Shallow Embedding of DSLs via Online Partial Evaluation

Roland Leißa Klaas Boesche Sebastian Hack
Saarland University, Germany
{leissa, boesche, hack}@cs.uni-saarland.de

Richard Membarth Philipp Slusallek
German Research Center for Artificial Intelligence, Germany
{richard.membarth, philipp.slusallek}@dfki.de

Abstract
This paper investigates shallow embedding of DSLs by means of online partial evaluation. To this end, we present a novel online partial evaluator for continuation-passing style languages. We argue that it has, in contrast to prior work, a predictable termination policy that works well in practice. We present our approach formally using a continuation-passing variant of PCF and prove its termination properties. We evaluate our technique experimentally in the field of visual and high-performance computing and show that our evaluator produces highly specialized and efficient code for CPUs as well as GPUs that matches the performance of hand-tuned expert code.

Categories and Subject Descriptors F.3.2 [Semantics of Programming Languages]: Partial evaluation

Keywords DSL Embedding, Partial Evaluation, Continuation-Passing Style

1. Introduction
To achieve optimum performance, programs have to be transformed in a way that is beyond the scope of ordinary compiler optimizations. These transformations have two goals: First, exploit domain knowledge that is lost in the implementation and not accessible to the compiler. Second, utilize features of the target hardware architecture to improve performance (vectorization, memory hierarchy, etc.).

One way of achieving this performance is to create a domain-specific language (DSL) that provides language constructs to express domain knowledge, and a compiler that leverages this knowledge to produce highly optimized code for a specific architecture. A popular approach to implement a DSL is to embed the DSL into a host language $H$. One typically distinguishes between two different styles of embedding [15]:

Deep. The DSL program is represented as a data structure in the host program.

Consider Figure 1a. In the host language $H$, the programmer writes a program $\text{pgen}$ that constructs the embedded program $e_{\text{spec}}$. Because $\text{pgen}$ constructs the embedded program, it can also construct a version of the embedded program that is partially evaluated with respect to the inputs $s$. Then, an optimizer $\text{opt}$ transforms $e_{\text{spec}}$ to $e_{\text{opt}}$ which is finally emitted to target code by compile. Note that $\text{opt}$ as well as compile are written in $H$.

Deep embeddings allow for powerful, domain-specific optimizations [7, 31, 39] because the embedded program is available as a data structure. For the same reason, deep embeddings can accommodate any embedded language. In terms of programming experience, one drawback of deep embeddings is that the programmer actually writes a program generator instead of a program. Modern deep embedding frameworks alleviate this problem by “virtualizing” the host language [7]: Overloading reinterprets a part of the language constructs to not perform the actual computation but to construct a representation of that computation. This virtualization is often not entirely faithful and compromises the illusion of writing the embedded program in $H$ in several ways: First, the overloading is not powerful enough to hide this construction entirely and leak implementation details of the embedding into the host language [27]. Second, the host language can usually not be virtualized entirely. Third, to reason about the embedded program (the result of $\text{pgen}$), the programmer ultimately has to understand how the generator works.

Shallow. The DSL constructs are defined by implementing their semantics in the host language directly.

Consider Figure 1b. The programmer directly writes the embedded program $e$ in language $H$. To perform partial evaluation, shallow embedding needs a partial evaluator $\text{pgen}$ to be available in the compiler of $H$. Both functions, $\text{opt}$ and compile are part of the compiler of $H$.

Like virtualization but unlike deep embedding, shallow embedding can accommodate only one embedded language $H$ itself. However, unlike virtualization, shallow embedding uses the entire language $H$. In contrast to deep embedding, shallow embedding cannot manipulate the embedded program because it is not available as a data structure. However, shallow embedding does not suffer from the programming experience problems that deep embedding does, because the programmer writes the embedded program directly and not a program generator. Nevertheless, only a few shallowly embedded high-performance DSLs (e.g. HIP$^+$ [35] and SYCL [30]) exist. One reason is that, if no partial evaluator is available for $H$ (which is usually the case), shallow embedding involves the unpleasant task of modifying an existing compiler.

1.1 Our Approach
In this paper, we present the continuation-passing style (CPS)-based language Impala together with a novel online partial evaluator. Impala enables shallow embedding without having to modify its compiler (usually). Embedding a DSL into Impala typically means that the domain-specific constructs are implemented as (higher-order) functions. These implementations essentially constitute a tagless interpreter [6]. We obtain the compiled DSL program by partially evaluating this interpreter with that program. The following
function constitutes such an “interpreter” to iterate several loop bodies in a fused manner:

```haskell
fn fused_iterate(iterate: fn(fn(int)) -> O, bodies: [fn(int)]) -> O {
    for i in iterate() {
        for body in #each(bodies) {
            body(i)
        }
    }
}
```

Specializing a call (via \texttt{\#}) to \texttt{fused_iterate} with a function range and an array of bodies yields the desired fused loop\footnote{Impala’s syntax borrows from Rust. \texttt{x} \texttt{\&} \texttt{\&} \texttt{x} \texttt{\&} \texttt{x}\texttt{\&} \texttt{x} means \texttt{dx}.}:

```haskell
@fused_iterate(body), range(a, b), bodies()
```

To generate hardware-specific code, the Impala compiler exposes hardware-specific paradigms through higher-order functions. These functions can be used to implement a certain DSL construct with respect to a specific kind of hardware. This way, we enable hardware-specific code generation without having to dig into Impala’s compiler. In the example above, we could replace the argument range with the compiler-known function \texttt{vectorize} that strip-mines the loop and vectorizes the resulting innermost loop:

```haskell
@fused_iterate(body), vectorize(length, a, b, bodies)
```

Using partial evaluation (PE), we are limited to optimizations that can be expressed by \textit{specializing} code. Optimizations that analyze and rewrite programs are not possible without modifying Impala’s compiler. We argue that many of such optimizations can be expressed by proper abstractions in the style of \texttt{fused_iterate}. Other optimizations that cannot be expressed this way have to be implemented in Impala’s compiler which has a sufficiently high-level IR [32] to facilitate this. For the DSLs we present in this paper, we did not have to modify Impala’s compiler.

### 1.2 Continuations

To enable embedded DSLs to use non-trivial control flow (see Section 5.1 for an example), Impala features continuations as first-class citizens. Impala represents all control flow (including functions) as continuations. The following example shows a simple \textit{for}-loop with unstructured control flow and its internal representation using CPS:

```haskell
for i in range(a, b) {
    let break: fn() = || "next"/
    if i == 23
        continue()
    else if i == 42
        break()
    else /*...*/
        /*...*/
    } "next"/
}
```

One important aspect of a partial evaluator is to determine where to resume partial evaluation after skipping code under a dynamic condition. Assume the partial evaluator wants to evaluate the call to \texttt{range} above. Furthermore, assume that \texttt{a} and \texttt{b} are dynamic, i.e. their values are not known at partial evaluation time. Because the partial evaluator cannot evaluate the condition of the \texttt{if} \texttt{i} == \texttt{23}, it should skip the call to \texttt{range}. However, at which continuation shall partial evaluation be resumed?

In a direct-style language a suitable resume point can be derived from the syntactic structure of the language, i.e. the statement after the skipped statement. In CPS, the “\textit{next instruction}” is passed as a parameter and can be an arbitrary continuation (closure) that might not be syntactically "close". In the example above, the resume point is the \texttt{break} continuation passed to \texttt{range}. In this paper, we extend the notion of a post-dominator (a well-known concept in first-order control flow) to higher-order programs to derive suitable resume points.

### 1.3 Contributions

In summary, this paper makes the following contributions:

- We present a novel, pragmatic algorithm for online PE. We discuss our algorithm on a CPS-based variant of Plotkin’s Programming

```plaintext
\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Partial evaluation and metaprogramming.}
\end{figure}
```

Computable Functions (PCF) that captures the semantics of full as well as partial evaluation. We formally describe a termination property of our partial evaluator and prove our partial evaluator is correct in that respect (Section 3).

- A crucial aspect of our PE algorithm is the computation of post-dominators in higher-order programs to designate resume points for PE. We present a novel control flow analysis (CFA) that locally computes useful post-dominators during PE (Section 4).

- We show how mapping to different hardware accelerators can be nicely expressed by higher-order functions. Our approach allows to weave in platform-specific mapping strategies such as executing code on a GPU or vectorizing code for a CPU by compiler-known higher-order functions. We demonstrate that our PE approach enables an efficient shallow embedding of high-performance DSLs for visual and high-performance computing in Impala (Section 5).

## 2. Background and Related Work

As running example to discuss prior work, we review how to specialize the power function to its exponent.

### 2.1 Partial Evaluation

In this paper we advocate online partial evaluation [9, 43, 44]. We directly specialize the source program \( P \) (Figure 3d) to the specialized or residual program \( P_{\text{res}} \) (Figure 2b). This corresponds to the first Futamura [14] projection: Specializing an interpreter \( P \) to an input program produces a compiled version of that program. The specializer is often called \textit{mix} in literature.

Specializing the specializer itself yields a compiler generator (\textit{cogen}): a tool that converts an interpreter to a compiler (the third Futamura projection). For a long time it was unclear how \textit{cogen} actually looks like and generating \textit{cogen} via \textit{mix} requires \textit{mix} to be self-applicable which turned out to be hard in practice. Building a self-applicable evaluator becomes easier when separating the input program into static and dynamic parts: the \textit{binding time}. In our example, a binding-time analysis (BTA) [23, 26] infers that everything which depends on \( x \) must stay dynamic and annotates the program accordingly (Figure 2a). Then, the specializer (\textit{mix}<sub>\textit{cogen}</sub>) runs on that annotated program as opposed to directly running the specializer (\textit{mix}<sub>\textit{mix}</sub>) on \( P \). For this reason, this technique is called offline partial evaluation [3, 21, 23]. Glück [16] discusses a self-applicable online partial evaluator.

Birkedal and Welinder [2] discovered that hand-writing \textit{cogen} is actually not more difficult than writing \textit{mix}. In particular, a handwritten \textit{cogen} does not require a bootstrapping process. Thus, \textit{cogen} does not necessarily need to be written in the same language as \( P \). Given the annotated program \( P_{\text{ann}} \), \textit{cogen} produces its generating extension \( P_{\text{gen}} \) (Figure 2c). All static parts of \( P_{\text{ann}} \) are copied over to \( P_{\text{gen}} \). Dynamic parts are converted into a program that generates the specialized program \( P_{\text{res}} \). Thus, \( P_{\text{gen}} \) is parametric in \( P_{\text{ann}} \).

```plaintext
\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Online partial evaluation and \textit{mix}<sub>\textit{cogen}</sub>.
\end{figure}
```
Figure 3. Synchronized the power function to its exponent with meta programming and partial evaluation.

let rec pow x n =
  if n = 0 then 1
  else if n mod 2 = 0 then
    let r = pow (x * x) (n / 2) in r * r
  else
    x * pow (x, n - 1)

let Pspec =
  fun x ->
  return [pow(x, 4)]

(a) MetaOCaml  (b) Terra  (c) Scala/LMS  (d) Impula

static input. Running $P_{pev}$ with a specific static input generates a program $P_{pec}$, which is parametric in $P_{pec}$'s dynamic input. For example, invoking $P_{gen}(4)$ generates $P_{pec}$ (Figure 2b). From a different point of view, cogent transforms the one-stage program $P$ into a two-stage program.

The cogent approach is prone to the overapproximation of the static BTA. It has to produce one $P_{pec}$ that must work for every static input of $P$. Thus, the BTA must always be $d$ dynamic in the following example:

```haskell
fn f(a, b, d) { if a == b (d = 42) } /*next*/
```

An online evaluator, however, which specializes this function for $a$ and $b$ can exploit the case where $a == b$ and set $d$ to 42.

2.2 Metaprogramming

Metaprogramming allows the programmer to write a program that generates another program. In other words, the programmer manually implements the generating extension $P_{pec}$. For this reason, metaprograms conceptually look like the pseudocode in Figure 2c and the programmer can explicitly stage a DSL interpreter [10]. Many projects implement $P_{pec}$ in a scripting language like Python. The script is invoked at build-time to generate $P_{pec}$—usually in a low-level language like C. As such scripts simply splice strings, the residual program may be ill-typed. Other approaches like C++ metaprogramming, Terra [13] (Figure 3b), quasi-quotiation and macros in Scheme/Lisp or Racket [49] increase programmer productivity by incorporating metaprogramming facilities into the language, but still may construct an ill-typed residual program. MetaML [43] and MetaOCaml (Figure 3a) on the other hand, guarantee well-typedness of the residual program if the metaprogram is well-typed. With PE, well-typedness of the residual program comes for free because type checking is independently performed prior to specialization. All these metaprogramming approaches have in common that the stage is a feature of the syntax and, hence, in contrast to PE, it is not possible to write a function which is polymorphic in the binding time of its parameters. Dynamic staging [11] tackles this problem by introducing the stage as a first-class citizen to the language at the cost of an unsound type system.

2.3 DSL Embedding

Carette et al. [6] and Hudak [19] lay the foundation of embedding a typed language by ordinary functions instead of object terms. Hofer et al. [17] picked up this idea and carried it to the Scala world while emphasizing modularity and the ability to abstract over and compose semantic aspects. Rumpf and Odersky [41] coined the term lightweight modular staging (LMS) and paved the way for performance-oriented embedded DSLs [7] like OptiML [46] or Liszt [12]. LMS does not rely on explicit staging capabilities of the host language Scala. Instead, executing the host program constructs a second domain-specific program representation like Delite [5] (deep embedding). Values of type $\tau$ are wrapped into a type operator $\text{Rep}$ that represents values which are computed in a deferred stage. Lancet [42] is an online partial evaluator for Java bytecode that serves as a front-end for LMS, as an alternative to explicit programing with $\text{Rep}$ types. Array Building Blocks [33] and Halide [39] leverage a similar staging mechanism to construct the actual program representation with C++ as host language. HiPAc [35] and SYCL [30] on the other hand, are shallowly embedded DSLs in C++. They rely on compiler plug-ins which directly manipulate the program representation to achieve performance.

Comparable to other explicit metaprogramming techniques, LMS essentially requires the programmer to write the generating extension. However, via overloading and type inference the staged program is somewhere between $P$ and $P_{pec}$ (Figure 3c). As the stage is encoded in the type system, Scala’s type inference works akin to a local BTA [37]. For example, $n \% 2$ is of type Int. Thus, the expression is executed when the host program runs. But since $r$ is of type $\text{Rep}[\text{Int}]$, executing $r*r$ results in a residual program containing a multiplication. The implementation of $\text{cogen}$ lies in LMS’ library which implements $a*b$ for $\text{Rep}[\text{Int}]$. The downside of this approach is that for data types unknown to LMS, the programmer must implement these overloads (“cogen for these types”) himself. To some extent, this limitation can be alleviated via parametric polymorphism [36] at the cost of introducing type variables for each desired staging combination:

```haskell
let f(x, y) =
  if x < y (else return x)
  fn f(x, y) =
  return [pow(x, 4)]

(a) MetaOCaml  (b) Terra  (c) Scala/LMS  (d) Impula
```

Jovanović et al. [27] present a technique based on Scala macros to generate a deep embedding from a shallow one. Our approach on the other hand uses shallow embedding and PE to achieve the performance of a deeply embedded DSL. Finally, LMS may suffer from induced divergence while our PE technique may only induce divergence when run-annotating recursive calls (see below and Section 3).

2.4 Divergence in Partial Evaluation

Katz and Weise [29] distinguish three classes of divergence that may occur during PE:

True divergence: If the full evaluation of the program does not terminate for some inputs, PE might also not terminate.

Hidden divergence: A program may contain unreachable code that is divergent. Partially evaluating this divergent code may cause the partial evaluator to diverge.

Induced divergence: A program may contain unreachable code that is divergent. Partially evaluating this divergent code may cause the partial evaluator to diverge.
call to the cached version instead. This technique does not prevent the counting loop problem.

LMS has special behavior for a while loop: If the termination condition is of type Boolean, the loop will be run when the host program runs. If it is of type Rep[Boolean], LMS will construct a residual loop. This approach only works since Scala does not provide a continue. LMS also leverages the aforementioned memoization technique for recursive calls. This has the effect that a counting while-loop with a dynamic conditional terminates when the host program runs, whereas a recursive implementation with an 1nt counter and a Rep[Int] bound diverges.

Both Similix [25] and Schism [8] are offline evaluators using BTA. These evaluators will not evaluate a cycle, if the condition which breaks the cycle remains dynamic. On the one hand, this is slightly more aggressive than our approach because our approach will also jump over an acyclic, dynamic conditional. On the other hand, both evaluators suffer from the inherent imprecisions caused by the BTA (see above). Furthermore, both evaluators depend on a CFA [58] which leads to further imprecisions: As argued in Section 4, in our setting an on-the-fly CFA is more precise than a CFA which runs once beforehand because evaluating the program brings full context-sensitivity for free. Lastly, we argue that a single rule—dynamic branches are skipped—is easier to understand for the programmer.

Other more complex termination heuristics, like monitoring the argument sizes of recursive calls have been applied in the past [22, 24]. We consider such heuristics hard to understand and account for by the programmer.

3. Partial Evaluation

In this section we first formally discuss our PE technique by studying the lambda language LMS (Figure 4). Then, we discuss how to embed and guide the partial evaluator from within a program and how this affects termination.

3.1 The CPS-Based, Simply-Typed Lambda Calculus

The syntax of LMS is similar to the simply-typed lambda calculus with an additional fixpoint operator (called letrec) to allow recursion. Furthermore, LMS uses CPS for function abstractions and applications. Thus, LMS is a CPS version of Plotkin’s PCF [38]. CPS introduces mainly two peculiarities compared to Plotkin’s original PCF:

1. As functions do not return in CPS, we do not allow functions to return a value. Instead, the actual function is formed by a body.
2. For the same reason, we cannot curry functions. Hence, we allow arbitrarily long parameter lists.

For the sake of presentation, we restrict the arithmetic to integer literals and addition. An if body tests the first expression—the condition—for zero. If the test yields true, evaluation will progress with the second expression as continuation or with the third one otherwise. Program execution ends with result e upon reaching an exit body exit e. A skip and skipping body is considered as expanded syntax, which only appears as intermediate results during evaluation. It cannot be used by the programmer directly.

We denote the expression language (using e as start symbol) by L. We sometimes refer to a body as program if we want to stress that a body may contain many sub-bodies. The syntactic structure of a language induces a syntax tree. We write a ≤ b if we require a to be a subtree of b. Parameters and identifiers bound in letrecs range over x. We require all names in the program to be unique in order to circumvent name capture in the rules. We use the common notation π to denote a list a₁,...,aₙ.

In examples we often elide type annotations as the reader can easily infer them from their surrounding context. As syntactic sugar, we use where to bind non-recursive functions and wherefunc to bind recursive ones. Finally, we sometimes use additional features in LMS examples like boolean types or more sophisticated branch constructs.

3.1.1 Typing

As functions do not return, typing rules checking bodies do not yield a type. A function type Fn(E) does not include a return type for the same reason. Apart from that, rules are standard.

Definition 1 (Expression Normal Form). An expression e is in normal form iff \( e = \) e. We write \( \text{nf} e \) if we require e to be in normal form.

Definition 2 (Exits). Let \( \text{exists}(b) := \{ \text{exit} \mid \exists e \leq b \} \) be the set of exits in the body b.

Definition 3 (Well-Typedness). We call a body b well-typed under \( \Gamma \) iff \( \Gamma \vdash b \) holds.

Definition 4 (Constant). We call an expression e a constant iff e is a normal form and \( + e : t \) for some type t.

Remark. These are all literals and functions without free variables.

Definition 5 (Valid Configuration). Let \( f := \lambda \text{letrec} \) b be a function constant. We call an argument list \( \pi \) of constants a valid configuration for \( f \) iff the application \( f(\pi) \) is well-typed. We denote \( C(f) \) as the set of all valid configurations for f.

3.1.2 Semantics

In the classic lambda calculus each argument to a function evaluates to a constant. Then, the function is substituted by its body while replacing all occurrences of the function’s parameters with its corresponding arguments. During PE however, arguments may not necessarily be constants.

Expression semantics. The function \( [e] \) evaluates an expression. If two literals occur in an addition they will be folded (operator \( + \) denotes the arithmetic addition as opposed to the syntactic terminal \( + \)). Other additions are evaluated by recursively applying evaluation. Other expressions yield identity. For example, the expression \( x + (1 + 2) \) reduces to \( x + 3 \).

Lemma 1 (Strong Normalization Property of Expressions). Evaluating an expression \([e] \) is strongly normalizing, that is, every \([e]\) eventually terminates with an expression in normal form.

Proof. Evaluation of variables, literals and functions is strongly normalizing by definition. By induction we show that addition is strongly normalizing, too. □

Body semantics. The semantics of LMS works as partial as well as for full evaluation. In contrast to expressions, body semantics is not strongly normalizing due to letrecs. On this account we use a small-step semantics for body evaluation rules of the form \( b \Rightarrow b' \). We read: A body b evaluates in one step to body b'. We write \( b \Rightarrow^{*} b' \) in order to specify that b evaluates to b' in n steps, and \( b \Rightarrow^{*} b' \) to specify that b evaluates in finitely many steps to b'.

Definition 6 (Body Normal Form). A body b is in normal form iff \( 3b' : b \Rightarrow b' \). We write \( \text{nf} b \) if we require b to be in normal form.

A function application first reduces its callee e, and arguments to normal form (E-App). The requirement \( \langle e, \pi \rangle \in [e, \pi] \) ensures that rules are deterministic; E-App will only trigger if at least the callee of the application or one of its arguments are not yet in normal form. If the callee and the arguments are in normal form, E-App will handle applications with a known function as callee. It will reduce to the function’s body while substituting all parameters with their arguments. If the callee is unknown, the application reduces to the artificial skip body (E-App_{skip}).

An if body first evaluates its arguments (E-Hz) and then selects either its true or false continuation if possible (E-Hz;₁ and E-Hz;₂). Similar to E-App, rule E-Hz only triggers if at least one of its expressions are
Remark. Not every function implies that a function does not have free variables.

### 3.2 Full and Partial Evaluation

If we evaluate a program with constants as arguments, we will never have to handle the case that an expression does not reduce to a constant. In particular, we will never trigger E-App nor E-Skip.

**Proof.** For the proof of Lemma 2 we have shown that all expressions collapse to constants during evaluation. Thus, the preconditions for E-App or E-Skip will never trigger.

If a program does not contain any letrecs, evaluation will always terminate.

**Lemma 4 (Termination without Letrec),** A function constant not containing letrecs terminates for all inputs.

**Proof sketch.** Tait [48] proved that the simply-typed lambda calculus is strongly normalizing. If we remove letrecs from $\lambda^m$, the resulting language is analogously strongly normalizing.

**Remark.** In particular, it is not possible to create a fixed-point operator like the Y combinator as in the untyped lambda calculus without relying on letrec. This also means, that all potential loops in a $\lambda^m$ program are syntactically recognizable by its letrecs. A $\lambda^m$ program cannot have other causes for divergence.

Now, we study the evaluation of a program $b$ which still contains free variables. If we lambda-lift all free variables into a function $f_b$, all variables in $f_b$ will be bound and we can perform a full evaluation (Lemma 3). If we perform $n$ steps on $b$ we will partially evaluate $b$ and obtain $b'$. If we lambda-lift $b'$ we can perform a full evaluation on that program. The following theorem states that both functions still compute the same result:
Theorem 1 (Correctness). Let \( b \) be a well-typed body under \( X(b) \). We define \( f' := \lambda X(b).b' \) with \( b \Rightarrow^n b' \) for some \( n \in \mathbb{N} \). If \( f(b) \) terminates with result \( i \) so does \( f'(\overline{f}) \):

\[
\forall \overline{c} \in C(f) : f(b) \Rightarrow^i \ overline{c} \Rightarrow^i \ overline{f} \Rightarrow^i \overline{f}(\overline{f}) \Rightarrow^i \overline{c} \Rightarrow^i \overline{f}.
\]

Remark. Note that \( f' \) uses the free variables of \( b \) and not \( b' \). This is because PE might eliminate some free variables but we would like both \( f_b \) and \( f' \) to have the same signature.

Proof sketch. By PCF being confluent [18].

In order to prove that PE does not induce divergence, we have to show that PE will always obtain a normal form in finitely many steps, if the original program terminates for at least one configuration.

Theorem 2 (Termination Guarantee). Let \( b \) be a well-typed body under \( X(b) \). If \( f_x(b) \) terminates for some valid configuration \( \overline{c} \) of \( f_x \), partially evaluating \( b \) terminates to a normal form \( b' \) in finitely many steps:

\[
(\exists \overline{c} \in \mathcal{C}(f) : f_x(b) \Rightarrow^i \overline{c} \Rightarrow^i b \Rightarrow^* b').
\]

Proof. Let \( i \) be the number of times rules E-App

or E-Hz skip are triggered during evaluation. We prove the induction by induction on \( i \).

By Lemma 5 evaluating \( f_x(b) \Rightarrow^i \overline{c} \) never triggers E-App

nor E-Hz. As an inductive base case, we assume that \( b \Rightarrow \overline{a} \) does neither trigger E-App

nor E-Hz for any \( n \). This implies that both evaluations \( f_x(b) \Rightarrow^i \overline{c} \Rightarrow^i \) \( l \) and \( b \Rightarrow^* b' \) will trace through the same functions till \( b' \). Merely, expressions which do not contribute to control flow might not be constants. Hence, there exists an \( n \) such that \( b' = \overline{e} \).

For the inductive step we have to show that triggering E-App

or E-Hz \( l + 1 \) times still leads to a normal form. Both E-App

and E-Hz, trigger E-Skip. By Definition \( b \) postdom(b) either retrieves \( l \) or a body \( p \leq b \) which preserves \( b' \)s termination behavior.

By the induction hypothesis evaluating \( p \) along E-Skipping will finally trigger E-Skipping, if the original program terminates. □

3.3 Run and Halt Annotations

Until now, we studied partial evaluation of a whole program. In practice, the programmer only wants to specialize certain parts of the program while other parts should be excluded from specialization (Section 5). The compiler searches the program for annotations from top to bottom. Assume the compiler reaches an annotation as in the following run-annotated Impala snippet:

\[
\texttt{letrec count = } \lambda \texttt{(i, n, ret). } \texttt{if(i < n) } \texttt{then( } \texttt{count(1+1, n, ret). } \texttt{else( } \texttt{count(1+1, n, ret)).}
\]

Its \( \lambda \) translation is presented on the right-hand side. During the search for annotations, all \texttt{letrecs} are substituted in—as in rule E-Letrec. In practice, we perform the substitution on demand when a free variable bound by a \texttt{letrec} is encountered. The run annotation \( \overline{0} \) then causes \( f(\texttt{args}, k) \) to be specialized by triggering rule E-App.

As PE should only run until a continuation (e.g. \( k \)) outside the body is called, all remaining free variables, together with any \texttt{exit} must be considered local \texttt{exit}s for the purpose of evaluation and the definition of post-domains in particular. Similarly, a \texttt{halt} annotation \( f(\texttt{args}, j) \) causes the evaluator to stop specialization at that point and resume at \( i \).

Run annotations impact the termination of PE. If a run-annotated code block is unreachable, the partial evaluator might be subject to hidden divergence:

\[
\texttt{if i_am_always_false } \{ \texttt{#will_not_terminate(42) \} }
\]

Under certain circumstances, run annotations might induce divergence. Reconsider the function count from Section 2.4 where the recursive call is annotated with \( B \). First, let us convert the function to \( \lambda \):

\[
\texttt{fn count(i: int, n: int) -> int \{ } \texttt{\texttt{letrec } count = } \lambda \texttt{(i, n, ret). } \texttt{if(i < n) } \texttt{then( } \texttt{count(1+1, n, ret). } \texttt{else( } \texttt{count(1+1, n, ret).}
\]

After substitution of count and application of E-App the following program results:

\[
\texttt{letrec count = } \lambda (i, n, ret). \texttt{if(i < n) } \texttt{then( } \texttt{count(1+1, n, ret). } \texttt{else( } \texttt{count(1+1, n, ret).}
\]

16
would then skip the remainder of the scope.

value for each outside function within the calling context of an in-

continuing evaluation from the computed post-dominator remains

required because any function passed to an inside function from

sensitive nodes

computed for

side function. This allows the

ments over all calls when constructing the

Partial context-sensitive CFA.

A symbolic CFA. The call to out may pass other functions to the

inside higher-order function C it receives as argument. For this

reason, we let the call to out symbolically represent any function

reachable from it. The CFA then determines the values B and out

for parameter pc. Similarly, the CFA obtains out for pa.

The CFA never tracks values for the parameters of an outside

function and does not propagate arguments of a call to it. Instead,

the CFA represents control flowing through an outside function with

a symbolic value. In the example, T calls out, which in turn may call

C. The analysis propagates the symbolic value out to C’s parameter

pc, and thus, C may jump back to out.

Merging control flow over all contexts for out leads to an

unacceptable loss in precision. The CFA in Figure 5c has only a

single node for out with edges to nodes A and C as these are the

arguments to any call to out. This yields the imprecise post-

dominator B

Partially context-sensitive CFA. Instead of merging the argu-

ments over all calls when constructing the CFG, we give one level of

to the calling context of an in-

side function. This allows the CFA to distinguish control flow from

multiple calls to an outside function to their function arguments. The

CFG in Figure 5d thus separates the calls to out in T, A and C and

their different arguments. The edge from outC to # and its absence

from other out shows the increase in precision. The post-dominator

computed for E is then C.

Figure 5d also shows that we need to add edges from context-

sensitive nodes to the outside node out. These edges are required

because any function passed to an inside function from

outside, might also call any other function leaked to the outside

function previously.

The CFA approximation is safe for computing post-dominators

because the only way for an outside function out to return to the

scope is via functions passed to it. The CFA does not model control-

flow for diverging or non-returning out. In these cases, however, the
evaluator cannot guarantee termination anyway (see Theorem 2) and

continuing evaluation from the computed post-dominator remains

sound for returning cases.

fn iterate(fld: Field; body: fn(int, int) -> () -> () { let vector_length = 8; for y in range(0, fld.rows) { vectorize(vector_length, 0, fld.cols, |x| body(x, y)); } (a) Iterator implementation for SIMD hardware

fn iterate(fld: Field; body: fn(int, int) -> () -> () { let grid = (fld.cols, fld.rows, 1); let block = (16, 1, 1); nvvm(grid, block, | { let x = tid.x()%grid.x(); tid.y()/grid.y(); body(x, y); }); (b) Iterator implementation for NVIDIA GPUs

5. Applications and Evaluation

The Impala compiler translates the source program into Thorin [32]:

a functional intermediate representation (IR) similar to Java.

Partial evaluation and other optimizations are performed at that level.

Finally, the compiler either translates to C/CUDA/OpenCL or

LLVM/SPIR/NVVM (see below).

Mapping Algorithms to Di

erent Architectures. In order to ab-

stract from specific target platforms, Impala provides intrinsic

higher-order functions. For example, invoking the following func-

tion vectorizes [28] body for SIMD width L and creates an appro-

riate loop from to b:

e = fn vectorize(l: int, a: int, b: int, body: fn(int) -> () -> ()

Likeewise, invoking the following function causes body to be exec-

uted via NVVM on an NVIDIA GPU:

The execution runs in parallel by the threads defined by grid with

the given blocking. Similarly, Impala supports code generation for

CUDA, OpenCL, and SPIR. In contrast to pragma-based solutions

like OpenACC or OpenMP, Impala’s intrinsics integrate seamlessly

into Impala’s type system. This allows the programmer to hide the

use of these functions behind other functions. Consider a higher-

order function iterate in order to iterate over a field:

iterate(fld, |x, y| { /* some loop body */ });

Figure 6 depicts one implementation of iterate which vectorizes

the loop body and one which schedules the loop body on the GPU.

Other iterator implementations may use other intrinsics and/or

more sophisticated blocking schemes

DSL Embedding. We demonstrate the effectiveness of our partial

evaluator on two examples. First, we present a small framework

for stencil computations in image processing: The framework

essentially is an “interpreter” that applies a stencil to an image.

The aspects of boundary handling, application of the stencil, and

the stencil itself are cleanly separated. PE composes those aspects

together and produces high-performance code that we specialize

for execution on CPU and GPU targets. Second, we present a DSL

for the V-cycle multigrid iteration; a multigrid method widely-used

in high-performance computing. The V-cycle employs different

stencils to smooth the error on different resolutions of the same data.

Passing the V-cycle components as functions to the DSL allows

us to merge multiple components in order to reduce high latency

memory accesses.

5.1 Stencil Computations

A linear filter convolves an image with a filter mask by applying the

filter mask to each pixel. Examples of linear filters are the Gaussian

blur filter, the Laplace operator, or the Sobel operator. Since the filter

mask for linear filters like the Gaussian blur or the Sobel operator

http://anydsl.github.io
All variants in Table 1 are specialized in this way for the Gaussian blur filter of size $5 \times 5$ and as input memory for the second component. Fusing multiple components is outlined in Section 5.2.

**Evaluation.** For the measurements we use a separated version of the Gaussian blur filter with a $5 \times 5$ filter mask and an image of 4096 x 4096 pixels. All specialized versions are generated from the same generic description using PE. Table 1 shows the median execution time in ms on the GTX 680 using the CUDA 6.5 drivers and toolkit, on the R9 290X using the Catalyst 15.7 drivers, on the Iris 5100, and on the Intel Core i5-4288U. On the GPU, the median of seven runs is used while 27 runs are used on the CPU.

The last line shows the execution for hand-tuned CUDA, OpenCL, and CPU (vectorized C++) implementations from OpenCV (version 2.4.10), a state-of-the-art image processing toolbox. The CUDA implementation in OpenCV is provided by NVIDIA experts and uses similar optimizations: the filter is separated, the iteration space is unrolled, border handling is limited to thread blocks at the image border, and fast on-chip scratchpad memory is used to stage data. The OpenCL implementation in OpenCV is provided by AMD experts and merges the row and column component in a single kernel: First, the data is loaded to fast on-chip scratchpad memory and the column component is executed, storing its results again to the scratchpad memory. Then, the row component is executed, loading its input from scratchpad and storing the result back to device memory. On the GPU, the row component is manually vectorized 8-fold using double-pumped SSE. The column component uses superword-level parallelism (SLP), unrolling multiple loop iterations so that the compiler can merge them easily into vector operations. The schedule applies always first the row component and then the column component. This allows to hold the intermediate results in cache. It can be seen from Table 1 that the specialized versions we obtain through PE even outperform the hand-tuned implementations in OpenCV.

At the same time, our implementation is more concise: The hand-tuned CUDA version from OpenCV consists of 251 lines of CUDA code plus 386 lines of kernel instantiations for different filter mask sizes and boundary handling modes. The kernel implementation in OpenCL requires 142 lines of OpenCL code; boundary handling is realized via macros that wrap memory accesses. For the CPU implementations, more than 1500 lines of code are required, providing specialized implementations for different data types, kernel sizes, and target instruction sets (SSE, Neon). Our Impala implementation requires 88 lines of code.

We use the *run* annotations highlighted in the sample codes to trigger PE; none of them annotates a “dangerous” recursive call (see Section 3.3). We used *half* annotations only to prevent loops iterating over the field to be evaluated at compile time.
5.2 The V-Cycle Multigrid Solver

The basic idea of the multigrid method is to smooth the error (e.g., using an iterative method like Jacobi or Gauss-Seidel) on different resolutions of the same data. The V-cycle describes one possible multigrid iteration [4, 50]. To transform data between different resolutions of the multigrid the algorithm uses the restrict and interpolate methods. On each level, the error is smoothed (smoother) and estimated (residual). This process is recursive and starts at the finest resolution.

For a V-cycle DSL we would like to have the different methods pluggable. Using Impala we achieve the same optimization by custom iterate functions that compute multiple components at once. As an example, consider the computation of the residual component followed by the restrict component: Instead of computing the residual for the whole field first and then restrict the field produced by the residual, we compute the residual only for two rows and restrict the residual before the next rows are processed. This pipelined processing allows to hold the result of the restrict component in cache on the GPU and allows to merge compute kernels on the GPU when using scratchpad (local or shared) memory. On the GPU, this has the same effect as loop fusion. Figure 8 illustrates this for the CPU. The index passed to the residual and restrict component refer to the temporary field. The offset to the current row of the other fields are tracked in the Field object and are used when accessing field elements. Merging the two components is only valid if the operation of the multigrid components is known: in our example, the restrict component is allowed to access two rows only. Otherwise, a larger temporary array has to be allocated and pre-computed before applying restrict.

Results. While we have shown in Section 5.1 that we achieve competitive performance for stencil codes, the multigrid iteration offers optimization opportunities when components are scheduled in a clever way. Figure 9 shows the speedup we get by merging the residual and restrict components for the first level of the V-cycle (smooth, residual, restrict, interpolate). The speedup is between 11 % on the CPU and up to 20 % on the GPU. Considering only the residual and restrict component, the computation is 25 % (72 %) faster on the CPU (AVX) and 42 % (45 %) faster on the GPU when using NVVM (SPIR). For AVX, we vectorize only the smooth and residual component since restrict and interpolate are otherwise slower due to their memory access pattern. On the Iris 5100, the execution takes 16 % longer when the two components are merged. Note that this is expected since the scratchpad memory is mapped to slow global memory in the Iris 5100 architecture. Consequently, the specialization for the Iris 5100 would not make use of scratchpad memory.

Furthermore, we compare the performance of our specialized V-cycle implementation against the performance of generated implementations by HIPAC, a DSL framework for stencil computations [35]. HIPAC provides CUDA and OpenCL backends for execution on GPUs. We use the HIPAC implementation from [34], which uses the same V-cycle components as our implementation.

For the first level of the V-cycle, our normal implementation has the same performance on the Iris 5100 (32.54 ms vs. 32.81 ms), is 8 % faster on the Radeo R9 290X (2.26 ms vs. 2.43 ms), and is 9 % slower on the GTX 680 (4.78 ms vs. 4.35 ms). Our merged imple-

```c

fn vcycle(field: Field, lvls: int, vsteps: int, ssteps: int, smoother: \(\text{fn}(*...*) \rightarrow \text{Field}\)) => Field {
    allocate memory for all lvls; Sol, RHS, Res, Temp
    if lvls == lvls-1 {
        for i in range(0, lvts) { // solve by vsteps smooths
            for x, y in iterate(Sol(lvts))
                solve(x, y, /\\text{fields}/);}
        }
        else {
            for i in range(0, lvts) { // pre-smoothing
                for x, y in iterate(Sol(lvts))
                    solve(x, y, /\\text{fields}/);}
            }
            for x, y in iterate(Rhs(lvts))
                restrict(x, y, /\\text{fields}/);}
            for x, y in iterate(Sol(lvts))
                residual(x, y, /\\text{fields}/);}
            for x, y in iterate(Sol(lvts))
                interpolate(x, y, /\\text{fields}/);}
        }
    }
    return field;
}

let res = @cycle(field, lvls, vsteps, ssteps, jacobi, residual, restrict, interpolate);

Figure 7. V-cycle implementation in Impala.

Figure 8. Merging residual and restrict on the CPU.

Figure 9. Speedup from merging the residual and restrict computation for the first level (4096 x 4096) of the V-cycle (smooth, residual, restrict, interpolate). The speedup over HIPAC implementations is also given where available.

Discussion. Our implementation can be easily extended to express different multigrid iterations. It is actually sufficient to change the recursion in the V-cycle implementation in order to get the schedule for the W-cycle multigrid iteration.

The evaluation has shown that we can map the same high-level description to different target platforms by providing target-specific mappings. Moreover, we merge multiple components as shown

\(\text{Figure 7. V-cycle implementation in Impala.}\)

\(\text{Figure 8. Merging residual and restrict on the CPU.}\)

\(\text{Figure 9. Speedup from merging the residual and restrict computation for the first level (4096 x 4096) of the V-cycle (smooth, residual, restrict, interpolate). The speedup over HIPAC implementations is also given where available.}\)
exemplarily for the residual and restrict components. This yields specialized implementations that outperform the implementations generated by HIPAP, which has no support for kernel fusion.

6. Conclusions

In this paper, we present a novel partial evaluation strategy and its application to shallow embedding of DSLs. To this end, we formally define full as well as partial evaluation and prove its correctness and termination property on a CPS-based variant of PCF. Every invocation of the partial evaluator will terminate if full evaluation of that program terminates. The termination behavior of the program is retained in a predictable way: The partial evaluator skips dynamic branches and continues at the post-dominator. To efficiently compute post-dominators on parts of higher-order programs with free variables we presented a local, partially context-sensitive CFA. In order to steer the evaluator from within the language, we introduce run- and halt-annotations. As long as run annotations are not placed on recursive calls, PE will not induce divergence. Annotating a call to a recursive function is not problematic.

We apply our partial evaluator on higher-order functions to embed high-performance DSLs and generate optimized code for different hardware architectures. We evaluate our technique experimentally in the field of visual and high-performance computing and show that our evaluator produces highly specialized and efficient code for CPUs as well as GPUs that matches the performance of vendor code for CPUs as well as GPUs that matches the performance of vendor code for CPUs as well as GPUs that matches the performance of vendor code for CPUs as well as GPUs that matches the performance of vendor code for CPUs as well as GPUs that matches the performance of vendor code for CPUs as well as GPUs that matches the performance of vendor code for CPUs as well as GPUs that matches the performance of vendor code for CPUs as well as GPUs that matches the performance of vendor code for CPUs as well as GPUs that matches the performance of vendor code.

Acknowledgments

This work is supported by the Federal Ministry of Education and Research (BMBF) as part of the ECOUSS project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken. It is also co-funded by the European Union (EU), as part of the Dreamspace project as well as by the Intel Visual Computing Institute Saarbrücken.

References