Pathless Scala: A Calculus for the Rest of Scala

Guillaume Martres
EPFL
Lausanne, Switzerland
guillaume.martres@epfl.ch

Abstract
Recent work on the DOT calculus successfully put core aspects of Scala on a sound foundation, but subtyping in DOT is structural and therefore not easily amenable to studying the parts of Scala that are deeply tied to its nominal subtyping system. On the other hand, the Featherweight Java calculus has proven to be a great basis for studying many aspects of Java and Java-like languages. Continuing this tradition, we present Pathless Scala: an extension of Featherweight Generic Java that closely models multiple inheritance and intersection types as they exist in the Scala language today. We define the semantics of Pathless Scala by erasing it to a simpler calculus in a way that closely models how Scala is compiled to Java bytecode in practice. More than a one-off, we believe that this calculus could be extended to describe many more features of Scala, although reconciling it with DOT remains an open problem.

CCS Concepts: • Software and its engineering → Formal language definitions; Inheritance; Polymorphism.

Keywords: Trait, Featherweight, Erasure, Dotty, Java

1 Introduction
Formalizing a programming language lets us reason about the behavior of programs in the language by developing its metatheory but it also means that the implementation strategies used by compilers can themselves be formalized. While the DOT calculus [Amin et al. 2016] has been very useful as a reasoning tool for various aspects of the Scala type system, it is not really suitable for answering questions such as "How do I compile this Scala program to Java bytecode?".

To answer this question our main source of inspiration will be [Igarashi et al. 2001] which defines two calculi: Featherweight Java (FJ) which models single-class inheritance and Featherweight Generic Java (FGJ) which adds type parameters to the language, and then proceeds to define a way to compile FGJ to FJ via erasure.

We define Pathless Scala (PS) as an extension of FGJ without casts (which we choose to not study), adding multiple inheritance via traits and intersection types in the style of DOT. Unlike DOT and as its name indicate, PS lacks type members and therefore path-dependent types.

Real Scala compilers erase traits to Java interfaces, but FJ does not model interfaces so cannot be directly used as a target for our erasure. Instead our target calculus is a fragment of FJλ [Bettini et al. 2018] which extends FJ with interfaces. FJλ also supports intersections and lambdas, but because these features are not present in Java bytecode, they are not useful for our purpose and we do not use them in our erasure mapping.

2 Syntax

C, D, E
X, Y, Z
N, P, Q := C[T]
S, T, U := X | N | T & T
L :=
  class C[X <: N](f : T) <: N(M)
  trait C[X <: N] <: N(H, M)
H := def M[X <: N](x : T) : T
M := H = e
f, g
x, y, z
e ::= x
e.f
e.m[T](v)

Figure 1. Syntax of Pathless Scala

1The answer to this question matters even when compiling Scala to a different backend such as JavaScript, because alternative backends strive to preserve the semantics of the JVM to ease cross-compilation [Doeraene 2018, § 2.1]
We do not need to support translating more complex constructors because of the restrictions imposed on well-formed classes by FGJ.

For convenience, we define auxiliary functions returning the parents and the method declarations of applied class types:

\[
\begin{align*}
\text{parents}(\text{Object}) & = \emptyset \\
\text{mdecls}(\text{Object}) & = \emptyset \\
\text{parents}(\text{C} \triangleleft \text{P}(\ldots), \overline{Q}(\overline{M})) & = \text{parents}(\text{C}[\overline{T}]) = [\overline{T}/\overline{X}](\overline{P}, \overline{Q}) \\
\text{mdecls}(\text{C} \triangleleft \text{P}(\ldots), \overline{Q}(\overline{M})) & = \text{mdecls}(\text{C}[\overline{T}]) = [\overline{T}/\overline{X}](\overline{P}, \overline{Q})
\end{align*}
\]

As an example, using the definition of A above we have:

\[
\begin{align*}
\text{parents}(\text{A[Object]}) & = \text{B[Object]} \\
\text{mdecls}(\text{A[Object]}) & = x : \text{C}, y : \text{D}
\end{align*}
\]

The set of base types of N is the transitive closure of parents(N) plus N itself.

3 Multiple Inheritance in Scala

The main difference between trait inheritance in Scala and interface inheritance in Java is that the order in which parent traits are inherited matters. In particular, Scala defines a canonical order of the base types of a class called its linearization.

3.1 \(\mathcal{L}(N)\): The Base Types of N in Linearization Order

[Odersky and Zenger 2005] defines linearization for class types C, but it is useful to generalize it to non-variable types N:

\[
\mathcal{L}(N) = N, \mathcal{L}(N_n) \dagger \ldots \dagger \mathcal{L}(N_1)
\]

Where \(\dagger\) denotes concatenation with elements on the right replacing identical elements of the left operand. It is illegal to inherit the same class twice if it is applied to different type arguments \(^2\) and so \(\dagger\) is undefined in that case:

\[
\bullet \dagger \overline{N} = \overline{N}
\]

\[
N_0 \in \overline{N_r}
\]

\[
(N_0, \overline{N_l}) \dagger \overline{N_r} = \overline{N_l} \dagger \overline{N_r}
\]

\[
N_0 = C_0[\ldots] \quad \overline{N_r} = C_r[\ldots] \quad C_0 \notin \overline{C_r}
\]

\[
(N_0, \overline{N_l}) \dagger \overline{N_r} = N_0, (\overline{N_l} \dagger \overline{N_r})
\]

We will use linearization to determine which base type of N contains the implementation of m that will be called at runtime which we dub the implementer of m in N.

3.2 mimpl(m, N): The Implementer of m in N

The following class table is legal:

\[
\begin{align*}
\text{class One; class Two} \\
\text{trait Base ( def foo(): Object )} \\
\text{trait Sub1 < Base ( def foo(): Object = new One )} \\
\text{trait Sub2 < Base ( def foo(): Object = new Two )} \\
\text{class A < Object, Sub1, Sub2}
\end{align*}
\]

\(^2\)In real Scala this is in fact possible with variant type parameters.
However, the equivalent class table in Java (using interface instead of trait) would be illegal: both Sub1 and Sub2 contain a concrete implementation of foo and neither trait overrides the other. By contrast, in Scala this typechecks and (new A).foo() will evaluate to new Two() because Sub2 precedes Sub1 in the linearization of A.

In general, concrete methods override abstract methods in both Java and Scala, but if we compare a concrete method M defined in C with another concrete method M’ defined in D then:

- In Java, M overrides M’ if D is a base type of C.
- In Scala, M overrides M’ in N if C precedes D in L(N).

Since a type P will always appear before its parent in any linearization involving P, this generalizes the Java rule.

From this it follows that mimpl(m, N) must be the first type in L(N) containing a concrete declaration of m. In the example above we had L(A) = A, Sub2, Sub1, Object and so we get mimpl(foo, A) = Sub2 as expected.

For a class C to be well-formed, it is not enough for mimpl to be defined for all its members, we must also check that the selected implementations are valid overrides.

3.3 isValid(m, C): The Implementation of m in C Is a Valid Override of All Declarations of m in the Base Types of C

Like in Java, for an override to be valid its type must match the type of all the overridden methods, meaning the type and term parameters must be identical (up to alpha-renaming which we do not model) and the result type is allowed to vary covariantly. Additionally, the override must not be accidental, a concept specific to Scala.

\[
\forall n \in L(C[X]), \text{override}_{m,C[N]}(m, N, n) \quad \text{and} \quad \text{noAccidentalOverride}(m, N, n) \quad \Rightarrow \quad \text{isValid}(m, C)
\]

3.3.1 override₄(m, Nᵢ, Nₒ): the Type of m in Nᵢ Overrides the Type of m in Nₒ in the Type Environment Δ.

The following class table is illegal:

```
class One; class Two
trait L { def foo(): One = new One }
trait R { def foo(): Two = new Two }
class LR < L, R ()
```

Even though mimpl(foo, LR) = R, the override is invalid because Two is not a subtype of One. This is enforced by:

\[\text{def } m[V <: P(x: T): T_i = \ldots \in mdecls(N_i)] \quad \text{def } m[V <: P(x: T): T_o \equiv \ldots \in mdecls(N_o)] \quad \Rightarrow \quad \text{override}_\Delta(m, N_i, N_o)\]

Interestingly, our definition of override is more expressive than the one in FGJ as it takes a type environment \(\Delta\) representing the type parameters of the class. This is needed to be able to typecheck:

```
class X
class Base { def foo(): X = ... }
class Sub[S <: X] < Base { def foo(): S = ... }
```

The corresponding Java code is valid and yet Sub is not well-formed in FGJ because the type parameter S <: X is not in the type environment when the override check is done [Igarashi et al. 2001, Figure 6].

3.3.2 noAccidentalOverride(m, Nᵢ, Nₒ): m in Nᵢ Does Not Accidentally Override m in Nₒ.

The following class table is not well-formed in Scala:

```
class One; class Two
trait Base { def foo(): Object }
trait Sub1 < Base { def foo(): Object = ... }
trait Unrelated { def foo(): Object }
trait Sub2 < Unrelated { def foo(): Object = ... }
class A < Object, Sub1, Sub2
```

Although we have mimpl(foo, A) = Sub2 and override(m, Sub2, Sub1), the compiler complains:

```
method foo in trait Sub2 cannot override a concrete member without a third member that's overridden by both (this rule is designed to prevent "accidental overrides")
```

In other words, when Nᵢ overrides a concrete member m defined in Nₒ, we must ensure that Nᵢ and Nₒ have a common base type which also declares m:

\[
\exists n \in N_i \quad \text{def } m ... \in mdecls(n) \quad \Rightarrow \quad \text{isAbstract}(m, N_i) \quad \text{noAccidentalOverride}(m, N_i, N_o)
\]

4 Typing

4.1 Subtyping and Well-Formedness

Most of the subtyping and well-formedness rules (Figures 2 and 3) are straightforward adaptations of the FGJ rules with S-CLASS and WF-CLASS generalized to handle traits. The

---

3To be precise, foo in Sub2 needs to be declared with the override keyword for A to compile, but we do not model this in our calculus: when translating code from PS into real Scala, override should be added everywhere it is legal to do so [Odersky et al. 2021, § 5.2.3].

4after adding override to the definition of foo in Sub2
\( \Delta \vdash S <: S \) [S-REFL]  \( \Delta \vdash X <: \Delta(X) \) [S-VAR]  
\[\begin{align*}
\Delta \vdash S <: T & \quad \Delta \vdash T <: U \\
\hline
\Delta \vdash S <: U & \quad \text{[S-TRANS]} \\
\hline
\end{align*}\]

\( P \in \text{parents}(N) \)
\( \Delta \vdash N <: P \) [S-CLASS]

\( \Delta \vdash T_1 <: T \)
\( \Delta \vdash T_1 & T_2 <: T \) [S-AND11]
\( \Delta \vdash T_2 <: T \)
\( \Delta \vdash T_1 & T_2 <: T \) [S-AND12]

\( \Delta \vdash T <: T_1 \quad \Delta \vdash T <: T_2 \)
\( \Delta \vdash T <: T_1 & T_2 \) [S-AND2]

**Figure 2.** Subtyping

\( \Delta \vdash \text{Object ok} \) [WF-OBJECT]

\( X \in \text{dom}(\Delta) \)
\( \Delta \vdash X \text{ ok} \) [WF-VAR]

\[
\begin{align*}
\left\{ \begin{array}{l}
\text{class} \\
\text{trait}
\end{array} \right\} C[X <: N] & \quad \Delta \vdash T \text{ ok} \\
\Delta \vdash \overline{T} \text{ ok} \\
\Delta \vdash [\overline{T}/X]N & \quad \text{[WF-CLASS]} \\
\hline
\Delta \vdash C[\overline{T}] \text{ ok} \\
\hline
\Delta \vdash T_1 \text{ ok} \\
\Delta \vdash T_2 \text{ ok} \\
\hline
\Delta \vdash T_1 & T_2 \text{ ok} \quad \text{[WF-AND]} \\
\end{align*}
\]

**Figure 3.** Well-formed types

rules for intersections however are lifted from the DOT calculus [Rompf and Amin 2016, Figure 1]. The subtyping relationship defined by these rules induces a partial order in which \( T_1 & T_2 \) is the greatest lower bound of \( T_1 \) and \( T_2 \).

Note that intersections in Scala are first-class types: they can appear in any position and members of an intersection can be arbitrary types, by contrast in Java the operands of the intersection cannot be type variables and the intersection itself can only appear in casts and upper-bounds of type parameters.

### 4.2 Typing Expressions

The typing rules for expressions (Figure 4) are also very close to the FGJ rules, but the helper functions getters (which corresponds to fields in FGJ) and mtype are different. Both of these functions take a subscript representing the type environment.

#### 4.2.1 getters \( \Delta(T) \): The List of Getters Accessible From a Value of Type \( T \)

We can read the getters of a class directly from its definition (we do not need to recurse on its parent class because GT-CLASS ensures that the class parameters of a well-formed class includes the class parameters of its parent):

\[
\begin{align*}
\text{getters}_\Delta(X) &= \text{getters}_\Delta(\Delta(X)) \\
\text{getters}_\Delta(\text{Object}) &= \bullet \\
\end{align*}
\]

\[
\begin{align*}
\text{class} C[X <: N](f : U) & \quad \text{getters}_\Delta(C[\overline{T}]) = [\overline{T}/X]f : U \\
\text{trait} C[...] & \quad \text{getters}_\Delta(C[\overline{T}]) = \bullet \\
\end{align*}
\]

In an intersection, the getters will usually be defined only on one side, but if they happen to be defined on both sides we can safely take the union without worrying about the same getter appearing with different types since GT-CLASS also ensures that we cannot inherit from multiple unrelated classes and that getters in sub-classes and super-classes have matching types:
What is the type of \( x.\text{foo()} \)? In Java (and FJ&\( \lambda \)) this would be an error, even though one can override both of these methods at once via covariant overriding. The problem is that there is no Java type representing the greatest lower bound of \( A \) and \( B \), whereas as we’ve seen above in Scala this is simply \( A \& B \). This means we can define:

\[
\text{mtype}_A(m, T_1) = [Y <: P] \rightarrow \text{mdecls}(N)
\]

4.3 Typing Declarations

Figure 5 lists the typing rules for declarations. Unlike FGJ override validation happens when typing the class (using the isValid judgment we defined in the previous section) and not the method since we need to check the validity of overrides defined in parents too. Unlike Scala, we do not perform override validation in traits since these checks are redundant with the ones done on the classes extending those traits, although in practice it’s of course better to find out about errors at the definition site rather than at the use site.

### 4.3.1 Checking for Abstract Methods in Classes

One might assume that a method is abstract in a class if there are no concrete implementation of this method among its base types. However, both Java and Scala 3 allow "re-abstracting" a method, for example in:

```
trait Base { def foo(): Object = ... }
trait Sub < Base { def foo(): Object }
class A < Object, Sub
class B < Object, Base, Sub
```

\( A \) and \( B \) have the same linearization so we’d expect them to be equivalent, but in fact an inherited method is considered abstract in a class if it is abstract among all the parents of this class, so \( A \) is not well-formed since it only inherits an abstract \( \text{foo} \) from \( \text{Sub} \). Although this concept exists in Java, it is not modeled in FJ&\( \lambda \) which does not allow an abstract method to override a concrete one.\(^5\) To represent this we define the mutually recursive \( \text{mnames}_{\text{abs}}(N) \) and \( \text{mnames}_{\text{con}}(N) \) to be the sets of names of respectively abstract and concrete members of \( N \):

\[
\text{mdecls}(N) = \text{def } m_{\text{abs}} ...; \text{def } m_{\text{con}} ... = ...
\]

\[ P = \text{parents}(N) \]

\[ \text{mnames}_{\text{con}}(N) = \text{mdecls}(N) \cup (\text{mnames}_{\text{con}}(P) \setminus \text{mdecls}(N)) \]

\[ \text{mnames}_{\text{abs}}(N) = \text{mdecls}(N) \cup (\text{mnames}_{\text{abs}}(P) \setminus \text{mnames}_{\text{con}}(P)) \]

\( \text{GT-CLASS} \) then takes care of checking that \( \text{mnames}_A \) is empty for proper classes. For convenience we also define the set of all method names in \( N \) as:

\[ \text{mnames}(N) = \text{mnames}_{\text{abs}}(N) \cup \text{mnames}_{\text{con}}(N) \]

5 Erasure

Our target calculus is FJ&\( \lambda \) without lambdas or intersections, we name the resulting fragment Featherweight Java with Default methods (FJD)\(^6\).

5.1 Type Erasure

Given a type environment \( \Delta \), we write |T|_\( \Delta \) for the type erasure of \( T \) which is defined in FGJ as:

\[
|X|_\Delta = |\Delta(X)|_\Delta
\]

\[ |C[...]|_\Delta = C \]

In general, we strive to have erasure preserve as much of the structure of the original program as possible to keep the translation simple and to allow interoperability between programs written in the source and target language. In particular, the mapping above preserves subtyping in FGJ: if

\(^5\)see the definition of \( \text{mdecls} \) in [Bettini et al. 2018, p. 15]

\(^6\)FJD was already taken by Featherweight Java with Inner classes [Igarashi and Pierce 2002].
\[
x : T \in \Gamma \\
\Delta, \Gamma \vdash x : T \\
\frac{\Delta, \Gamma \vdash e_0 : T_0 \text{ getters}_\Delta(T_0) = f : T}{\Delta, \Gamma \vdash e_0.f_j : T_j} \\
\frac{\Delta, \Gamma \vdash e_0 : T_0 \text{ mtype}_\Delta(m, T_0) = [\bar{V} <: \bar{P}] \rightarrow \bar{U} \rightarrow U_0}{\Delta \vdash \bar{V} \text{ ok} \quad \Delta \vdash \bar{V} <: [\bar{V}/\bar{Y}] \bar{P} \quad \Delta, \Gamma \vdash \bar{z} : \bar{S} \quad \Delta \vdash \bar{S} <: [\bar{V}/\bar{Y}] \bar{U}} \quad \frac{\Delta, \Gamma \vdash e_0, m[\bar{V}](\bar{z}) : [\bar{V}/\bar{Y}] U_0}{\Delta, \Gamma \vdash \bar{z} : \bar{S} \quad \Delta \vdash \bar{S} <: \bar{T}} \\
\Delta \vdash N \text{ ok} \quad \text{getters}_\Delta(N) = f : T \quad \Delta, \Gamma \vdash \bar{z} : \bar{S} \quad \Delta \vdash \bar{S} <: \bar{T} \quad \Delta, \Gamma \vdash \text{new } N(\bar{T}) : N \quad \frac{\Delta = X <: N, Y <: \bar{P}}{\Delta \vdash T, T, \bar{P} \text{ ok} \quad \Delta \vdash x : T, \text{ this } : C[X] \vdash e_0 : S \quad \Delta \vdash S <: T} \quad \frac{\text{def } m[\bar{Y} <: \bar{P}](x : \bar{T}) : T = e_0 \text{ OK IN } C[X <: N]}{\frac{\text{If } \text{mimpl}(m, C[X]) \text{ is defined then } \text{isVal}(m, C) \text{ must also be defined}}{\text{mnames}_{abs}(C) = \bullet}} \quad \frac{\text{class } C[X <: N](g : U, f : \bar{T}) \ll P(\bar{g}), Q \{\bar{M}\} \text{ OK}}{\frac{X <: N \equiv N, \bar{Q} \text{ OK}}{\frac{\text{trait } C[X <: N] \ll \bar{Q} \{\bar{H}, \bar{M}\} \text{ OK}}{\frac{\text{class } C[X <: N]}{\text{proper class}} \quad \text{interface } C[X <: N] \quad \text{abstract method}} \quad \text{concrete method}} \quad \text{class field}} \quad \text{variable}} \quad \text{expression}} \quad \text{object}} \quad \text{cast}} \\
\frac{\text{type}}{\text{class declaration}}
\]

Figure 4. Syntax Directed Typing Rules

\[
\frac{\Delta = X <: N, Y <: \bar{P} \quad \Delta \vdash T, T, \bar{P} \text{ ok} \quad \Delta \vdash x : T, \text{ this } : C[X] \vdash e_0 : S \quad \Delta \vdash S <: T}{\text{def } m[\bar{Y} <: \bar{P}](x : \bar{T}) : T = e_0 \text{ OK IN } C[X <: N]} \quad \frac{\text{If } \text{mimpl}(m, C[X]) \text{ is defined then } \text{isVal}(m, C) \text{ must also be defined}}{\text{mnames}_{abs}(C) = \bullet} \quad \frac{\text{class } C[X <: N](g : U, f : \bar{T}) \ll P(\bar{g}), Q \{\bar{M}\} \text{ OK}}{\text{trait } C[X <: N] \ll \bar{Q} \{\bar{H}, \bar{M}\} \text{ OK}}
\]

Figure 5. Declaration Typing Rules

Unfortunately, no matter how we erase intersection types, we cannot preserve subtyping in general because although \(T_1 \land T_2\) is the greatest lower bound of \(T_1\) and \(T_2\), there might not exist a specific type in FJD representing the greatest lower bound of \(T_1\land\|\land\) and \(T_2\land\|\land\). Nevertheless, since we’re trying to preserve as much structure as possible, it seems logical to define:

\[
[T_1 \land T_2]_{\|\land} = \text{erasedGlb}(\spadesuit, T_1, T_2) \quad \text{where} \quad \text{erasedGlb} \quad \text{always} \quad \text{returns} \quad \text{one} \quad \text{of} \quad \text{its} \quad \text{arguments.} \quad \text{In} \quad \text{fact} \quad \text{this} \quad \text{is} \quad \text{what} \quad \text{both} \quad \text{Java} \quad \text{and} \quad \text{Scala} \quad \text{do}, \quad \text{but} \quad \text{they} \quad \text{differ} \quad \text{on} \quad \text{the} \quad \text{implementation} \quad \text{of} \quad \text{erasedGlb}: \]

- Java simply defines \(\text{erasedGlb}(T_1, T_2) = T_1 \quad \text{[Gosling \ et \ al. \ 2015, \ § 4.6].} \quad \text{This} \quad \text{means} \quad \text{the} \quad \text{user} \quad \text{can} \quad \text{tweak} \quad \text{the} \quad \text{eraser} \quad \text{by} \quad \text{reordering} \quad \text{types} \quad \text{which} \quad \text{can} \quad \text{be} \quad \text{useful} \quad \text{for} \quad \text{evolving} \quad \text{code} \quad \text{in} \quad \text{a} \quad \text{binary-compatible} \quad \text{way.} \)

\[
\Delta \vdash S <: FGD T \text{ then } |S|_{\|\land} <: FJ |T|_{\|\land} \quad \text{(Lemma A.3.5}_{FDG} \text{) which reduces the amount of casts that need to be inserted when erasing expressions to a minimum (Theorem 4.5.3}_{FDG})).
\]

\[
\text{Technically, subtyping would be preserved if we erased all types to Object, but this would defeat the point since it would require many more casts in expression erasure and impede interoperability between Scala and Java.}
\]
\[ |x|_{\Delta, \Gamma} = x \text{ [E-VAR]} \]
\[ \Delta; \Gamma \vdash e_0 : T_0 \quad |T_0|_{\Delta} = C \quad |e_0|_{\Delta, \Gamma} = f \quad \text{[E-FIELD]} \]
\[ \Delta; \Gamma \vdash e_0 : T_0 \quad \text{erasedReceiver}_{\Delta}(m, T_0) = C \quad \text{ntype}_{\Delta}(m, C, \Gamma) = \overline{U}_0 \quad e'_i = |e_i|_{\Delta, \Gamma} \quad |N|_{\Delta} = C \quad \text{fields}_{\Delta}(C) = f : T \quad e'_i = |e_i|_{\Delta, \Gamma} \quad \text{[E-INVK]} \]

**Figure 7. Expression Erasure**

\[ \Gamma = \overline{x : T}, \text{this : C} | \overline{x} \quad \Delta = \overline{X <: N, Y <: P} \quad \text{[E-METHOD]} \]
\[ \text{def } m | \overline{Y : P} | \overline{x : T} : T_0 \equiv e_0 | \overline{X < N, C} = |T_0|_{\Delta} m_C | \overline{T} x \{ \text{return } |e_0|_{\Delta, \Gamma} \} ; \]
\[ \Delta = \overline{X <: N} \quad K = C | \overline{U} g, \overline{T} f \{ \text{super } \overline{g}; \text{this } \overline{T} = f \} ; \quad \text{M} = |M|_{\Delta, C} \cup \{ \text{bridges } (m, C) \mid m \in \text{names } (C) \} \quad \text{[E-CLASS]} \]
\[ \text{class } C | \overline{X <: N} (g : U, f : T) < p | \overline{Q} \{ M \} = \text{class } C < |P|_{\Delta}, |Q|_{\Delta} \{ |T|_{\Delta} f ; K ; M \} \quad \text{[E-METHOD]} \]
\[ \Delta = \overline{X <: N} \quad \text{M} = |M|_{\Delta, C} \quad |\text{trait } C | \overline{X <: N} < \overline{Q} \{ M \} = \text{interface } C < |Q|_{\Delta} \{ M \} \quad \text{[E-TRAIT]} \]

**Figure 8. Class Table Erasure**

- On the other hand, Scala 2 defines `erasedGlb` to prefer subtypes over supertypes (thus actually returning the greatest lower bound of the erased types) and proper classes over traits (because both casting and method call are usually faster on classes than on interfaces [Click and Rose 2002; Shipilev 2020]). Unfortunately, completely specifying the behavior of Scala 2 here is extremely hard because it inadvertently depends on implementation details of the compiler.

- Scala 3 preserves the two properties from Scala 2 mentioned above and additionally ensures that erasure preserves commutativity of intersection (\(|T_1 \& T_2|_{\Delta} = |T_2 \& T_1|_{\Delta}|) by applying a tie-break based on the lexicographical order of the names of the compared types. The following pseudo-code accurately specifies its behavior:

```
1 def erasedGlb(tp1: Type, tp2: Type): Type =
2   if tp1.isProperClass && !tp2.isProperClass then
3     return tp1
4   if tp2.isProperClass && !tp1.isProperClass then
5     return tp2
6   if tp1 <: tp2 then return tp1
7   if tp2 <: tp1 then return tp2
8   if tp1.name <= tp2.name then tp1 else tp2
```

The Scala 3 algorithm preserves most interesting properties of intersections but has one non-obvious shortcoming: it does not preserve associativity, consider:

\[ \text{trait } X; \text{trait } Y; \text{trait } Z \text{ extends } X \]

Then \(|X \& Y \& Z| = Z \text{ but } |X \& (Y \& Z)| = X\). The problem is that while the lexicographic ordering by itself is total, it is applied inconsistently because incompatibility of subtyping is not transitive: in our example neither \(X <: Y \) nor \(Y <: X\) making \(X\) and \(Y\) incomparable, but even though \(Y\) and \(Z\) are also incomparable it is not true that \(X\) and \(Z\) are incomparable.

To rectify this we propose ordering classes by the number of base types they have. In other words, we replace the subtyping checks on lines 6 and 7 in the listing above by:

```
val relativeLength = L(tp1).length - L(tp2).length
if relativeLength > 0 then return tp1
```

See [https://github.com/lampepl/dotty/blob/master/compiler/src/dotty/tools/dotc/core/TypeErasure.scala](https://github.com/lampepl/dotty/blob/master/compiler/src/dotty/tools/dotc/core/TypeErasure.scala) for the unsavory details.

\(^8\)The complete implementation also special-cases value types and array types which we do not model in our calculus, see `erasedGlb` in [https://github.com/lampepl/dotty/blob/master/compiler/src/dotty/tools/dotc/core/TypeErasure.scala](https://github.com/lampepl/dotty/blob/master/compiler/src/dotty/tools/dotc/core/TypeErasure.scala)

\(^9\)Since this change would break binary compatibility, it will have to wait until the next major version of Scala.
if relativeLength < 0 then return tp2

This still means we prefer subtypes over supertypes since a subclass necessarily has more base types than any of its parent, but incomparability is now transitive which is enough to make erasedGl1b itself transitive.

In the rest of this section, we will assume erasedGl1b prefers classes over traits as well as subtypes over supertypes but otherwise will stay independent of any particular implementation.

5.2 Expression Erasure

Because type erasure does not preserve subtyping we might need to insert casts both on prefixes of calls as well as on method arguments. To keep the typing rules in Figure 7 readable, we delegate casting $|e|_{\Delta_\Gamma}$ to $T$ to an auxiliary judgment $|e|^T_{\Delta_\Gamma}$ which is mutually recursive with the main judgment:

$$|e|^T_{\Delta_\Gamma} = \begin{cases} e' & \text{ if } S \leq\text{FC} T \\ (T)e' & \text{ otherwise} \end{cases}$$

Casting the prefix of a getter call to the appropriate type is easy: we know that erasedGl1b will always return the most specific class type in an intersection and that traits do not contain getters, therefore if geterss$_\Gamma(T_\Delta) = f : T$ then fields$_\text{FC}(|T|_\Delta) = f : |T|_\Delta$ and E-FIELD is straightforward, but finding the right cast for the receiver of a method call is more involved.

5.2.1 erasedReceiver$_\Delta(m, N)$: The First Erased Parent Type Where $m$ Is Defined.

Given $x : L \& R$ and the class table:

trait L { def l(): Object }

trait R { def r(): Object }

Then the type of $|x|_{\Delta_\Gamma}$ will be either $L$ or $R$ (depending on the definition of erasedGl1b), but that means that one of $x.1()$ and $x.r()$ will require casting the receiver, therefore E-INVK relies on the following auxiliary function:

$$\text{erasedReceiver}_\Delta(m, X) = \text{erasedReceiver}_\Delta(m, \Delta(X))$\
\text{erasedReceiver}_\Delta(m, C[...]) = C$\
\text{erasedReceiver}_\Delta(m, T_1 \& T_2) =$$

$$\begin{cases} \text{erasedReceiver}_\Delta(m, T_1) & \text{ if } \text{type}_\Delta(m, T_1) \text{ is defined} \\ \text{erasedReceiver}_\Delta(m, T_2) & \text{ otherwise} \end{cases}$$

Additionally, erasure does not preserve method names: $m$ is erased to $m_C$ where $C$ is the type of the receiver, this is justified in the following section.

5.3 Class Table Erasure

Given the class table:

trait X; class Y extends X

trait L[T] { def foo(): T }

trait R[T <: X] { def foo(): T }

class A < Object, L[Y], R[Y] { 
  def foo(): Y = new Y }

One might hope we could erase it just by erasing each type and expression appearing in it:

interface L { Object foo() }

interface R { X foo() }

class A < Object, L, R { 
  Y foo() { return new Y(); } 
}

But that would be incorrect: a method in FJD must have exactly the same type as the methods it overrides (just like in Java bytecode). Compilers normally handle this by generating synthetic bridge methods [Bracha et al. 2003]:

interface L { Object foo() }

interface R { X foo() }

class A < Object, L, R { 
  Y foo() { return overloads of foo returning Y(); } 
  X foo() { return <overloads of foo returning Y>(); } 
}

Notice that the types of the new methods added in $A$ match the types of the overridden methods in $L$ and $R$ and simply forward to the actual implementation of $\text{Foo}$ in $A$, thus restoring the semantics present in the source program. But we cannot directly reuse this technique since our target calculus does not support overloading, faced with the same problem FGJ adopted the following strategy:

In [Generic Java], the actual erasure is somewhat more complex, involving the introduction of bridge methods [...] instead, the rule E-METHOD merges two methods into one by inline-expanding the body of the actual method into the body of the bridge method.

But this works because FGJ only supports single-class inheritance, whereas in the example above we need two bridges in $A$ corresponding to the two traits containing an overridden $\text{Foo}$. Like FGJ, we shy away from introducing overloading in our target calculus and instead employ the following scheme:

- When erasing a call to $m$, we replace it by a call to $m_C$ where $C$ is the erased receiver of $m$ (see the previous section).
- When erasing the declaration of $m$ in $C$, we rename it to $m_C$.
- When erasing a class $C$, we add enough bridge methods so that erased calls to $m$ always end up being forwarded to the implementer of $m$ in $C$.

For our example this means we get:

interface L { Object foo(); } 

interface R { X foo(); } 

class A < Object, L, R { 
  Y foo() { return new Y(); } 
}
This scheme wouldn’t be practical in a real compiler since it would make it much harder for Java and Scala code to interoperate, but as a model we believe it’s close enough to the real thing to be useful. The exact rules are described in Figure 8 which makes use of the following judgments:

\[
\begin{align*}
& \text{mtype}_{\text{FD}}(m_E, E) = \overline{T} \rightarrow T_0 \\
& \text{mtype}_{\text{FD}}(m_D, D) = U \rightarrow U_0 \\
& x_i = \text{this}.m_D(e) \quad e_i = \begin{cases} x_i & \text{if } T_i = U_i \\ (U_i)x_i & \text{otherwise} \end{cases} \\
& \text{bridge}(m_E, m_D) = T_0 m_E(\overline{T}x) \{\text{return } e_0;\} \\
& \text{minimpl}(m, N) = D[\overline{T}] \\
& \overline{E[...]} = \{n \in \mathcal{L}(N) \land D[\overline{T}] | \text{def } m... \in \text{mdecls}(n)\} \\
& \text{bridges}(m, N) = \text{bridge}(m_E, m_D)
\end{align*}
\]

Note that this definition of \textit{bridges} can generate unnecessary bridges since it does not take into account that a parent class might already have defined an equivalent bridge.

### 6 Related Work

#### 6.1 Multiple Inheritance and the Diamond Problem

What should happen when multiple matching methods from unrelated classes are inherited? There is no standard solution here but languages usually pick one of the following approaches:

- In Java and C++ with virtual inheritance, the class definition is considered invalid and an error is emitted.
- In C++ with non-virtual inheritance, the ambiguity resolution is delayed until the method call site, where the user can upcast the receiver to manually resolve the ambiguity. See [Wasserrab et al. 2006] for a precise treatment of inheritance in C++ including a soundness proof but make sure to prepare a pot of coffee first.
- Several languages like Scala will attempt to determine a linearization order for the parent classes and use that to resolve the ambiguity. The C3 linearization algorithm [Barrett et al. 1996] originally defined for Dylan is especially popular, being notably used by Python and Raku. This form of linearization is guaranteed to be monotonic: two classes will always appear in the same order in any given linearization, this isn’t true in Scala when traits are involved which lets us define class hierarchies more freely at the cost of making linearization harder to reason about.

#### 6.2 Related Calculi

**Featherweight Java** was first extended with interfaces and intersection types faithful to Java semantics in FJ&\& [Bettini et al. 2018]. The semantics of intersection types were then generalized beyond what Java supports in FJP&\& [Dezani-Ciancaglini et al. 2019] to allow intersections in any position (like Scala) and not just as target of casts, finally [Dezani-Ciancaglini et al. 2020] showed how to erase FJP&\& into FJ&\&. Pathless Scala can be seen as a generalization of FJP&\&, but we found it easier to extend FGJ with traits and intersections rather than to extend FJP&\& with polymorphism and generalize its interfaces to traits. However, FJ&\& stripped of intersections and lambdas makes for a great target calculus as it closely models most of the important aspects of Java bytecode, although we would really need to extend it with overloading to describe Scala’s erasure faithfully.

**Featherweight Scala** (FS) [Cremet et al. 2006] is not an extension of Featherweight Java despite its name: it does have nominal classes (including inner classes) but uses type members and path-dependent types rather than type parameters and is therefore more closely related to DOT. FS also includes multiple inheritance via traits, but it does not precisely model the overriding rules of Scala like we do in Section 3.

DOT was first described in [Amin et al. 2012] but wasn’t proved sound until [Amin et al. 2016], although this version of DOT lacked union and intersection types. A soundness proof for DOT with intersections was then presented in [Rompf and Amin 2016], and unions finally made a comeback in [Giarrusso et al. 2020]. DOT has also been extended in multiple ways to bring it closer to Scala [Kabir and Lhoták 2018; Rapport 2019; Stucki and Giarrusso 2021] but the gap between the two remains large.

### 7 Conclusion and Future Work

We have presented Pathless Scala, a convenient calculus for formalizing the semantics and compilation schemes of parts of Scala which we found to be understudied. In particular, we believe its important to specify language features and their erasure together rather than leaving the latter as an implementation detail. They inevitably leak to the user (e.g., via Java reflection) and interoperability (of Scala 2 code with Scala 3 code, or of Scala code with Java code) requires the same type to be erased in the same way by multiple different compilers. We know from having had to reverse-engineer how Scala erasure works that this can end up being much harder than it needs to be. Therefore, we are particularly interested in extending Pathless Scala to cover other aspects of Scala with non-trivial erasures such as union types or polymorphic function types. Eventually, this could serve as a basis for a more precise version of the Scala Language Specification [Odersky et al. 2021].

In this work we’ve focused on erasing Scala types into “bytecode Java” types, but in practice we also need to worry about erasing Scala types into “source Java” types: the bytecode format defines a Signature attribute [Lindholm et al. 2015, § 4.7.8] which lets us specify a polymorphic Java...
method signature that will be ignored by the JVM at runtime but used by the Java compiler for typechecking, thus improving the interoperability between Scala and Java. It would be useful to specify an erasure from PS into full FJ&\lambda as a way to model this process. The Java compiler will also use this attribute if it is available to compute the erased signature it will emit when invoking the method, therefore we should also define an erasure of FJ&\lambda into FJD based on the semantics of Java erasure and verify that the composition of these two mappings are equivalents to the erased mapping of PS into FJD to avoid issues such as https://github.com/scala/bug/issues/4214.

We did not define evaluation semantics for PS, instead we described erasure rules to a simpler calculus known to be sound. For the sake of rigor, it would be good to follow the FGJ model: give evaluation rules to our calculus independent of its erasure, prove soundness, and show that directly evaluating a PS program is equivalent to erasing and then evaluating it. Given that our calculus intentionally excludes the hard parts of DOT, we believe that the existing proofs given in the FJ paper can be extended in a straight-forward way to achieve this, but we have not completed this work yet.

Of course, eventually we should also strive to reconcile Pathless Scala and DOT, but that is likely to be a much longer-term project given how difficult it has been to extend the meta-theory of DOT so far, meanwhile the rest of Scala awaits us!

References


