$$

b.) Typing for the expression  $(+) ?a ?b$

$$\begin{array}{c}
 (app) \quad \frac{\Delta \quad \Gamma_0, x : \text{int} @ \{?b : \text{int}\} \vdash ?b : \text{int}}{\Gamma_0, x : \text{int} @ \{?a : \text{int}, ?b : \text{int}\} \vdash (+) ?a ?b : \text{int}} \\
 (abs) \quad \frac{}{\Gamma_0 @ \mathbf{r}_1 \vdash \lambda x. (+) ?a ?b : \text{int} \xrightarrow{\mathbf{r}_2} \text{int}}
 \end{array}$$

c.) Possible splittings of the implicit parameters

$$\begin{array}{ll}
 \mathbf{r}_1 = \{\} & \mathbf{r}_2 = \{?a : \text{int}, ?b : \text{int}\} \\
 \mathbf{r}_1 = \{?a : \text{int}\} & \mathbf{r}_2 = \{?a : \text{int}, ?b : \text{int}\} \\
 \mathbf{r}_1 = \{?b : \text{int}\} & \mathbf{r}_2 = \{?a : \text{int}, ?b : \text{int}\} \\
 \mathbf{r}_1 = \{?a : \text{int}, ?b : \text{int}\} & \mathbf{r}_2 = \{?a : \text{int}, ?b : \text{int}\} \\
 \mathbf{r}_1 = \{?a : \text{int}, ?b : \text{int}\} & \mathbf{r}_2 = \{\} \\
 \mathbf{r}_1 = \{?a : \text{int}, ?b : \text{int}\} & \mathbf{r}_2 = \{?a : \text{int}\} \\
 \mathbf{r}_1 = \{?a : \text{int}, ?b : \text{int}\} & \mathbf{r}_2 = \{?b : \text{int}\} \\
 \mathbf{r}_1 = \{?a : \text{int}\} & \mathbf{r}_2 = \{?b : \text{int}\} \\
 \mathbf{r}_1 = \{?b : \text{int}\} & \mathbf{r}_2 = \{?a : \text{int}\}
 \end{array}$$

---

Figure 34: Coeffect typing for implicit parameters

a.) Typing for the expression  $(\lambda x.42) y$

$$\begin{array}{c}
 \text{(const)} \quad \frac{}{y:\tau, x:\tau @ \mathbf{D} \vdash 42 : \text{int}} \quad \text{(var)} \quad \frac{}{y:\tau @ \mathbf{L} \vdash y : \tau} \\
 \text{(abs)} \quad \frac{}{y:\tau @ \mathbf{D} \vdash \lambda x.42 : \tau \xrightarrow{\mathbf{D}} \text{int}} \\
 \text{(app)} \quad \frac{}{y:\tau @ \mathbf{D} \sqcup (\mathbf{L} \sqcap \mathbf{D}) \vdash (\lambda x.42) y : \text{int}} \\
 \hline
 y:\tau @ \mathbf{D} \vdash (\lambda x.42) y : \text{int}
 \end{array}$$

b.) Typing for the expression `twoTimes 42`

$$\begin{array}{c}
 \text{(var)} \quad \frac{}{\Gamma_0 @ \mathbf{L} \vdash \text{twoTimes} : \text{int} \xrightarrow{\mathbf{L}} \text{int}} \quad \text{(var)} \quad \frac{}{\Gamma_0 @ \mathbf{D} \vdash 42 : \text{int}} \\
 \text{(app)} \quad \frac{}{\Gamma_0 @ \mathbf{L} \sqcup (\mathbf{D} \sqcap \mathbf{L}) \vdash \text{twoTimes } 42 : \text{int}} \\
 \hline
 \Gamma_0 @ \mathbf{L} \vdash \text{twoTimes } 42 : \text{int}
 \end{array}$$

c.) Typing for the expression  $(\lambda x.x) 42$

$$\begin{array}{c}
 \text{(var)} \quad \frac{}{x:\text{int} @ \mathbf{L} \vdash x : \text{int}} \quad \text{(const)} \quad \frac{}{() @ \mathbf{D} \vdash 42 : \text{int}} \\
 \text{(abs)} \quad \frac{}{() @ \mathbf{L} \vdash \lambda x.x : \text{int} \xrightarrow{\mathbf{L}} \text{int}} \\
 \text{(app)} \quad \frac{}{() @ \mathbf{L} \sqcup (\mathbf{L} \sqcap \mathbf{D}) \vdash (\lambda x.x) 42 : \text{int}} \\
 \hline
 () @ \mathbf{L} \vdash (\lambda x.x) 42 : \text{int}
 \end{array}$$

---

Figure 35: Coeffect typing for liveness

a.) Typing for sub-expressions of the conditional and for the argument

$$\begin{array}{c}
 \text{(prev)} \frac{\Gamma_0 @ 0 \vdash \text{tick} : \text{int}}{\Gamma_0 @ 1 \vdash \text{prev tick} : \text{int}} \\
 \hline
 \Gamma_0 @ 1 \vdash (\text{prev tick}) = 0 : \text{bool}
 \\
 \\
 \text{(prev)} \frac{\text{(var)} \frac{}{\Gamma_0, x : \text{int} @ 0 \vdash x : \text{int}}}{\Gamma_0, x : \text{int} @ 1 \vdash \text{prev } x : \text{int}} \\
 \text{(abs)} \frac{}{\Gamma_0 @ 1 \vdash \lambda x. \text{prev } x : \text{int} \xrightarrow{1} \text{int}}
 \\
 \\
 \text{(abs)} \frac{\text{(sub)} \frac{\text{(var)} \frac{}{\Gamma_0, x : \text{int} @ 0 \vdash x : \text{int}}}{\Gamma_0, x : \text{int} @ 1 \vdash x : \text{int}}}{\Gamma_0 @ 0 \vdash \lambda x. x : \text{int} \xrightarrow{1} \text{int}}
 \\
 \\
 \text{(prev)} \frac{\text{(var)} \frac{}{\Gamma_0 @ 0 \vdash \text{counter} : \text{int}}}{\Gamma_0 @ 1 \vdash \text{prev counter} : \text{int}}
 \end{array}$$

b.) Typing for the composed expression

$$\begin{array}{c}
 \text{(if)} \frac{(\dots)}{\Gamma_0 @ 1 \vdash (\text{if } \dots \text{ then } \dots \text{ else } \dots) : \text{int} \xrightarrow{1} \text{int}} \quad \Gamma_0 @ 1 \vdash (\dots) : \text{int} \\
 \text{(app)} \frac{\Gamma_0 @ 1 \vdash (\text{if } \dots \text{ then } \dots \text{ else } \dots) : \text{int} \xrightarrow{1} \text{int}}{\Gamma_0 @ \max(1, 1 + 1) \vdash (\text{if } \dots \text{ then } \dots \text{ else } \dots) (\dots) : \text{int}} \\
 \hline
 \Gamma_0 @ 2 \vdash (\text{if } \dots \text{ then } \dots \text{ else } \dots) (\dots) : \text{int}
 \end{array}$$

Figure 36: Coeffect typing for data-flow





## APPENDIX B

This appendix provides additional details for some of the proofs for equational theory of flat coeffect calculus from Chapter 4 and structural coeffect calculus from Chapter 5.

## B.1 SUBSTITUTION FOR FLAT COEFFECTS

In Section 4.4.3, we stated that, for a bottom-pointed flat coeffect algebra (i.e.  $\forall r \in \mathcal{C} . r \geq \text{use}$ ), the call-by-name substitution preserves type if all operators of the flat coeffect algebra coincide (Lemma 5). This section provides the corresponding proof.

**Lemma** (Bottom-pointed substitution). *In a bottom-pointed flat coeffect calculus with an algebra  $(\mathcal{C}, \otimes, \oplus, \wedge, \text{use}, \text{ign}, \leq)$  where  $\wedge = \otimes = \oplus$  and the operation is also idempotent and commutative and  $r \leq r' \Rightarrow \forall s. r \otimes s \leq r' \otimes s$  then:*

$$\begin{aligned} \Gamma @ S \vdash e_s : \tau_s \quad \wedge \quad \Gamma_1, x : \tau_s, \Gamma_2 @ r \vdash e_r : \tau_r \\ \Rightarrow \quad \Gamma_1, \Gamma, \Gamma_2 @ r \otimes S \vdash e_r[x \leftarrow e_s] : \tau_r \end{aligned}$$

*Proof.* Assume that  $\Gamma @ S \vdash e_s : \tau_s$  and we are substituting a term  $e_s$  for a variable  $x$ . Note that we use upper-case  $S$  to distinguish the coeffect of the expression that is being substituted into an expression. Using structural induction over  $\vdash$ :

(VAR) Given the following derivation using (var):

$$\frac{}{\Gamma_1, y : \tau, \Gamma_2 @ \text{use} \vdash y : \tau}$$

There are two cases depending on whether  $y$  is the variable  $x$  or not:

- If  $y = x$  then also  $\tau = \tau_s$  and thus  $y[x \leftarrow e_s] = e_s$ . Using the assumption, implicit weakening and the fact that  $\text{use}$  is a unit of  $\otimes$ :

$$\frac{\frac{\Gamma @ S \vdash y[x \leftarrow e_s] : \tau_s}{\Gamma_1, \Gamma, \Gamma_2 @ S \vdash y[x \leftarrow e_s] : \tau}}{\Gamma_1, \Gamma, \Gamma_2 @ \text{use} \otimes S \vdash y[x \leftarrow e_s] : \tau}$$

- If  $y \neq x$  then  $y[x \leftarrow e_s] = y$ . Using the fact that  $\text{use}$  is the bottom element and sub-coeffecting:

$$\frac{\Gamma_1, y : \tau, \Gamma_2 @ \text{use} \vdash y : \tau}{\Gamma_1, y : \tau, \Gamma_2 @ \text{use} \otimes S \vdash y : \tau}$$

(CONST) Similar to the (var) case when the variable is not substituted.

(SUB) Given the following typing derivation using (sub):

$$\frac{\Gamma_1, x : \tau_s, \Gamma_2 @ r' \vdash e : \tau}{\Gamma_1, x : \tau_s, \Gamma_2 @ r \vdash e : \tau} \quad (r' \leq r)$$

From the induction hypothesis, we have that  $\Gamma_1, \Gamma, \Gamma_2 @ r' \otimes S \vdash e[x \leftarrow e_s] : \tau$ . The condition on  $\leq$  means that  $r' \otimes S \leq r \otimes S$  and so we can apply the (sub) rule to obtain  $\Gamma_1, \Gamma, \Gamma_2 @ r \otimes S \vdash e[x \leftarrow e_s] : \tau$ .

(ABS) Given the following typing derivation using *(abs)*:

$$\frac{\Gamma_1, x : \tau_s, \Gamma_2, y : \tau_1 @ \mathbf{r} \wedge \mathbf{s} \vdash e : \tau_2}{\Gamma_1, x : \tau_s, \Gamma_2 @ \mathbf{r} \vdash \lambda y. e : \tau_1 \xrightarrow{s} \tau_2}$$

Assume w.l.o.g. that  $x \neq y$ . From the induction hypothesis, we have that:

$$\Gamma_1, \Gamma, \Gamma_2, y : \tau_1 @ \mathbf{r} \otimes \mathbf{s} \vdash e[x \leftarrow e_s] : \tau_2$$

Now using the fact that  $\wedge = \otimes$ , associativity and commutativity and *(abs)*:

$$\frac{\frac{\Gamma_1, \Gamma, \Gamma_2, y : \tau_1 @ (\mathbf{r} \wedge \mathbf{s}) \otimes \mathbf{S} \vdash e[x \leftarrow e_s] : \tau_2}{\Gamma_1, \Gamma, \Gamma_2, y : \tau_1 @ (\mathbf{r} \otimes \mathbf{S}) \wedge \mathbf{s} \vdash e[x \leftarrow e_s] : \tau_2}}{\frac{\Gamma_1, \Gamma, \Gamma_2 @ \mathbf{r} \otimes \mathbf{S} \vdash \lambda y. (e[x \leftarrow e_s]) : \tau_1 \xrightarrow{s} \tau_2}{\Gamma_1, \Gamma, \Gamma_2 @ \mathbf{r} \otimes \mathbf{S} \vdash (\lambda y. e)[x \leftarrow e_s] : \tau_1 \xrightarrow{s} \tau_2}}$$

(APP) Given the following typing derivation using *(app)*:

$$\frac{\Gamma_1, x : \tau_s, \Gamma_2 @ \mathbf{r} \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \quad \Gamma_1, x : \tau_s, \Gamma_2 @ \mathbf{s} \vdash e_2 : \tau_1}{\Gamma_1, x : \tau_s, \Gamma_2 @ \mathbf{r} \oplus (\mathbf{s} \otimes \mathbf{t}) \vdash e_1 e_2 : \tau_2}$$

From the induction hypothesis, we have that:

$$\begin{aligned} \Gamma_1, \Gamma, \Gamma_2 @ \mathbf{r} \otimes \mathbf{S} \vdash e_1[x \leftarrow e_s] : \tau_1 \xrightarrow{t} \tau_2 \\ \Gamma_1, \Gamma, \Gamma_2 @ \mathbf{s} \otimes \mathbf{S} \vdash e_2[x \leftarrow e_s] : \tau_1 \end{aligned} \quad (*)$$

Now using the *(app)* rule and the fact that  $\oplus = \otimes$ , associativity, commutativity and idempotence (note that all three properties are needed):

$$\frac{(*)}{\frac{\Gamma_1, \Gamma, \Gamma_2 @ (\mathbf{r} \otimes \mathbf{S}) \oplus ((\mathbf{s} \otimes \mathbf{S}) \otimes \mathbf{t}) \vdash e_1[x \leftarrow e_s] e_2[x \leftarrow e_s] : \tau_2}{\Gamma_1, \Gamma, \Gamma_2 @ (\mathbf{r} \oplus (\mathbf{s} \otimes \mathbf{t})) \otimes \mathbf{S} \vdash (e_1 e_2)[x \leftarrow e_s] : \tau_2}}$$

(LET) Given the following typing derivation using *(let)*:

$$\frac{\Gamma_1, x : \tau_s, \Gamma_2 @ \mathbf{r} \vdash e_1 : \tau_1 \quad \Gamma_1, x : \tau_s, \Gamma_2, y : \tau_1 @ \mathbf{s} \vdash e_2 : \tau_2}{\Gamma_1, x : \tau_s, \Gamma_2 @ \mathbf{s} \oplus (\mathbf{s} \otimes \mathbf{r}) \vdash \mathbf{let} \ y = e_1 \ \mathbf{in} \ e_2 : \tau_2}$$

From the induction hypothesis, we have that:

$$\begin{aligned} \Gamma_1, \Gamma, \Gamma_2 @ \mathbf{r} \otimes \mathbf{S} \vdash e_1[x \leftarrow e_s] : \tau_1 \\ \Gamma_1, \Gamma, \Gamma_2, y : \tau_1 @ \mathbf{s} \otimes \mathbf{S} \vdash e_2[x \leftarrow e_s] : \tau_2 \end{aligned} \quad (\dagger)$$

Now using the *(let)* rule and similarly to the *(app)* case:

$$\frac{(\dagger)}{\frac{\Gamma_1, \Gamma, \Gamma_2 @ (\mathbf{s} \otimes \mathbf{S}) \oplus ((\mathbf{s} \otimes \mathbf{S}) \otimes (\mathbf{r} \otimes \mathbf{S})) \vdash \mathbf{let} \ y = e_1[x \leftarrow e_s] \ \mathbf{in} \ e_2[x \leftarrow e_s] : \tau_2}{\Gamma_1, \Gamma, \Gamma_2 @ (\mathbf{s} \oplus (\mathbf{s} \otimes \mathbf{r})) \otimes \mathbf{S} \vdash (\mathbf{let} \ y = e_1 \ \mathbf{in} \ e_2)[x \leftarrow e_s] : \tau_2}}$$

□

## B.2 SUBSTITUTION FOR STRUCTURAL COEFFECTS

In order to prove that  $\beta$ -reduction and  $\eta$ -expansion preserve the type of an expression in Section 5.4.2, we required a multi-nary form of the substitution lemma (Lemma 14). This section provides the corresponding proof.

**Lemma** (Multi-nary substitution). *Given an expression with multiple holes filled by variables  $x_{T_i} : \tau_{T_i}$  with coefficients  $s_k$ :*

$$\Gamma @ \mathbf{r} [x_{T_1} : \tau_{T_1} @ \langle s_1 \rangle | \dots | x_{T_k} : \tau_{T_k} @ \langle s_k \rangle] \vdash e_r : \tau_r$$

*and a expressions  $e_{T_i}$  with free-variable contexts  $\Gamma_{T_i}$  annotated with  $\mathbf{T}_i$ :*

$$\Gamma_1 @ \mathbf{T}_1 \vdash e_{T_1} : \tau_{T_1} \quad \dots \quad \Gamma_k @ \mathbf{T}_k \vdash e_{T_k} : \tau_{T_k}$$

*then substituting the expressions  $e_{T_i}$  for variables  $x_{T_i}$  results in an expression with a context where the original holes are filled by contexts  $\Gamma_{T_i}$  with coefficients  $s_i \otimes \mathbf{T}_i$ :*

$$\Gamma @ \mathbf{r} [\Gamma_{T_1} @ s_1 \otimes \mathbf{T}_1 | \dots | \Gamma_{T_k} @ s_k \otimes \mathbf{T}_k] \vdash e_r[x_{T_1} \leftarrow e_{T_1}] \dots [x_{T_k} \leftarrow e_{T_k}] : \tau_r$$

*Proof.* Assume that  $\Gamma @ \mathbf{T}_i \vdash e_{T_i} : \tau_{T_i}$  and we are substituting terms  $e_{T_i}$  for variables  $x_{T_i}$ . Furthermore, we assume that are variables that are being substituted for actually appear in the original term (which means that  $k$  is at most the number of variables). Note that we use upper-case  $\mathbf{T}$  to distinguish the coefficient of the expression that is being substituted into an expression. Using structural induction over  $\vdash$ :

## SYNTAX-DRIVEN TYPING RULES

(VAR) Given the following derivation using (var):

$$\frac{}{y : \tau @ \langle \text{use} \rangle \vdash y : \tau}$$

Here, the context contains exactly one variable and so  $k = 1$ . There are two cases depending on whether  $y$  is the (only substituted) variables  $x_{T_1}$  or not:

- If  $y = x_{T_1}$  then also  $\tau = \tau_{T_1}$  and thus  $y[x_{T_1} \leftarrow e_{T_1}] = e_{T_1}$ . In this case, the context contains only a single hole  $\Gamma @ \mathbf{r} = - @ -$ . Using the assumption and the fact that  $\text{use}$  is a unit of  $\otimes$ :

$$\frac{\frac{\Gamma_{T_1} @ \mathbf{T}_1 \vdash y[x_{T_1} \leftarrow e_{T_1}] : \tau_r}{\Gamma @ \mathbf{r} [\Gamma_{T_1} @ \mathbf{T}_1] \vdash y[x_{T_1} \leftarrow e_{T_1}] : \tau_r}}{\Gamma @ \mathbf{r} [\Gamma_{T_1} @ \text{use} \otimes \mathbf{T}_1] \vdash y[x_{T_1} \leftarrow e_{T_1}] : \tau_r}$$

- If  $y \neq x_{T_1}$  then there is no substitution that could be performed ( $y$  does not appear in the context) and so (trivially):

$$\Gamma @ \mathbf{r} [] \vdash y : \tau_r$$

(CONST) Similar to the (var) case when the variable is not substituted.

(APP) In the (app) rule, the context  $\Gamma$  is obtained as a tensor product  $\Gamma_1, \Gamma_2$ . Given  $\Gamma$  of length  $k$ , we assume that  $\Gamma_1$  has length  $l$  and  $\Gamma_2$  has length  $k - l$ . Now, given the following typing derivation using (app):

$$\frac{\begin{array}{l} \Gamma_1 @ \mathbf{s}_1 [x_{T_1} @ \langle s_1 \rangle | \dots | x_{T_l} @ \langle s_l \rangle] \vdash e_1 : \tau_1 \xrightarrow{t} \tau_2 \\ \Gamma_2 @ \mathbf{s}_2 [x_{T_{l+1}} @ \langle s_{l+1} \rangle | \dots | x_{T_k} @ \langle s_k \rangle] \vdash e_2 : \tau_1 \end{array}}{\Gamma_1, \Gamma_2 @ \mathbf{s}_1 \times (t \otimes \mathbf{s}_2) [x_{T_1} @ \langle s_1 \rangle | \dots | x_{T_l} @ \langle s_l \rangle, \\ x_{T_{l+1}} @ \langle t \otimes s_{l+1} \rangle | \dots | x_{T_k} @ \langle t \otimes s_k \rangle] \vdash e_1 e_2 : \tau_2}$$

Here, we use  $\mathbf{t} \otimes \mathbf{s}_2$  as a pointwise extension of  $\wedge$  that is additionally applied to holes such that  $\mathbf{t} \wedge - = -$ . That is, a hole in the context remains a hole and so the second part of the context is filled with  $\mathbf{t} \otimes \mathbf{s}_i$  for all  $i \in \{l+1 \dots k\}$ . Next, from the induction hypothesis, we have that:

$$\begin{aligned} \Gamma_1 @ \mathbf{s}_1 [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 | \dots | \Gamma_{Tl} @ \mathbf{s}_l \otimes \mathbf{T}_l] \vdash e_1 [\dots] : \tau_1 \xrightarrow{\mathbf{t}} \tau_2 \\ \Gamma_2 @ \mathbf{s}_2 [\Gamma_{Tl+1} @ \mathbf{s}_{l+1} \otimes \mathbf{T}_{l+1} | \dots | \Gamma_{Tk} @ \mathbf{s}_k \otimes \mathbf{T}_k] \vdash e_2 [\dots] : \tau_1 \end{aligned} \quad (*)$$

Now using the (*app*) rule in the first step and associativity of  $\otimes$  on the second part of the context in the second step:

$$\begin{array}{c} (*) \\ \hline \Gamma_1, \Gamma_2 @ \mathbf{s}_1 \times (\mathbf{t} \otimes \mathbf{s}_2) [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 | \dots | \Gamma_{Tl} @ \mathbf{s}_l \otimes \mathbf{T}_l, \\ \Gamma_{Tl+1} @ \mathbf{t} \otimes (\mathbf{s}_{l+1} \otimes \mathbf{T}_{l+1}) | \dots | \Gamma_{Tk} @ \mathbf{t} \otimes (\mathbf{s}_k \otimes \mathbf{T}_k)] \vdash (e_1 e_2) [\dots] : \tau_2 \\ \hline \Gamma_1, \Gamma_2 @ \mathbf{s}_1 \times (\mathbf{t} \otimes \mathbf{s}_2) [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 | \dots | \Gamma_{Tl} @ \mathbf{s}_l \otimes \mathbf{T}_l, \\ \Gamma_{Tl+1} @ (\mathbf{t} \otimes \mathbf{s}_{l+1}) \otimes \mathbf{T}_{l+1} | \dots | \Gamma_{Tk} @ (\mathbf{t} \otimes \mathbf{s}_k) \otimes \mathbf{T}_k] \vdash (e_1 e_2) [\dots] : \tau_2 \end{array}$$

(ABS) Without the loss of generality, we can assume that the bound variable is not one of the variables being substituted. Thus, the last variable (and the corresponding coeffect) are not holes. The typing derivation using (*abs*) then looks as follows:

$$\frac{\Gamma, x : \tau_1 @ \mathbf{r} \times \langle \mathbf{s} \rangle [x_{T1} : \tau_{T1} @ \langle \mathbf{s}_1 \rangle | \dots | x_{Tk} : \tau_{Tk} @ \langle \mathbf{s}_k \rangle] \vdash e : \tau_2}{\Gamma @ \mathbf{r} [x_{T1} : \tau_{T1} @ \langle \mathbf{s}_1 \rangle | \dots | x_{Tk} : \tau_{Tk} @ \langle \mathbf{s}_k \rangle] \vdash \lambda x. e : \tau_1 \xrightarrow{\mathbf{s}} \tau_2}$$

From the induction hypothesis, we have that:

$$\Gamma, x : \tau_1 @ \mathbf{r} \times \langle \mathbf{s} \rangle [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 | \dots | \Gamma_{Tk} @ \mathbf{s}_k \otimes \mathbf{T}_k] \vdash e [\dots] : \tau_2$$

Because the last position in the vector of variables is an actual variable rather than a hole, we just need to apply the (*abs*) rule:

$$\frac{\Gamma, x : \tau_1 @ \mathbf{r} \times \langle \mathbf{s} \rangle [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 | \dots | \Gamma_{Tk} @ \mathbf{s}_k \otimes \mathbf{T}_k] \vdash e [\dots] : \tau_2}{\Gamma @ \mathbf{r} [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 | \dots | \Gamma_{Tk} @ \mathbf{s}_k \otimes \mathbf{T}_k] \vdash \lambda x. e [\dots] : \tau_1 \xrightarrow{\mathbf{s}} \tau_2}$$

(LET) In the structural coeffect calculus let-binding can be viewed as a syntactic sugar for abstraction/application and so the case follows from (*abs*) and (*app*).

Compared to the similar proof for the flat coeffect calculus, the proof for the structural system requires fewer properties of the coeffect algebra. In particular, we only needed associativity of  $\otimes$  in the (*app*) rule the fact that *use* is a unit of  $\otimes$  in (*var*).

#### STRUCTURAL RULES

(CONTR) In case of contraction, we can assume that the two variables to be contracted (in the assumption) are not the variables that are being substituted. However, the resulting variable (in the conclusion) can be one of the variables being substituted for.

- Assuming that  $x \neq x_{Ti}$  for all  $i$ , the original derivation is:

$$\frac{\Gamma_1, y : \tau_1, z : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s}, \mathbf{t} \rangle \times \mathbf{q} [x_{T1} : \tau_{T1} @ \langle \mathbf{s}_1 \rangle | \dots | x_{Tk} : \tau_{Tk} @ \langle \mathbf{s}_k \rangle] \vdash e : \tau}{\Gamma_1, x : \tau_1, \Gamma_2 @ \mathbf{r} \times \langle \mathbf{s} \oplus \mathbf{t} \rangle \times \mathbf{q} [x_{T1} : \tau_{T1} @ \langle \mathbf{s}_1 \rangle | \dots | x_{Tk} : \tau_{Tk} @ \langle \mathbf{s}_k \rangle] \vdash e[z, y \leftarrow x] : \tau}$$

Applying (*contr*) to the induction hypothesis gives the required result:

$$\frac{\Gamma_1, y:\tau_1, z:\tau_1, \Gamma_2 @ \mathbf{r} \times \langle s, t \rangle \times \mathbf{q} \quad [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 \mid \dots \mid \Gamma_{Tk} @ \mathbf{s}_k \otimes \mathbf{T}_k] \vdash e[\dots] : \tau}{\Gamma_1, x:\tau_1, \Gamma_2 @ \mathbf{r} \times \langle s \oplus t \rangle \times \mathbf{q} \quad [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 \mid \dots \mid \Gamma_{Tk} @ \mathbf{s}_k \otimes \mathbf{T}_k] \vdash e[z, y \leftarrow x][\dots] : \tau}$$

- In the other case,  $x = x_{Ti}$  for some  $i$ . The original typing is:

$$\frac{\Gamma_1, -, -, \Gamma_2 @ \mathbf{r} \times \langle -, - \rangle \times \mathbf{q} \quad [x_{T1}:\tau_{T1} @ \langle \mathbf{s}_1 \rangle \mid \dots \mid y:\tau_1 @ \langle \mathbf{s} \rangle \mid z:\tau_1 @ \langle \mathbf{t} \rangle \mid \dots \mid x_{Tk}:\tau_{Tk} @ \langle \mathbf{s}_k \rangle] \vdash e : \tau}{\Gamma_1, -, \Gamma_2 @ \mathbf{r} \times \langle -, - \rangle \times \mathbf{q} \quad [x_{T1}:\tau_{T1} @ \langle \mathbf{s}_1 \rangle \mid \dots \mid x:\tau_1 @ \langle \mathbf{s} \oplus \mathbf{t} \rangle \mid \dots \mid x_{Tk}:\tau_{Tk} @ \langle \mathbf{s}_k \rangle] \vdash e[z, y \leftarrow x] : \tau}$$

Applying (*contr*) to the induction hypothesis gives the following result:

$$\frac{\Gamma_1, -, -, \Gamma_2 @ \mathbf{r} \times \langle -, - \rangle \times \mathbf{q} \quad [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 \mid \dots \mid \Gamma_{Ti} @ \mathbf{s} \otimes \mathbf{T}_k \mid \Gamma_{Ti} @ \mathbf{t} \otimes \mathbf{T}_k \mid \dots \mid \Gamma_{Tk} @ \mathbf{s}_k \otimes \mathbf{T}_k] \vdash e[\dots] : \tau}{\Gamma_1, -, \Gamma_2 @ \mathbf{r} \times \langle -, - \rangle \times \mathbf{q} \quad [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 \mid \dots \mid \Gamma_{Ti} @ (\mathbf{s} \otimes \mathbf{T}_k) \oplus (\mathbf{t} \otimes \mathbf{T}_k) \mid \dots \mid \Gamma_{Tk} @ \mathbf{s}_k \otimes \mathbf{T}_k] \vdash e[\dots] : \tau}$$

Here the  $\oplus$  operation represents a pointwise extension of  $\oplus$ . Thus for  $i^{th}$  substituted coeffect, we have  $(\mathbf{s} \otimes \mathbf{T}_i) \oplus (\mathbf{t} \otimes \mathbf{T}_i)$ . Using the distributivity law of structural coeffect algebra, we obtain the required structure:  $(\mathbf{s} \oplus \mathbf{t}) \otimes \mathbf{T}_i$ .

(*SUB*) As in the (*contr*) case, in the (*sub*) case we distinguish two situations. If the sub-coeffecting is applied to variable that is *not* being substituted for, then the case is easy (sub-coeffecting does not interact with substitution), so we only consider the case when  $x$  is one of the variables being substituted for:

$$\frac{\Gamma_1, -, \Gamma_2 @ \mathbf{r} \times \langle -, - \rangle \times \mathbf{q} \quad [x_{T1}:\tau_{T1} @ \langle \mathbf{s}_1 \rangle \mid \dots \mid x:\tau_1 @ \langle \mathbf{s}' \rangle \mid \dots \mid x_{Tk}:\tau_{Tk} @ \langle \mathbf{s}_k \rangle] \vdash e : \tau}{\Gamma_1, -, \Gamma_2 @ \mathbf{r} \times \langle -, - \rangle \times \mathbf{q} \quad [x_{T1}:\tau_{T1} @ \langle \mathbf{s}_1 \rangle \mid \dots \mid x:\tau_1 @ \langle \mathbf{s} \rangle \mid \dots \mid x_{Tk}:\tau_{Tk} @ \langle \mathbf{s}_k \rangle] \vdash e : \tau} \quad (\mathbf{s}' \leq \mathbf{s})$$

From the induction hypothesis, we have the following:

$$\Gamma_1, -, \Gamma_2 @ \mathbf{r} \times \langle -, - \rangle \times \mathbf{q} \quad [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 \mid \dots \mid \Gamma_{Ti} @ \mathbf{s}' \otimes \mathbf{T}_k \mid \dots \mid \Gamma_{Tk} @ \mathbf{s}_k \otimes \mathbf{T}_k] \vdash e : \tau$$

To complete the case, we need to apply (*sub*) repeatedly on each of the variables in  $\Gamma_{Ti}$ . For  $i^{th}$  variable  $x_i$ , the coeffect annotation is  $\mathbf{s}' \otimes \mathbf{T}_i$ . Using the fact that combining coeffects with  $\otimes$  preserves the ordering, we get that  $(\mathbf{s}' \otimes \mathbf{T}_i) \leq (\mathbf{s} \otimes \mathbf{T}_i)$  and so the conditions of (*sub*) are satisfied.

(*WEAK*) We again need to consider whether the removed variable is one of the variables that are being substituted for. If this is not the case, the proof is easy, so we only look at the other case:

$$\frac{\Gamma @ \mathbf{r} \quad [x_{T1}:\tau_{T1} @ \langle \mathbf{s}_1 \rangle \mid \dots \mid x_{Tk}:\tau_{Tk} @ \langle \mathbf{s}_k \rangle] \vdash e : \tau}{\Gamma, - @ \mathbf{r} \times \langle -, - \rangle \quad [x_{T1}:\tau_{T1} @ \langle \mathbf{s}_1 \rangle \mid \dots \mid x_{Tk}:\tau_{Tk} @ \langle \mathbf{s}_k \rangle \mid x:\tau_1 @ \mathbf{ign}] \vdash e : \tau}$$

Now, we use the induction hypothesis, apply the (*weak*) rule and use properties of the structural coefficient algebra:

$$\begin{array}{c}
 \frac{\Gamma @ \mathbf{r} [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 \mid \dots \mid \Gamma_{Tk-1} @ \mathbf{s}_{k-1} \otimes \mathbf{T}_{k-1}] \vdash e [\dots] : \tau}{\Gamma, - @ \mathbf{r} \times \langle - \rangle [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 \mid \dots \mid \Gamma_{Tk-1} @ \mathbf{s}_{k-1} \otimes \mathbf{T}_{k-1} \mid x : \tau_1 @ \mathbf{ign}] \vdash e [\dots] : \tau} \\
 \text{(sub)} \quad \frac{\Gamma, - @ \mathbf{r} \times \langle - \rangle [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 \mid \dots \mid \Gamma_{Tk-1} @ \mathbf{s}_{k-1} \otimes \mathbf{T}_{k-1} \mid x : \tau_1 @ \mathbf{ign} \otimes \mathbf{T}_k] \vdash e [\dots] : \tau}{\Gamma, - @ \mathbf{r} \times \langle - \rangle [\Gamma_{T1} @ \mathbf{s}_1 \otimes \mathbf{T}_1 \mid \dots \mid \Gamma_{Tk-1} @ \mathbf{s}_{k-1} \otimes \mathbf{T}_{k-1} \mid x : \tau_1 @ \mathbf{ign} \otimes \mathbf{T}_k] \vdash e [\dots] : \tau}
 \end{array}$$

The derivation first applies the standard (*weak*) rule and then uses sub-coeffecting rule and the property  $\mathbf{ign} \leq (\mathbf{ign} \otimes \mathbf{r})$  to obtain conclusion of the required form.

(EXCH) In the (*exch*) case, the property follows directly from the induction hypothesis. The required conclusion is obtained by applying (*exch*) repeatedly (as we now need to exchange not just two individual variables, but two contexts, possibly containing multiple variables).

□