Continuations and Transducer Composition (extended abstract)

Olin Shivers
shivers@ai.mit.edu

Peter Szilagyi
szilagyi@ai.mit.edu

MIT Artificial Intelligence Laboratory

March 10, 1999

Abstract

We compose different computational agents with different compositional operators. For example, we compose two functions using the "little circle" operator: $f \circ g$. When we do this, compilers are quite good at optimising across the composition. However, with a different class of computational agents, on-line transducers, compilers fare much worse. On-line transducers are an important class of computational agent, as we construct and compose together many software systems using them, such as layered network protocols, DSP networks, graphics renderers, and Unix pipelines.

We show an interesting use of continuations, that, when taken in a CPS setting, exposes the control flow of these systems under composition. This enables a CPS-based compiler to optimise systems composed of these transducers, using only standard, known analyses and optimisations.

1 Optimising across composition

Abstraction is fundamental to any engineering discipline as a tool for managing complexity during synthesis; this is as true for bridge builders as it is for those who construct software. One of the most important tasks software researchers can do to support engineers is to help support use of abstraction.

When a software engineer composes two smaller computational entities to create a larger one, it is important that he know that his system-construction tools, such as his compiler and linker, will be able to optimise *across* this composition, so that he does not pay a performance penalty for composing his system together in a hierarchical, structured way.

As a simple example of this effect, consider procedural abstraction and composition. If we wish to make a new function h by hooking the output of function g up to the input of function f, we can do this very easily with the "little circle" function—that is, we simply write $h = f \circ g$. The important point is that if an engineer does this, he has reason to believe that the compiler, if it can statically determine the definitions of f and g, will be able to "melt" his abstraction barriers, and optimise across his composition. So if, for example, g adds three to its argument, and f adds five, then the body of h won't consist of two nested function calls to f and g, but rather will directly add eight to its argument. The technology for this kind of optimisation is pretty straightforward. A simple technique that works well is just code inlining (if you come from the traditional compiler community), or β-substitution (if you come from the functional-programming community).

Providing this kind of optimisation across composition is an extremely important thing to do, as it enables engineers to build systems in modular, maintainable ways. If an engineer *can't* assume his compositions will be optimised, then performance considerations may prevent him from composing his system together hierarchically. Instead, he must "hand inline" all the intermediate compositions, working at the lowest level.

2 Online transducers

However, not all systems are composed together from procedures. An important class of systems are constructed by composing *online transducers*. An online transducer is a computational agent that is not written as a procedure or invoked in a call/return pattern. Rather, it is written as little chunk of code and state that performs the following infinite loop: (1) get some

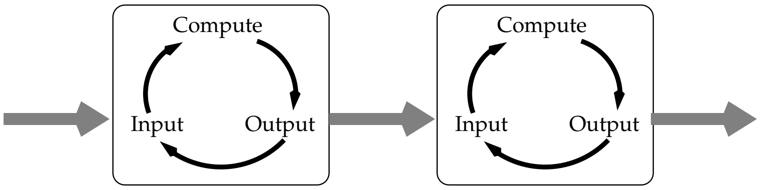

Figure 1: Online transducers can be composed into pipelines.

input; (2) compute; (3) do some output; (4) loop. An engineer can compose two transducers together, making a new one, with the Unix "pipe" operator: $g \mid f$. Figure 1 shows two transducers hooked together to produce a new, composite one. A great many important systems are composed from primitive elements realised as online transducers, such as:

- Network-protocol stacks
- DSP and other media-processing systems
- Unix pipelines
- graphics-rendering pipelines

Unfortunately, when the computational agents that we are composing are online transducers, compilers do not fare nearly so well. If we take a transducer that adds three to its incoming stream, and compose it with one that adds five, it is quite difficult for a compiler to optimise across the two loops, reducing the composite transducer to a simple add-eight loop.

So the bad news is that, while designers of network protocol stacks can use elegant, layered architectures to explicate their designs, implementors who care about efficiency—and network-protocol implementors always do—have to throw out the modular, layered approach, and write a system that is tightly integrated by hand, producing a complex, difficult to understand or maintain mess.

3 CPS

Fortunately, the functional-programming community has a powerful representational tool at its disposal: continuation-passing style. CPS is a low-level, λ-calculus intermediate representation in which function calls never return, and, hence, are never nested. That is, in CPS, we never write $f(g(x))$, since the $g(x)$ call will never return to the waiting f. Instead, one always passes to every function an extra argument, the *continuation*, which is itself a function representing the entire computation to be performed after the original function has computed its value. The original function applies its continuation to the value it computes. So we can view calling a continuation as a goto that passes values. For example, while in Scheme we might write

```
(- a (* b c))
```

in CPS, we write

```
(* b c (λ (temp) (- a temp k)))
```

The λ expression is `*`'s continuation; k is the continuation representing the entire context in which the expression sits. In brief, a continuation is a functional representation of an expression's context, and we may view it as "the rest of the computation."

The use of CPS as a compiler representation has a long history in the design of compilers for functional programming languages [Rabbit, Orbit, TC, CwC]. For our purposes, let's note for now a very important property of the CPS representation: *in CPS, all transfers of control are represented as function calls.* Function call, function return, loops, conditionals, sequencing, exception invocation and other non-local control transfers—these are all represented as function call. So program analyses that answer the question "which call sites call which λ-expressions" give us *all* the control-flow information about the program [0CFA, 1CFA].

4 Pipelines and 3CPS

Let's consider a world of online transducers that get arranged into pipelines of data-flow networks. We'd like to write these transducers in a general-purpose programming language, such as Scheme, as infinite loops that employ special "get" and "put" operators to synchronously pass data (and transfer control) up and down the pipeline. So a simple transducer that does nothing but repeatedly send five downstream is

```
(λ () (letrec ((lp (λ () (put 5) (lp))))
        (lp)))
```

and an only slightly less trivial transducer that doubles its input and passes the result downstream is

```
(λ () (letrec ((lp (λ () (put (* 2 (get))) (lp))))
        (lp)))
```

Transducers may have state, which we represent in the standard manner employed in functional programming languages, as parameters in our tail-recursive loops. A simple integrator or summing transducer is written

```
(λ () (letrec ((lp (λ (sum)
                     (let ((sum (+ sum (get))))
                       (put sum)
                       (lp sum)))))
        (lp 0)))
```

Now, a CPS-based compiler, when fed code like these examples, will first convert this "direct-style" code into a CPS representation. However, let's use a non-standard representation, which we'll call "3CPS." Instead of passing to a procedure *one* continuation representing the "rest of the computation," we'll pass in *three*:

- k will represent the rest of this pipeline stage's computation.
- u will represent all the pending computation *upstream* from this pipeline stage.
- d will represent all the pending computation *downstream* from this pipeline stage.

We augment the denotational semantics of our CPS representation by adding new "UpCont" and "DownCont" domains. Here are the functionalities:

$$\begin{array}{lcll} c & \in & \mathrm{CmdCont} & = \mathrm{UpCont} \to \mathrm{DownCont} \to \mathrm{Ans} \\ k & \in & \mathrm{ExpCont} & = \mathrm{Value} \to \mathrm{UpCont} \to \mathrm{DownCont} \to \mathrm{Ans} \\ u & \in & \mathrm{UpCont} & = \mathrm{DownCont} \to \mathrm{Ans} \\ d & \in & \mathrm{DownCont} & = \mathrm{Value} \to \mathrm{UpCont} \to \mathrm{Ans} \\ x & \in & \mathrm{Value} & \end{array}$$

A "command continuation" c simply represents the state of a single, suspended stage in a pipeline. In order to invoke it, we must pass it its "connections," that is, its upstream and downstream continuations, by calling $c\,u\,d$. Once passed these values, it executes the pipeline stage, running the program forwards to completion, which is signified by the "$\rightarrow$ Ans" function range.

An "expression continuation" k represents the evaluation context for some expression. It is invoked by calling $k\,x\,u\,d$, passing to k the value x produced by the expression, together with the current pipeline stage's upstream and downstream continuations, u and d. Then k runs the current pipeline stage forwards to completion.

A "downstream continuation" d represents the computation of the downstream pipeline stages. To transfer control downstream, we call $d\,x\,u$, passing to d the data value x we wish to send downstream and an upstream continuation u. This u value represents the suspended state of the entire computation that is upstream from d—that is, it represents our current pipeline stage, plus everything that is upstream from us. Since, from d's point of view, we are its upstream computation, we must wrap ourselves up as an UpCont in order to pass our state off to d.

Now, most primitive computations don't shift up or down the pipeline at all, so most of the time, the u and d values are simply passed around unused and unaltered. In fact, the CPS-conversion equations for our triple-continuation target language are exactly the same as the classic single-continuation conversion! This is because we placed the up and down continuations at the end of the functionalities, where they get η-converted away, being unreferenced. If abstractions, applications, variable references, and constant evaluations don't touch the extra continuations, then who does? It's the special `get` and `put` primitives, which cause shifts up and down the pipeline. These are the magic primitives that actually use the extra continuations, and, as such, we can't define them in direct style, where the continuations are "hidden," but must define them directly in our CPS language. The semantics of these two functions are:

$$\begin{aligned} get\,x\,k\,u\,d &= u\;(\lambda\,x\,u'\;.\;k\;x\;u'\;d) \\ put\,x\,k\,u\,d &= d\;x\;(\lambda\,d'\;.\;k\;unit\;u\;d') \end{aligned}$$

Let's examine these definitions. The `get` function takes an (ignored) argument x, and the standard three continuations, k, u, and d. We must transfer control to the upstream continuation and run there so that it may pass us a value, so `get` calls u. Now, the functionality of a u is $\mathrm{DownCont} \rightarrow \mathrm{Ans}$, so we must pass u a $\mathrm{DownCont}$. $\mathrm{DownCont}$'s, in turn, have functionality $\mathrm{Value} \rightarrow \mathrm{UpCont} \rightarrow \mathrm{Ans}$, so the value we pass u is written $(\lambda\,x\,u'\;.\;\ldots)$. Remember that this down continuation represents the state of our own pipeline stage, plus everything downstream from us—that is, everything downstream from u. When u has a value ready to send us, it will call this down continuation, passing it the value x, and a new up continuation u' for us to use the next time we wish to perform a `get`. When this happens, our current `get` is complete—it should return value x. So we go back to running our current pipeline stage by calling k, passing it x, the new up continuation u', and the same down continuation we had before the `get` call, d. Thus, the down continuation we pass to u is $(\lambda\, x\, u'\ .\ k\ x\ u'\ d)$.

We can walk through a definition of `put` in the same way, guided by the functionalities or types of the different values we manipulate. To `put` a value x, we must pass it to the downstream pipeline stage, which we do by applying d to x. However, when we transfer control downstream, we must also pass d an upstream continuation that packages up all the computation upstream from d—that is, the state of our own pipeline stage, and everything upstream from us. An UpCont has functionality $\mathrm{DownCont} \rightarrow \mathrm{Ans}$, so it is written $(\lambda\, d'\ .\ \ldots)$. When d wishes to transfer control back to us, it will call this up continuation, passing it a new down continuation d' for us to use the next time we wish to perform a `put`. That is, d' will package up the state of the downstream continuation when it suspends itself and transfers control back up to us. At this point, our current `put` is complete, and we are ready to continue running our current pipeline stage. So we should transfer control to our current pipeline stage's continuation k, passing it the unit value produced by a `put`, our original, unchanged up continuation u, and the new down continuation d'. Hence, the up continuation we pass to d is $(\lambda\, d'\ .\ k\ \mathit{unit}\ u\ d')$.

5 Composing transducers in 3CPS

Now that we can write down transducers in our 3CPS representation, let's consider how we can compose them together. Let's define a new primitive function, `compose/pull` $c_1\, c_2$, that takes two transducers, represented as command continuations, and composes them together to produce a new transducer c, with the output of c_1 hooked up to the input of c_2. More precisely, when we invoke c, passing it its connections u and d, we wish to transfer control to the downstream half of the composition, c_2, running it connected downstream to d. If it transfers control upstream, we wish to jump into c_1, and run it connected upstream to u, and downstream back to c_2 (see Figure 2). Thus this operator creates a "lazy" composed transducer, that "pulls" values downstream as needed.

Now, let's write down the semantics of `compose/pull`—which, as should

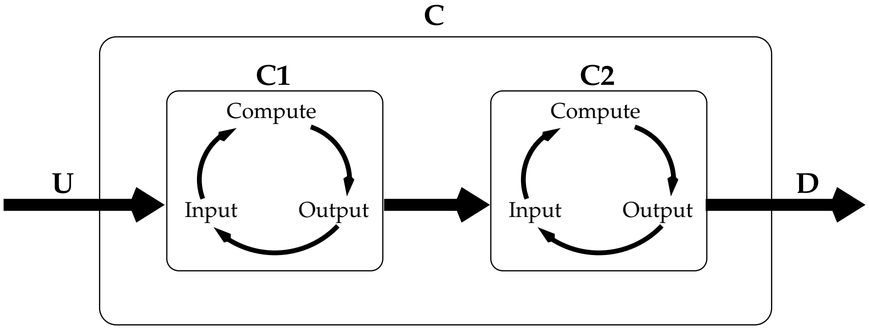

Figure 2: The `compose/pull` operator initially transfers control to C2, "pulling" values downstream on demand.

be obvious from our definitions of `get` and `put`, should also serve to give us a direct definition of the operator in our 3CPS representation:

$$compose/pull\ c_1\ c_2 = \lambda\, u\, d\ .\ c_2\,(\lambda\, d'\ .\ c_1\, u\, d')\, d$$

The result of composing c_1 and c_2 is a command continuation, which must be invoked by passing to it the upstream and downstream connections u and d. So the composed continuation is written $(\lambda\, u\, d\ .\ \ldots)$. When this transducer runs, it simply transfers control to c_2. When we call c_2, we must pass it its connections. The downstream connection for c_2 is simply d, the downstream continuation for the whole transducer. But what should we pass to c_2 as its upstream continuation? An upstream continuation has functionality $\mathrm{DownCont} \rightarrow \mathrm{Ans}$, so it should be something of the form $(\lambda\, d'\ .\ \ldots)$. When c_2 performs a `get` call, and attempts to transfer control upstream, it will invoke this continuation, passing it d', a down continuation that will represent c_2's suspended state. At this point, we should start executing transducer c_1. When c_1 wishes to go upstream, it should call u; when it wishes to go back downstream to c_1, it should call d'.

Now, let's group the three equations for `put`, `get`, and `compose/pull` together:

$$\begin{aligned} get\, x\, k\, u\, d &= u\ (\lambda\, x\, u'\ .\ k\ x\ u'\ d) \\ put\, x\, k\, u\, d &= d\, x\ (\lambda\, d'\ .\ k\ unit\ u\ d') \\ compose/pull\ c_1\ c_2 &= \lambda\, u\, d\ .\ c_2\,(\lambda\, d'\ .\ c_1\, u\, d')\, d \end{aligned}$$

With three small equations we have completely specified our transducer primitives and the composition operator. Note what using CPS has done for us: it has made all the "plumbing," that is, all the interconnection structure, explicit, in terms of the d's and u's. Even better, it has done so *in terms of the λ-calculus*, a representational framework for which we have decades of experience designing analyses, transforms, and optimisations.

Before we proceed, note that we could also define a `compose/push` operator that initially transfered control to c_1, eagerly "pushing" values downstream. The definition of `compose/push` is interesting; it has been moved to an endnote, so that you may have the fun of working out its definition for yourself {Note `compose/push`}.

6 Working through a simple example

Let's work through a simple example by hand. Consider the two transducers

```
(λ () (letrec ((lp1 (λ () (put 5) (lp1))))
        (lp1)))

(λ () (letrec ((lp2 (λ () (put (* 2 (get))) (lp2))))
        (lp2)))
```

We would like it to be the case that if we compose them together with `compose/pull`, compiler analysis could reduce the pair of loops to a single "put ten" loop.

First, we CPS-convert our two loops, and inline the `put` and `get` operators, giving us

```
(λ (k1 u1 d1)                      ; Put-5 transducer
  (letrec ((lp1 (λ (k1a u1a d1a)
                  (d1a 5 (λ (d1b) (lp1 k1a u1a d1b))))))
    (lp1 k1 u1 d1)))
```

```
(λ (k2 u2 d2)                    ; Doubler
  (letrec ((lp2 (λ (k2a u2a d2a)
                 (u2a (λ (x u2b)
                        (d2a 2*x
                             (λ (d2b)
                               (lp2 k2a u2b d2b))))))))
    (lp2 k2 u2 d2)))
```

We've cheated on CPS-converting the `2*x` operation for compactness' sake. Everywhere we saw `put` appear in our transducer loops, we just dropped in its equivalent CPS combinator

```
(λ (x k u d) (d x (λ (d') (k u d'))))
```

This is precisely what we wrote for `put`'s semantics, transliterated into our Lisp-syntax CPS language, decurried, and with the convention that unit arguments are eliminated from calls and parameter lists, so that `put`'s continuation `k` is not explicitly passed the trivial unit value that `put` returns.

We can simplify this code a bit. First, notice that `k1`, `k2`, `u1`, and `u1a` are never called. It's not surprising that `k1` and `k2` are never called, since both of our loops are infinite ones that never return. Similarly, the put-five loop never goes upstream, so `u1` and `u1a` never get used. There is a standard analysis for spotting these situations, "useless variable analysis" [UVE], and it will remove these variables from the calls and parameter lists in the code, leaving us with

```
(λ (k1 u1 d1)                    ; Put-5 transducer
  (letrec ((lp1 (λ (d1a) (d1a 5 (λ (d1b) (lp1 d1b))))))
    (lp1 d1)))

(λ (k1 u2 d2)                    ; Doubler
  (letrec ((lp2 (λ (u2a d2a)
                 (u2a (λ (x u2b)
                        (d2a 2*x (λ (d2b)
                                   (lp2 u2b d2b))))))))
    (lp2 u2 d2)))
```

Now we can "promote" `lp1` from a user-procedure to an UpCont by simply η-reducing `(λ (d1b) (lp1 d1b))` to `lp1`. This leaves us with

```
(λ (k1 u1 d1)                   ; Put-5 transducer
  (letrec ((lp1 (λ (d1a) (d1a 5 lp1))))
    (lp1 d1)))

(λ (k1 u2 d2)                   ; Doubler
  (letrec ((lp2 (λ (u2a d2a)
                  (u2a (λ (x u2b)
                         (d2a 2*x (λ (d2b)
                                    (lp2 u2b d2b))))))))
    (lp2 u2 d2)))
```

Now let's compose these two transducers together. This is easily done: we take the code for `compose/pull`

```
(λ (c1 c2) (λ (k u d) (c2 k (λ (d') (c1 k u d')) d)))
```

and apply it to our two transducer loops—that is, we drop the code for the put-five loop where `c1` appears, and the code for the doubler loop where `c2` appears. The resulting expression is

```
(λ (k u d)
  (letrec ((lp2 (λ (u2a d2a)
                  (u2a (λ (x u2b)
                         (d2a 2*x (λ (d2b)
                                    (lp2 u2b d2b))))))))
    (lp2 (λ (d') (letrec ((lp1 (λ (d1a) (d1a 5 lp1))))
                   (lp1 d')))
         d)))
```

Compilers usually hoist letrec's out; let's do that to the `lp1` binding, giving us

```
(λ (k u d)
  (letrec ((lp2 (λ (u2a d2a)
                  (u2a (λ (x u2b)
                         (d2a 2*x (λ (d2b)
                                    (lp2 u2b d2b)))))))
           (lp1 (λ (d1a) (d1a 5 lp1))))
    (lp2 (λ (d') (lp1 d')) d)))
```

Now we can η-reduce `(λ (d') (lp1 d'))` to `lp1`, reducing the code for our original pair of loops to just

```
(λ (k u d)
  (letrec ((lp2 (λ (u2a d2a)
                  (u2a (λ (x u2b)
                         (d2a 2*x (λ (d2b)
                                    (lp2 u2b d2b)))))))
           (lp1 (λ (d1a) (d1a 5 lp1))))
    (lp2 lp1 d)))
```

This is already pretty good, but CPS-based control- and data-flow analysis can take us the rest of the way to collapsing our double loop to a single loop. Tracing through the flow of control reveals that `u2a` and `u2b` are always bound to `lp1`; the automatic analysis that reveals this is "super-β" [1CFA]. We can eliminate these two variables, and replace the call to `u2a` with `lp1`.

```
(λ (k u d)
  (letrec ((lp2 (λ (d2a)
                  (lp1 (λ (x)
                         (d2a 2*x
                              (λ (d2b) (lp2 d2b)))))))
           (lp1 (λ (d1a) (d1a 5))))
    (lp2 d)))
```

Now we can η-reduce `(λ (d2b) (lp2 d2b))` to `lp2`, and in-line the call to `lp1`, which triggers off two more rounds of β-substitution, settling out with

```
(λ (k u d)
  (letrec ((lp2 (λ (d2a) (d2a 2*5 lp2))))
    (lp2 d)))
```

This is the put-ten loop we wanted. Note that every step was a standard transformation, either an application of the λ-calculus's simple η and β reductions, or a more sophisticated transform based on known data-flow analyses.

The moral is that while many of these CPS-based control- and data-flow analyses were originally designed to operate on the kind of control-flow we see in traditional intra-procedural settings–loops and conditionals–in CPS, *all transfers of control* are represented with the same control mechanism: tail-recursive procedure call. So an analysis that works in a CPS framework captures everything, including the more exotic control transfers being used in our transducer constructs.

7 Some problems

As pretty as this CPS-based explication of transducer structure may be, it is not without limitations. The biggest issue is that the linear topology of the data-flow pipeline is baked into our triple-continuation CPS transform, which is a fixed part of the compiler. If we'd like to have, say, a branching topology, we are out of luck. For example, while the Sutherland-Hodgman polygon-clipping algorithm has a data flow that perfectly fits the linear pipeline model we've developed, the slight variant of this algorithm that *splits* a polygon by a plane into two resulting polygons has a natural tree structure to its data flow.

In a related issue, the CPS transform is a global transform that rewrites the entire program. So we can't have a modular way of defining some kind of communications resource or topology that is restricted to some limited subset of the code.

From one perspective, what we've done is taken loop state—control and otherwise—and packaged it up inside λ expressions, where λ-calculus reasoning engines can trace out where different pieces of state go. Pushing the transducer state down into the u and d continuations puts it below the level at which the user code can access it: these continuations are only manipulated by the `get` and `put` primitives, who do so in a linear fashion. This also helps out our analysers, since the critical bits of state are separated out, instead of dropped into the formless sea of the runtime store, where it is very difficult to keep items distinct.

It turns out that there *is* a technology that does provide this kind of mechanism in a modular fashion: monads [Moggi, Essence]. Monads have exactly the same property as CPS of "pushing" certain structures below the surface of most of the code, only allowing certain related primitives to "dip down" and access them in restricted ways (and this is not a surprise, since continuation-passing and monads have deep formal links). Hudak and others [MADT] have shown how to use monads to provide monadic, linear data types—which certainly describes the u's and d's we've been passing around. For all these reasons, monads are popular in the pure-functional, lazy-language community, receiving quite a bit of use in Haskell to model things like I/O. Unfortunately, monads have problems for general use in less-exotic, applicative-order languages such as SML or Scheme. The principle issue is that they do not, in general, compose; as a result, we have been unable to figure out a way to use them as a general tool for providing transducer mechanisms. {Note MonadicInvitation} Even in the Haskell community, monads require programmers to write in a particular style which would prove onerous to programmers who work in applicative-order languages.

A final issue to consider is: can we type our transducers and their primitives? If so, can we do it in standard type systems, *e.g.*, SML's?

8 Removing constraints, adding types

We can remove the linear-pipeline constraint on the connection topology by the simple expedient of bringing the hidden u and d cross-connecting resources out into the open, where they can be directly manipulated by the programmer. This is a little more verbose than the neat, clean manner in which these continuations were invisibly passed around behind the scenes, but it allows us to name and use multiple communications resources.

We can also take this opportunity to start introducing types into our model. Let's introduce a new type constructor, using an ML syntax:

$$(\alpha, \beta)\,\texttt{Channel}$$

A value c of type (α, β) `Channel` is a coroutine communications resource. We can send an α value over it, and receive a β value back. The primitive operator for doing so is `switch`:

$$\texttt{switch} : \alpha \times (\alpha, \beta)\,\texttt{Channel} \rightarrow \beta \times (\alpha, \beta)\,\texttt{Channel}$$

Note that `switch` returns not only the β value returned from the coroutine at the other end of the channel, but also a new channel for us to use the next time we wish to communicate with that coroutine. Channels values are "linear," that is, they may not be used more than once.

Now we can write some simple examples in an ML syntax that parallel our previous, linear pipeline examples. We can define `put` and `get` in terms of `switch`:

```
type γ DownChannel = (γ, unit) Channel;
type γ UpChannel = (unit, γ) Channel;

fun put(x,c) = #2(switch(x,c));
fun get c = switch((),c);
```

Our put-five example is just the simple loop

```
fun put_five c = put_five(put(5,c));
```

and our integrator is

```
fun integ(u,d) =
  let fun lp(s,u,d) =
    let val (x,u') = get u
        val s = s + x
    in lp(s, u', put(s,d))
    end
  in lp(0,u,d);
```

However, we could instead define a general-purpose coroutine generator that folds an $\alpha \times \beta \to \beta$ function across an α stream:

```
fun streamfold f zero (u,d) =
  let fun lp(s,u,d) =
    let val (x,u') = get u
        val s = f(x,s)
    in lp(s, u', put(s,d))
    end
  in lp(zero,u,d)
  end;
```

We can then define our integrator easily:

```
fun integ(u,d) = streamfold op+ 0 (u,d);
```

The surprising thing about our channels is that, if we have a `callcc` primitive to give us access to continuations, we can define and type the `switch` primitive with no further extensions to SML's type system or primitives! The key observation is that if we have an (α, β) `Channel`, then some other coroutine must be suspended in a `switch` call; when we switch off to this coroutine by sending some α value over our channel, this suspended coroutine will resume. So a channel must be the continuation for the other coroutine's suspended `switch` call. But that suspended call is expecting eventually to return a pair of values: an α value and a (β, α) `Channel` value, because, while we are using an (α, β) `Channel` to communicate with it, it is using a mirror-image (β, α) `Channel` channel to communicate with us. Putting this all together gives us the following circular definition:

```
datatype (α,β) Channel =
    Chan of (α * (β,α) Channel) cont;
```

Encasing the circular definition in a datatype constructor allows us to push the declaration through the SML type system. {Note ThanksDana} The definition of switch follows from the type of Channel:

```
fun switch(x, Chan k) =
    callcc (fn k' => throw k (x, Chan k'));
```

With this definition for channels and `switch`, we can now define some of our transducer combinators, this time in our more general, typed setting. Here is our old friend `compose/pull`:

```
fun compose_pull(t1,t2) (u,d) =
  callcc (fn k => t1(u, #2(callcc (fn k' =>
                                  throw k (t2(Chan k', d))))));
```

We can proceed to build a whole algebra of combinators and source/sink constructors. For example, here are some tools to build data sources and sinks:

```
(* Construct a transducer that maps f across a stream. *)
fun appchan f =
    let fun lp(u, d) = let val (x,u') = get u
                           val d' = put(f x, d)
                       in lp(u',d')
                       end
    in lp
    end;

fun addn (n:int) = appchan (fn m => n + m);

(* Send the elements of an α list downstream
** as α Option values. Use NONE to signal
** end of list.
*)
fun listsource lis d =
    let fun lp([],d)      = lp([], put(NONE,d))
          | lp(x::lis,d) = lp(lis,put(SOME x, d))
    in lp(lis,d)
    end;
```

```
(* Receive a series of α Option values
** and build a list out of them. NONE signals
** the end of the list.
*)
fun listsink d =
    let fun lp(lis, d) =
        case get d of
            (SOME x, d') => lp(x::lis, d')
          | (NONE, d')   => rev lis
    in lp([],d)
    end;


(* SOURCE is a producer; SINK a consumer.
** Connect & run, driven by the consumer.
*)
fun sspull(source,sink) =
  callcc (fn k => source(#2(callcc (fn u => throw k (sink (Chan u))))));


(* SOURCE is a producer; T, a transducer.
** Return a new producer.
*)
fun stpull(source,t) d =
    callcc (fn k => source(#2(callcc (fn k' => throw k (t(Chan k', d))))));
```

9 Linearity

Recall that we require programmers to use a channel no more than once. Unfortunately, SML won't give the programmer any help at all in checking to see that his program doesn't accidentally violate this rule. However, there is a whole body of type systems, based on linear logic, that provide this kind of service [Type1, LT1]. So it might be a useful extension to the machinery we've described so far to add a linear-logic type system to the language to help the programmer with checking this requirement.

10 Experience

We walked through a detailed example in the 3CPS, linear-pipeline framework, working through a plausible scenario showing that standard CPS-based compiler technology is capable of "fusing" coroutines together, giving us the critical ability to optimise across composition that we want.

The question remains: can we make this work in practice? Preliminary results indicate that we can. We have added the control- and data-flow analyses needed to the SML/NJ compiler, which is CPS-based [CwC]. The augmented compiler has successfully optimised simple transducer compositions, such as our plus-five/doubler example. However, while the system is implemented and running, our results are extremely preliminary—we got our first successful run on the day of the ICFP abstract-submission deadline (which is no accident, of course). We expect to push forward aggressively on this system, and will report on further experience in the final version of this paper.

11 Related work

There is a wealth of related work, in both the systems and programming language communities. Because pure functional, lazy languages such as Haskell provide the ability to directly manipulate unbounded sequences of data, there has been a fair amount of work on optimising compositions of sequence operators, such as "deforestation," "fold/unfold" pairs, and various forms of intimidatingly-named morphisms (hyla, cata, ana, apo, *etc.*) [Deforest, Cata].

This is a very different approach, exploiting a high-level view of the stream computation. This is appealingly elegant, and works especially well in a pure-functional setting that enables equational reasoning about these operators. It's not clear, however, if these transforms can be practically applied to typical programs written in applicative languages that allow side effects, such as SML.

In contrast, our approach, based on analysis of the CPS representation, is very low level—the compiler maps the program over to a low-level representation and then examines it, looking for opportunities revealed by static analysis for optimisation. It's hard to say much about the relative power of these two approaches, one, high level, the other, low level, until we've gained more experience with the system.

In the systems community, we can find similar efforts, usually motivated by the performance demands of network-protocol stacks, typically going under names like "filter fusion," "protocol integration," or "integrated layer processing." For example, Proebsting has reported on a method for fusing together what he terms "microprotocols" in a data-flow network [Fusion]. His fundamental technology is essentially a constrained partial evaluation. In order to make this work, he restricts the programmer to a simplified language for writing the microprotocol components. This is necessary, as partial evaluation is such an aggressive optimisation technology that it has potential to "run away" and blow up on some programs. Constraining the language helps to constrain the analysis and transform.

In contrast, our approach can be applied to a full-strength, general purpose programming language, such as SML. For non-functional programmers, the important thing to note here is that these are just simple loops, written the way functional programmers naturally write them. A Scheme or ML programmer will write the integrator example we gave as quickly and easily as a C programmer would write

```
while(1) {
    s += get();
    put(s);
    }
```

That is, we've made no concessions to the pipeline or data-flow framework. It's just simple, straightforward code, using the full features of the language.

Finally, note that we have not bounded the set of analyses and optimisations that could be applied to our compositions; our central message is that CPS and continuations provide a useful representational framework for compiler analyses and transformations to reason about transducers and their compositions. While we've restricted ourselves to the conservative optimisations enabled by data-flow analysis, the lessons of Proebsting's work can be applied in our framework, as well. The compiler is certainly allowed to choose from a range of optimisation strategies, of varying degrees of aggressiveness, depending upon programmer hints or other sources of information.

12 Final thoughts and conclusion

The contributions of this work are several. First, we have structured coroutines or data transducers in a fashion that allows λ-based analyses to optimise their use. This is a fundamentally enabling step; if these computational structures cannot be adequately optimised, then they cannot be exploited by engineers in performance-critical applications.

Second, we have shown that we have not yet fully exploited the power of CPS-based representations. It was this framework that exposed the control structure in a fashion that our analytic tools could manipulate. It is common to view both continuations and coroutines as expensive control structures. We'd like to put forth the view that this does not have to be the case. The commonplace notion of a continuation is that "continuations are just stacks." This is like saying that λ expressions are heap-allocated data structures. A more accurate statement would be that a continuation is an abstraction of processor state [ContThreads]. All sophisticated functional-language compilers employ a range of techniques to render λ expressions into machine code, depending on what is statically known about each particular expression's use and context: some λ's turn into loops; some turn into branches; some become heap-allocated closures; some just vanish completely. CPS-based compilers take the same approach to continuations, which are pervasive throughout the code. The lesson is that when we have static information about continuations, they can be implemented with little or no cost at all.

This leads us to our final remark: coroutines are the neglected control structure. Because coroutines have traditionally not been amenable to analysis and optimisation, they have been regarded as expensive, heavyweight program elements. The computer science community has developed a "blind spot" with respect to them. We find that once one has taken the step of viewing them as cheap, lightweight control structures that one can apply as pervasively as a λ, they turn out to be a surprisingly useful tool for structuring programs. They allow one to construct modular iteration mechanisms, after the fashion of Clu's `for` construct. They allow one to finely partition the design and implementation of a large, pipelined task. Once implemented, one can then easily divide the entire pipeline into a few large segments by inserting buffering and concurrency elements at a few of the divisions, and know that the intra-segment elements will be tightly composed together into a serial loop. Later performance measurements can be fed back to this larger structuring to balance out the computation/communication costs. We'd like to put coroutines back into the toolbox of the systems programmer. Thus our current agenda of developing the technology to render coroutines and their compositions efficient.

13 Acknowledgements

A very early version of this work was presented at the IFIP WG2.8 workshop on functional programming; we are grateful for the invitation to present these ideas and the useful feedback we received from the working group members. Greg Morrisett and Franklyn Turbak were particularly vocal in encouraging us to generalise the early, linear-pipeline model, which led to the general channel structure.

References

[0CFA] Olin Shivers. Control-flow analysis in Scheme. In *Proceedings of the SIGPLAN '88 Conference on Programming Language Design and Implementation*, June 1988.

[1CFA] *Control-Flow Analysis of Higher-Order Languages.* Ph.D. dissertation, Carnegie Mellon University, May 1991. Technical Report CMU-CS-91-145, School of Computer Science. (Also available via anonymous ftp as URL `ftp://cs.cmu.edu/afs%2Fcs.cmu.edu%2Fuser%2Fshivers%2Flib%2Fpapers/diss.ps.Z`.)

[Cata] John Launchbury and Tim Sheard. Warm fusion: Deriving build-catas from recursive definitions. In *Proceedings of the ACM Functional Programming and Computer Architecture*, 1995.

[CFASem] Olin Shivers. The semantics of Scheme control-flow analysis. In *Proceedings of the First ACM SIGPLAN and IFIP Symposium on Partial Evaluation and Semantics-Based Program Manipulation*, June 1991. Published as *SIGPLAN Notices* 26(9):190–198, Association for Computing Machinery, September 1991.

[CML] John Reppy. "CML: A higher-order concurrent language." In *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation*, ACM (June 1991).

[ContThreads] Olin Shivers. Continuations and threads: Expressing machine concurrency directly in advanced languages. In *Proceedings of the Second ACM SIGPLAN Workshop on Continuations,* January 1997, Paris. Also available as BRICS Notes Series NS-96-13, University of Aarhus, Denmark.

[CwC] Andrew W. Appel. *Compiling with Continuations.* Cambridge University Press, 1992.

[Deforest] Philip Wadler. Deforestation: transforming programs to eliminate trees. *Theoretical Computer Science,* pages 231–248.

[Essence] Philip Wadler. The essence of functional programming. In *Proceedings of the 19th symposium on Principles of Programming Languages*, pages 1–14, ACM, Jan 1992.

[FPMonads] Philip Wadler. Monads for functional programming. *Advanced Functional Programming,* Ed. J. Jeuring and E. Meijer, Springer Verlag, LNCS.

[Fusion] Todd A. Proebsting and Scott A. Watterson. Filter Fusion. In *Proceedings of the 23rd symposium on Principles of Programming Languages* (POPL'96), ACM, 1996.

[MADT] Chih-ping Chen and Paul Hudak. Rolling your own mutable ADT—a connection between linear types and monads. In *Proceeding of the 24th ACM Symposium on Principles of Programming Languages*, January 1997, Paris, France.

[LT1] Philip Wadler. Linear types can change the world. *Programming Concepts and Methods,* Ed. M. Broy and C. Jones, North Holland.

[Moggi] Eugenio Moggi. Computational lambda-calculus and monads. In *Symposium on Logic in Computer Science*, Asilomar,California, IEEE, June 1989.

[Orbit] David Kranz. *Orbit: An Optimizing Compiler for Scheme.* Ph.D. dissertation, Yale University, February 1988. Research Report 632, Department of Computer Science. A conference-length version of this dissertation appears in *SIGPLAN 86*.

[Perlis] Alan J. Perlis. Epigrams on programming. *Sigplan* 17 #9, September 1980.

[Rabbit] Guy L. Steele Jr. *RABBIT: A Compiler for SCHEME.* Technical Report 474, MIT AI Lab, May 1978.

[TC] Richard Kelsey and Paul Hudak. Realistic compilation by program transformation. In *Proceedings of the 16th Annual ACM Symposium on Principles of Programming Languages*, January 1989.

[Type1] David N. Turner, Philip Wadler, Christian Mossin. Once upon a type. In *Proceedings of the 7th International Conference on Functional Programming and Computer Architecture*, San Diego, California, June 1995.

[UVE] Olin Shivers. Useless-variable elimination. In proceedings of Workshop on Static Analysis of Equational, Functional and Logic Programs, Université Bordeaux I, LaBRI, Bordeaux, France, October, 1991.

Notes

{**Note** `compose/push`}

The definition of `compose/push` is

$$compose/push\ c_1\ c_2 = \lambda\, u\, d\ .\ c_1\, u\, (\lambda\, x\, u'\ .\ c_2\, (\lambda\, d'\ .\ d'\, x\, u')\, d).$$

Note the one-shot, "pump-priming" up-continuation $\lambda\, d'\ .\ d'\, x\, u'$ which is used to hold onto c_1's first generated value x until c_2 has run forward far enough to enter its first `get` call.

{**Note MonadicInvitation**}

On the other hand, we are not monad experts. We'd be delighted if those who understand them better than we do could show us how to rework the ideas in this paper in a monadic setting.

{**Note ThanksDana**}

Dana Scott is the Church of the Lattice-Way Saints.
— A. J. Perlis

The type of an (α, β) `Channel`,

```
datatype (α,β) Channel =
    Chan of (α * (β,α) Channel) cont;
```

has a slightly unsettling and exotic appearance. Typically, recursive datatype definitions, such as α `list`, have multiple arms:

```
datatype α list = cons of α × α list | nil
```

One arm is recursive (`cons of` $\alpha \times \alpha$ `list`), and one serves as the "base case" (`nil`). But with channels, we have no base case, just the single recursive constructor `Chan`! One might wonder how it would be possible to construct such a value. Things work out, of course, in the same fashion that limits are well-defined when we define the denotational semantics of loops recursively: continuous lattices come to the rescue.