

Type-Driven Partial Evaluation without Code Duplication

No Author Given

No Institute Given

Abstract.

Keywords: Partial Evaluation

1 Introduction

Partial evaluation [?] is an optimization technique that identifies *statically known* program parts and pre-computes them at compile time. The compile-time computation yields a *residual program* that does not contain the, previously identified, statically known parts of the program. Partial evaluation has been intensively studied and successfully applied for: removing abstraction overheads in high-level programs [?,?], domain-specific languages [?,?], and converting language interpreters into compilers [?,?,?]. Applying partial evaluation in these domains often improves program performance by several orders of magnitude [?,?].

Unlike other compiler optimizations partial evaluation is not *safe* as it might lead to *code explosion* and might not *terminate*. Due to compile-time execution, computing **folds** and loops over data structures of static size can produce arbitrarily large programs. Furthermore, in a Turing-complete language assuring termination is undecidable.

Automatically assuring safety of partial evaluators necessarily leads to lack of *predictability*. To illustrate, let us define a function **dot** for computing a dot-product of two vectors that contain numeric values .

```
def dot[V:Numeric](v1: Vector[V], v2: Vector[V]): V =  
  (v1 zip v2).foldLeft(zero[V]){ case (prod, (c1, cr)) =>  
    prod + c1 * cr  
  }
```

When **dot** is called with vectors of static size (*e.g.* **dot(Vector(2, 4), Vector(1, 10))**) the abstraction overhead of **zip** and **foldLeft** can be completely removed. However, the partial evaluator must apply extensive analysis to conclude that vectors are static in size and that this can be later used to unroll the recursion inside **foldLeft**. Even if the analysis is successful the evaluator must be conservative about unrolling the **foldLeft**. The vector sizes, and thus the produced code,

⁰ We use Scala for all code examples in this paper. In order to comprehend the paper the reader is required to know the mere basics of the language

can unacceptably large in a general case. What if we know that vector sizes are relatively small and we would like to predictably unroll `dot` into a flat sum of products?

Lack of predictability and danger of code explosion are the reason that successful partial evaluators [?, ?, ?, ?] are programmer controlled. We categorize the existing solutions in three categories (for further discussion *c.f.* §??):

- Programming languages Idris and D provide allow placing the `static` annotation on function arguments. Since `static` is placed on terms, it denotes that the *whole term* is static. This restricts the number of programs that can be expressed, *e.g.* , we could not express in the signature of `dot` that vector parameters are static in size.
- Type-directed partial evaluation [?] and Lightweight Modular Staging [?] use types to communicate the programmer’s intent about partial evaluation. By changing the types of parameters to be (*e.g.*

2 Formalization $F_{i<}$:

$t ::=$	Terms:	$S, T, U ::=$	Types:
x, y	identifier	$iS \Rightarrow jT$	function type
$(x : iT) \Rightarrow t$	function	$\{x : iS\}$	record type
$t(t)$	application	$[X <: iS] \Rightarrow jT$	universal type
$\{x = t\}$	record	Any	top type
$t.x$	selection	$iT, jT, kT, lT ::=$	Binding-Time Types:
$[X <: iT] \Rightarrow t$	type abstraction	X	type identifier
$t[iT]$	type application	$T, dynamic\ T$	dynamic type
$inline\ t$	inline view	$static\ T$	static type
$dynamic\ t$	dynamic view	$inline\ T$	inline type
$v ::=$	Values:	$\Gamma ::=$	Contexts:
$x \Rightarrow t$	function value	\emptyset	empty context
$\{x = t\}$	record value	$\Gamma, x : iT$	term binding
		$\Gamma, X <: iT$	type binding

Fig. 1. Syntax of $F_{i<}$.

We formalize the essence of our inlining system in a minimalistic calculus based on $F_{<}$ with lazy records. To accommodate predictable partial evaluation we introduce binding-time annotations into the type system as first-class types that represent three kinds of bindings:

1. **Dynamic binding.** These are the types which express computation at runtime. All types written in the end user code are considered to be dynamic by default if no other binding-time annotation is given.
2. **Static binding.** Values of static terms can be computed at compile-time (*e.g.* constant expressions) but are still evaluated at runtime by default. All language literals are static by default.
3. **Inline binding.** And finally the types that correspond to terms that are hinted to be computed at compile-time whenever possible.

2.1 Composition

An interesting consequence of encoding of binding times as first-class types is ability to represent values which are partially static and partially dynamic. For example let's have a look at simple record that describes a complex number with two possible representations encoded through *isPolar* flag:

$$complex : static \{isPolar : static\ Boolean, a : Double, b : Double\} \in \Gamma$$

This type is constructed out of a number of components with varying binding times. Representation encoding is known in advance and is static according to

the signature. Coordinates a and b do not have any binding-time annotation meaning that they are dynamic.

Given this binding to *complex* in our environment Γ we can use *inline* to obtain a compile-time view to evaluate access to *isPolar* field at compile-time:

inline complex.isPolar : inline Boolean

Any statically known expression can be promoted via *inline*. Selection of dynamic fields on the other hand will return dynamic values despite the fact that record is statically known. In practice this can be used to specialize a particular execution path in the application to a particular representation by selectively inlining statically known parts.

Once you have inline view of the term it's also possible to demote it back to runtime evaluation through *dynamic* view.

Not all type and binding time combinations are correct though. We restrict types to disallow nesting of more specific binding times into less specific ones.

$$\begin{array}{ll}
 \text{wff } iAny & \text{(W-ANY)} \\
 \frac{i <: j \quad i <: k \quad \text{wff } jT_1 \quad \text{wff } kT_2}{\text{wff } i(jT_1 \Rightarrow kT_2)} & \text{(W-ABS)} \\
 \frac{i <: j \quad i <: k \quad \text{wff } jS \quad \text{wff } kT}{\text{wff } i([X <: jS] \Rightarrow kT)} & \text{(W-TABS)} \\
 \frac{\forall j. \quad i <: j \quad \text{wff } jT}{\text{wff } i\{x : jT\}} & \text{(W-REC)}
 \end{array}$$

Fig. 2. Well formed types wff iT

This restriction allows us to reject programs that have inconsistent annotations. For example the following function has incorrectly annotated parameter binding time:

$$(x : inline Int) \Rightarrow x + 1$$

This is inconsistent because the body of the function might not be evaluated at compile-time (as the function is not inline.) As described in (W-ABS) functions may only have parameters that are at most as specific as the function binding-time. In our example this doesn't hold as *inline* is more specific than implicit *static* annotation on function literal.

2.2 Subtyping

Another notable feature of our binding-time analysis system is deep integration with subtyping. We believe that such integration is crucial for an object-oriented language that wants to incorporate partial evaluation.

At core of the subtyping relation we have a subtyping relation on binding-time information with *dynamic* as top binding-time.

$i <: \text{dynamic}$ (I-DYNAMIC)	$\text{inline} <: \text{static}$ (I-STATIC2)
$\text{static} <: \text{static}$ (I-STATIC1)	$\text{inline} <: \text{inline}$ (I-INLINE)

Fig. 3. Binding-time subtyping.

We proceed by threading binding time information throughout regular $F_{<}$ subtyping rules augmented with standard record types.

$\Gamma \vdash iS <: \text{Any}$	(S-TOP)
$\Gamma \vdash iS <: iS$	(S-REFL)
$\Gamma \vdash iS <: jU \quad \Gamma \vdash jU <: kT$	(S-TRANS)
$\Gamma \vdash iS <: jU$	
$X <: iT \in \Gamma$	(S-TVAR)
$\Gamma \vdash X <: iT$	
$\{x_p : i_p T_p \mid p \in 1..n+m\} <: \{x_p : i_p T_p \mid p \in 1..n\}$	(S-WIDTH)
$\Gamma \vdash kT_1 <: iS_1 \quad \Gamma \vdash jS_2 <: lT_2$	(S-ARROW)
$\Gamma \vdash iS_1 \Rightarrow jS_2 <: kT_1 \Rightarrow lT_2$	
$\forall p \in 1..n. i_p S_p <: j_p T_p$	(S-DEPTH)
$\{x_p : i_p S_p \mid p \in 1..n\} <: \{x_p : j_p T_p \mid p \in 1..n\}$	
$i <: j \quad \Gamma \vdash S <: T$	(S-INLINE)
$\Gamma \vdash iS <: jT$	
$\Gamma, X <: iU_1 \vdash jS_2 <: kT_2$	(S-ALL)
$\Gamma \vdash [X <: iU_1] \Rightarrow jS_2 <: [X <: iU_1] \Rightarrow kT_2$	
$\{x_p : i_p S_p \mid p \in 1..n\}$ is permutation of $\{y_p : j_p T_p \mid p \in 1..n\}$	(S-PERM)
$\{x_p : i_p S_p \mid p \in 1..n\} <: \{y_p : j_p T_p \mid p \in 1..n\}$	

Fig. 4. Subtyping.

Integration between binding-time subtyping and subtyping on regular types is expressed through (S-INLINE) rule that merges the two into one coherent relation on binding-time types.

2.3 Generics

Crucial consequence of our design choices made in the system manifests in ability to use regular generics as means to abstract over binding-time without any additional language constructs.

For example given a generic identity function:

$$\text{identity} : \text{static} ([X <: \text{Any}] \Rightarrow \text{static} (X \Rightarrow X)) \in \Gamma$$

We can instantiate it to both in static and dynamic contexts through corresponding type application:

$$\begin{aligned} \text{identity}[\text{static } Int] &: \text{static } (\text{static } Int \Rightarrow \text{static } Int) \\ \text{identity}[Int] &: \text{static } (Int \Rightarrow Int) \end{aligned} \quad (1)$$

In practice this allows us to write code that is polymorphic in the binding time without any code duplication which is quite common in other partial evaluation systems.

This is possible due to the fact that we've integrated binding time information into types and augmented subtyping relation with subtyping

2.4 Typing

To enforce well-formedness of types in a context of partial evaluation we customize standard typing rules with additional constraints with respect to binding time.

$$\begin{aligned} &\frac{x : iT \in \Gamma}{\Gamma \vdash x : iT} && \text{(T-IDENT)} \\ &\frac{\forall t. \quad \Gamma \vdash t : jT \quad \text{wff } i\{x : jT\}}{\Gamma \vdash i\{x = t\} : i\{x : jT\}} && \text{(T-REC)} \\ &\frac{\Gamma \vdash t_1 : i(jT_1 \Rightarrow kT_2) \quad \Gamma \vdash t_2 : jT_1}{\Gamma \vdash t_1(t_2) : kT_2} && \text{(T-APP)} \\ &\frac{\Gamma \vdash t : i\{x = jT_1, y = kT_2\}}{\Gamma \vdash t.x : jT_1} && \text{(T-SEL)} \\ &\frac{t \text{ is not literal} \quad \Gamma \vdash t : \text{static } T}{\Gamma \vdash \text{inline } t : \text{inline } T} && \text{(T-INLINE)} \\ &\frac{t \text{ is not literal} \quad \Gamma \vdash t : iT}{\Gamma \vdash \text{dynamic } t : \text{dynamic } T} && \text{(T-DYNAMIC)} \\ &\frac{\Gamma \vdash t : iS \quad \Gamma \vdash iS <: jT}{\Gamma \vdash t : jT} && \text{(T-SUB)} \\ &\frac{\Gamma, x : jT_1 \vdash t : kT_2 \quad \text{wff } i(jT_1 \Rightarrow kT_2)}{\Gamma \vdash i((x : jT_1) \Rightarrow t) : i(jT_1 \Rightarrow kT_2)} && \text{(T-FUNC)} \\ &\frac{\Gamma, X <: jT_1 \vdash t_2 : kT_2 \quad \text{wff } i([X <: jT_1] \Rightarrow kT_2)}{\Gamma \vdash i([X <: jT_1] \Rightarrow t_2) : i([X <: jT_1] \Rightarrow kT_2)} && \text{(T-TABS)} \\ &\frac{\Gamma \vdash t_1 : i([X <: jT_{11}] \Rightarrow kT_{12}) \quad \Gamma \vdash lT_2 <: jT_{11}}{\Gamma \vdash t_1[lT_2] : [X \mapsto lT_2]kT_{12}} && \text{(T-TAPP)} \end{aligned}$$

Fig. 5. Typing.

The most significant changes lie in:

- Additional checks in literal typing that ensure that constructed values correspond to well-formed types (T-FUNC, T-REC, T-TABS). To do this we typecheck literals together with possible binding-time term that might enclose it.
- New typing rules for binding-time views (T-INLINE, T-DYNAMIC). These rules only cover non-literal terms as composition of binding-time view and literal itself is handled in corresponding typing rule for given literal.

2.5 Partial Evaluation

In order to simplify partial evaluation rules we erase all of the type information before partial evaluation. This means that all functions become function values, type abstraction and application are complete eliminated.

$$\begin{array}{c}
\frac{t \rightsquigarrow t'}{x \Rightarrow t \rightsquigarrow x \Rightarrow t'} \quad (PE-FUNC) \\
\frac{\bar{t} \rightsquigarrow \bar{t}'}{\{\bar{x} = t\} \rightsquigarrow \{\bar{x} = t'\}} \quad (PE-REC) \\
\frac{t_1 \rightsquigarrow t'_1 \quad t'_1 \text{ is not inline} \quad t_2 \rightsquigarrow t'_2}{t_1(t_2) \rightsquigarrow t'_1(t'_2)} \quad (PE-APP) \\
\frac{t_1 \rightsquigarrow inline \ x \Rightarrow t \quad t_2 \rightsquigarrow t'_2 \quad [x \mapsto t'_2]t \rightsquigarrow t'}{t_1(t_2) \rightsquigarrow t'} \quad (PE-IAPP) \\
\frac{t \rightsquigarrow t' \quad t' \text{ is not inline}}{t.x \rightsquigarrow t'.x} \quad (PE-SEL) \\
\frac{t \rightsquigarrow inline \ \{x = t_x, \ \bar{y} = t_y\} \quad t_x \rightsquigarrow t'_x}{t.x \rightsquigarrow t'_x} \quad (PE-ISEL) \\
\frac{t \text{ is not literal} \quad t \rightsquigarrow t' \quad t' \Downarrow v}{inline \ t \rightsquigarrow inline \ v} \quad (PE-INLINE)
\end{array}$$

Fig. 6. Partial evaluation $t \rightsquigarrow t'$

2.6 Evaluation

Once partial evaluation is complete we strip all binding-time terms and use regular untyped lambda calculus evaluation rules extended with lazy records.

$$\begin{array}{c}
\frac{v \Downarrow v}{t_1 \Downarrow x \Rightarrow t \quad t_2 \Downarrow v \quad [x \mapsto v]t \Downarrow v'} \quad (E-VALUE) \\
\frac{t_1(t_2) \Downarrow v'}{t \Downarrow \{x = t_x, \ \bar{y} = t_y\} \quad t_x \Downarrow v} \quad (E-APP) \\
\frac{t \Downarrow \{x = t_x, \ \bar{y} = t_y\} \quad t_x \Downarrow v}{t.x \Downarrow v} \quad (E-SEL)
\end{array}$$

Fig. 7. Evaluation $t \Downarrow v$

2.7 Conjectures

1. Progress.
2. Preservation.
3. Static terms are closed over statically bound variables.
4. Inline terms will be replaced with canonical value of corresponding type after partial evaluation.

3 Translating Object Oriented Features Into $F_{i<}$

The core calculus §?? captures the essence of user-controlled predictable partial-evaluation. In practice, though, it requires an inconveniently large number of `inline` calls. Moreover, the calculus does not provide a way to define data structures that would correspond to *classes* in modern multi-paradigm languages. In this section we formalize convenient implicit conversions for the calculus §??, a scheme for translating classes into the calculus how to promote constructs classes and *methods* into their partially-evaluated versions §??.

The core rules of the calculus do not support effect-full computations and each `inline` term is trivially converted to a dynamic term after erasure. In case of languages that do support mutable state and side-effects this needs to be treated specially. For simplicity, we omit side-effects from our discussion and assume that all partially evaluated code is side-effect free and that each `inline` term can be converted to dynamic code.

3.1 Implicit Conversions

One of the core assumptions of the calculus is that all `static` terms can always be promoted to `inline`. In case of function interfaces, e.g. the `dot` function in §??, that user should not manually promote all arguments to `inline`. As a convenience for a library user we provide a type-driven conversion by default (C-StaticInline in Figure ??).

$$\frac{\Gamma \vdash t_1 : i(\text{inline } T_1 \Rightarrow kT_2) \quad \Gamma \vdash t_2 : \text{static } T_1}{\Gamma \vdash \sim t_1(t_2) : kT_2 t_1(\text{inline } t_2) : kT_2} \text{ (C-STATICINLINE)}$$

Fig. 8. Type-driven conversions from `static` terms to `inline` terms.

3.2 Making Object Oriented Constructs Partially Evaluated

4 Case Studies

4.1 Integer Power Function

- Explain what happens.
- Typical partial evaluation example. Can be handled by D and Idris and not without duplication with type-driven partial evaluation.

$\llbracket \text{let } x : T_x = t_x \text{ in } t \rrbracket = ((x : T_x) \Rightarrow t)(t_x)$
 $\llbracket \text{let type } T_1 = T_2 \text{ in } t \rrbracket = ([T_1 <: T_2] \Rightarrow t)[T_2]$
 $\llbracket \text{let class } C[A](x : T_x)\{ \text{def } f[B](y : T_y) = t_f \} \text{ in } t \rrbracket =$
 $\text{let type } C = [A] \Rightarrow \{x : T_x, f : [B] \Rightarrow T_y \Rightarrow T_f\} \text{ in}$
 $\text{let } C : [A] \Rightarrow (x : T_x) \Rightarrow C[A] =$
 $[A] \Rightarrow (x : T_x) \Rightarrow \{x = x, f = [B] \Rightarrow (y : T_y) \Rightarrow t_f\} \text{ in } t$

Fig. 9. Desugaring of classes into $F_{i<}$.

$$\begin{array}{c}
\frac{\Pi \vdash T \in \Pi}{\Pi \vdash \sim iTiT} \text{ (CT-TVAR)} \qquad \frac{\Pi \vdash T \notin \Pi}{\Pi \vdash \sim iTinline T} \text{ (CT-T-VAR)} \\
\frac{\Pi \vdash \sim tt'}{\Pi \vdash \sim i\{x = t\}inline \{x = t'\}} \text{ (CT-REC)} \\
\frac{\Pi \vdash \sim iTjT}{\Pi \vdash \sim \{x : iT\}inline \{x : jT\}} \text{ (CT-T-REC)} \\
\frac{\Pi \vdash \sim iTjT \quad \Pi \vdash \sim kSlS}{\Pi \vdash \sim iT \Rightarrow kSjT \Rightarrow lS} \text{ (CT-T-ARROW)} \\
\frac{\Pi \vdash \sim jTkT}{\Pi \vdash \sim [X <: iS] \Rightarrow jT[X <: iS] \Rightarrow kT} \text{ (CT-T-UNIV)} \\
\frac{\Pi \vdash \sim tt' \quad \Pi \vdash \sim iTjT}{\Pi \vdash \sim i(x : iT) \Rightarrow tinline (x : jT) \Rightarrow t'} \text{ (CT-FUNG)} \\
\frac{\Pi, X \vdash \sim tt'}{\Pi \vdash \sim i([X <: jT_1] \Rightarrow t)inline ([X <: jT_1] \Rightarrow t')} \text{ (CT-TABS)} \\
\frac{\Pi \vdash \sim tt' \quad \Pi \vdash \sim iTjT}{\Pi \vdash \sim t[iT]t'[jT]} \text{ (CT-TAPP)}
\end{array}$$

Fig. 10. Translation of a regular class to a compile-time version.

```

@i? def pow(base: Double, exp: Int @i?): Double =
  if (exp == 0) 1 else base * pow(base, exp)

```

Fig. 11. Function for computing the non-negative power of a real number.

4.2 Variable Argument Functions

- `@i?` in argument position is a macro that expands the function to an underlying function `@i? def min_underlying[T: Numeric](values:Seq[T]@i?): T` and a macro that will call it according to the input parameters.
- Comparison to other approaches.

```
@i? def min[T: Numeric](@i? values:T*): T =
  values.tail.foldLeft(values.head)((min, el) => if (el < min) el else min)
```

Fig. 12. Function for computing the non-negative power of a real number.

4.3 Butterfly Networks

- Reference LMS. Discuss `@i!` annotation on classes. Works for both dynamic and static inputs.
- Comparison to LMS. Mention a pervasive number of annotations. Discuss duality of `Exp[T]` and `@i!`.

4.4 Dot Product

- Explain the removal of type classes together with inline. Explain how type classes are `@i?` and how they will completely evaluate if they are passed a static value.
- Comparison to other approaches.

5 Evaluation

6 Related Work

7 Conclusion

```

object Numeric {
  @i! implicit def dnum: Numeric[Double] @i! = DoubleNumeric
  @i! def zero[T](implicit num: Numeric[T]): T = num.zero
  object Implicits {
    @i! implicit def infixNumericOps[T](x: T)(implicit num: Numeric[T]): Numeric[T]#Ops = new
  }
}

trait Numeric[T] {
  def plus(x: T, y: T): T
  def times(x: T, y: T): T
  def zero: T

  @i! class Ops(lhs: T) {
    @i! def +(rhs: T) = plus(lhs, rhs)
    @i! def *(rhs: T) = times(lhs, rhs)
  }
}

object DoubleNumeric extends Numeric[Double] {
  @i! def plus(x: Double @i?, y: Double @i?): Double = x + y
  @i! def times(x: Double @i?, y: Double @i?): Double = x * y
  @i! def zero: Double = 0.0
}

```

Fig. 13. Function for computing the non-negative power of a real number.