Lightweight linear types in System F°

Karl Mazurak  
*University of Pennsylvania*

Jianzhou Zhao  
*University of Pennsylvania*

Stephan A. Zdancewic  
*University of Pennsylvania, stevez@cis.upenn.edu*

Follow this and additional works at: [http://repository.upenn.edu/cis_papers](http://repository.upenn.edu/cis_papers)

Part of the [Computer Sciences Commons](http://repository.upenn.edu/cis_papers)

**Recommended Citation**

doi>10.1145/1708016.1708027
© ACM, 2010. This is the author’s version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in *ACM SIGPLAN International Workshop on Types in Languages Design and Implementation*, ((2010))
http://doi.acm.org/10.1145/1708016.1708027* Email permissions@acm.org

This paper is posted at ScholarlyCommons. [http://repository.upenn.edu/cis_papers/591](http://repository.upenn.edu/cis_papers/591)
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).
Lightweight linear types in System F°

Abstract
We present F°, an extension of System F that uses kinds to distinguish between linear and unrestricted types, simplifying the use of linearity for general-purpose programming. We demonstrate through examples how F° can elegantly express many useful protocols, and we prove that any protocol representable as a DFA can be encoded as an F° type. We supply mechanized proofs of F°’s soundness and parametricity properties, along with a nonstandard operational semantics that formalizes common intuitions about linearity and aids in reasoning about protocols.

We compare F° to other linear systems, noting that the simplicity of our kind-based approach leads to a more explicit account of what linearity is meant to capture, allowing otherwise-conflicting interpretations of linearity (in particular, restrictions on aliasing versus restrictions on resource usage) to coexist peacefully. We also discuss extensions to F° aimed at making the core language more practical, including the additive fragment of linear logic, algebraic datatypes, and recursion.

Disciplines
Computer Sciences

Comments

doi>10.1145/1708016.1708027

© ACM, 2010. This is the author’s version of the work. It is posted here by permission of ACM for your personal use. Not for redistribution. The definitive version was published in ACM SIGPLAN International Workshop on Types in Languages Design and Implementation, {(2010)} http://doi.acm.org/10.1145/1708016.1708027 Email permissions@acm.org

This other is available at ScholarlyCommons: http://repository.upenn.edu/cis_papers/591
Lightweight Linear Types in System F°

Karl Mazurak  Jianzhou Zhao  Steve Zdancewic
University of Pennsylvania
{mazurak,jianzhou,stevez}@cis.upenn.edu

Abstract

We present System F°, an extension of System F that uses kinds to distinguish between linear and unrestricted types, simplifying the use of linearity for general-purpose programming. We demonstrate through examples how System F° can elegantly express many useful protocols, and we prove that any protocol representable as a DFA can be encoded as an F° type. We supply mechanized proofs of System F°’s soundness and parametricity properties, along with a nonstandard operational semantics that formalizes common intuitions about linearity and aids in reasoning about protocols.

We compare System F° to other linear systems, noting that the simplicity of our kind-based approach leads to a more explicit account of what linearity is meant to capture, allowing otherwise-conflicting interpretations of linearity (in particular, restrictions on aliasing versus restrictions on resource usage) to coexist peacefully. We also discuss extensions to System F° aimed at making the core language more practical, including the additive fragment of linear logic, algebraic datatypes, and recursion.

Categories and Subject Descriptors D.3.3 [Programming Languages]: Language Constructs and Features

General Terms Design, Languages, Theory

Keywords Linear logic, Polymorphism, Type systems

1. Introduction

Linear logic [15, 16] models resource usage by restricting the properties of contraction (the ability to duplicate a resource) and weakening (the ability to discard a resource). In the context of programming languages, ideas from linear logic were quickly adopted, at first to eliminate garbage collection [21] and shortly thereafter to handle mutable state [29].

Since their introduction, variants, refinements, and improvements on linear type systems have been proposed for many applications, including explicit memory management and control of aliasing [2, 13, 17, 28, 36], capabilities [9, 10], and tracking state changes for program analysis [11, 32]. Of particular interest is work on typestate, which ensure that a sequence of API calls is well-behaved [13, 12], and on session types, which check that the endpoints of a channel agree on the next message to be sent or received [18, 25, 27]. Walker has a more comprehensive survey [33].

Given these success stories, it is perhaps surprising that we have yet to see general linear types seriously considered for inclusion in a mainstream functional programming language.¹ But alas, linear types can easily lead to awkward programming models and potentially complicated language designs that are difficult both to implement and to program with. This paper seeks to address these issues by introducing System F°—pronounced “F-pop”—a language that is intended to be a simple foundation for practical linear programming. Rather than aiming at one particular problem, System F° lets programmers enforce their own protocol abstractions through the power of linearity and polymorphism, yet its typing discipline is lightweight enough to expose in a surface language. System F° is simply the Girard–Reynolds polymorphic λ-calculus [14, 23] extended with two base kinds: ⋆, classifying ordinary, unrestricted types, and ◦, classifying linear types. A subkinding relation ⋆ ≤ ◦ makes explicit the observation that values of unrestricted types may safely be treated linearly; i.e., since variables of unrestricted type may be used any number of times and linear variables must be used exactly once, it is always safe to use unrestricted variables as though they were linear. Any System F° expression with kinds erased is a well-typed System F expression.

In introducing System F°, this paper contributes the following:

- The design of System F°, and in particular its use of kinds and kind subsumption, which is lightweight and structured so as to integrate well with existing functional languages. (Section 2)
- Mechanized proofs of standard soundness and parametricity results for System F°, which ensure that the properties functional programmers rely on continue to hold. (Section 2.1)
- A second, linearity-aware semantics for System F°, which formalizes common intuitions about linearity and shows that these intuitions do indeed hold. (Section 4)
- Several examples—including all regular languages—along with extensions and proposals for compiler support, remarkable primarily for their simplicity, which showcase System F°’s potential usefulness. (Sections 3 and 5)

In the rest of this section we discuss the design of System F° in the context of prior work, and we showcase System F°’s ability to enforce programmer-defined protocols with a familiar example.

1 Clean is often seen as the exception to this, but there are subtle differences between its uniqueness types, which concern aliasing, and standard linear types (see Section 3.1).
semantics and programming model of System F. This division between linear and unrestricted types is inspired by that proposed for monomorphic systems by Wadler [29] and Benton [5]. Here we contrast our approach with some alternatives from the literature.

Broadly speaking, there are two other approaches to distinguishing linear from unrestricted variables. The first approach, which closely follows linear logic, is to treat all types as linear and introduce the modal constructor ! to account for unrestricted terms, which must be closed with respect to linear variables. This interacts easily enough with polymorphism, since all types are treated uniformly—or any type can be substituted for any type variable.

Unfortunately, assuming linearity by default requires very explicit handling of unrestricted terms; the standard approach uses a let ! construct to introduce unrestricted variables from suspended computations ≤ [6, 30]. Because unrestricted values are common in practice, this can prove quite cumbersome, and the burden extends even to base types—the constant 3 would be of type Int, not !Int! Further, full inference of !s has been shown to be impossible [22].

A second approach distinguishes linear from unrestricted types by means of qualifiers lin (for “linear”) and un (for “unrestricted”) applied to a collection of pre-types [1, 33]. These qualifiers (which do not nest) constrain usage (or aliasing), while the pre-types determine the introduction and elimination forms of values. This separation facilitates implicit copying and discarding for unrestricted types, yielding a less burdensome programming model.

Type qualifiers, however, have more complex interactions with polymorphism. To retain both soundness and expressivity, one is led to introduce quantification over qualifiers, pre-types, and types independently. This quickly leads to large and complex types; for instance, the type of plus has five qualifiers:

\[
\text{plus} : (q_1 \text{Int}) \to \text{lin} (q_2 \text{Int}) \to \text{lin} (q_3 \text{Int})
\]

The relationships among such qualifiers are often nontrivial (e.g. \(q_1\) should be lin if \(q_2\) is), which can be captured (at the expense of additional complexity) by qualifier-level bounded quantification. Qualifiers thus ease the use of unrestricted types but are too unwieldy for a polymorphic source language—indeed, others have argued against qualifiers even for intermediate languages [34].

Our use of kinds in lieu of ! or type qualifiers strikes a good balance on these issues. As with qualifiers, programming with unrestricted types is natural; as with !, polymorphism remains simple. Subkinding also plays well with base types: 3 has type Int, which has kind \(\text{lin}\) (and, by subsumption, \(\circ\)), while the type of plus is the simple Int \(\text{lin} \to \text{lin}\). We thus have flexible polymorphic types without the need for bounded quantification or other complexities of subtyping in a higher-order setting.

Closest in design to System F\(^\circ\) is probably Ahmed, Fluet, and Morrisett’s language for substructural state [1], though they use qualifiers as described above. Due to their focus on aliasing in mutable state, their language does not admit subtyping, which would be the analog in their setting of our subkinds. Our subkind relation \(\ast \leq \circ\) agrees with the interpretation of linearity as related to usage; it does not reflect linearity as alias-freedom; in which linear types are analogous to uniqueness types [17, 28]. Nevertheless, we show in Section 3.1 that matters of aliasing can indeed be addressed in System F\(^\circ\) given an appropriate representation of references.

Ahmed et al. also admit affine types, for which only weakening is permitted, and relevant types, which allow only contraction. Such concepts would fit well with our subkinding relation, but would increase the complexity of typing context management. Other systems have considered notions of usage much more fine-grained than ours [19], but the types in such systems can quickly become overwhelming if exposed to the programmer. As we are incorporating the full power of System F\(^\circ\) exposing at least some types is unavoidable; we also believe that System F\(^\circ\) convincingly demonstrates that just the simple distinction between linear and unrestricted types has much to offer.

1.2 System F\(^\circ\) by example: Types for filehandles

Linearity lets us specify a filesystem interface that requires a filehandle to be closed exactly once and forbids its use thereafter. A first approximation\(^2\) for an idealized linear filesystem might be

\[
\begin{align*}
\text{FileHandle} & : \circ \\
\text{open} : \text{String} \to \text{FileHandle} \\
\text{read} : \text{FileHandle} \to \text{Char, FileHandle} \\
\text{write} : \text{Char} \to \text{FileHandle} \to \text{FileHandle} \\
\text{close} : \text{FileHandle} \to \text{Unit}
\end{align*}
\]

\[
\begin{align*}
\text{readLine} : \text{FileHandle} \to \text{String, FileHandle}
\end{align*}
\]

Here open, read, write, and close are intended to be primitive operations over the linear (note the kind ascription) type FileHandle, while readLine is one of many library functions defined to make file access more convenient. The \(\ast\) decorating the arrow in a type like open \(\to\) FileHandle indicates that open is an unrestricted function, which may be used more than once; each time open is invoked, it will return a new FileHandle value that must be used linearly. Unrestricted functions are also free to take arguments of linear type, as can be seen in the other operations. Similarly, a linear function of type \(\tau_1 \to \tau_2\) should be invoked exactly once, but this has no bearing on the kind of either \(\tau_1\) or \(\tau_2\).

Because operations like read and write consume a FileHandle as input and return a linear FileHandle as output, and because such values cannot be duplicated, client programs are forced to sequence the calls to these functions, fixing their order of evaluation. Since FileHandle values cannot be discarded, the program—unless its overall type indicates that it contains a filehandle—must eventually use close to dispose of any FileHandle it has created. Linearity ensures that no aliased or duplicated FileHandles representing closed files remain to be improperly accessed. Clients of this interface are thus constrained, after opening a file, to access that file according to the regular protocol (read/write/\(\ast\)) close

Unfortunately, today’s operating systems do not understand linearity and instead provide unrestricted interfaces, which make no guarantees about correct filehandle usage:

\[
\begin{align*}
\text{UnsafeFH} & : \ast \\
\text{unsafeOpen} & : \text{String} \to \text{UnsafeFH} \\
\text{unsafeRead} & : \text{UnsafeFH} \to \text{Char} \\
\text{unsafeWrite} & : \text{Char} \to \text{UnsafeFH} \to \text{Unit} \\
\text{unsafeClose} & : \text{UnsafeFH} \to \text{Unit}
\end{align*}
\]

System F\(^\circ\) makes it easy to create a safe interface protecting the above from misuse. First, for \(\alpha\) of kind \(\circ\)—our abstract representation of an actual filehandle—we define a record of safe file operations:

\[
\begin{align*}
\text{File}(\alpha) & = \{ \text{read} : \alpha \to \text{Char}, \alpha, \\
\text{write} : \text{Char} \to \alpha \to \alpha, \\
\text{close} : \alpha \to \text{Unit} \}
\end{align*}
\]

\(^2\) This example uses roughly the same interface given by DeLine and Fähndrich to motivate Vault [11] and discussed by Kiselyov and Shan [20] as an alternative to their filehandle regions.
Now we can define open to return both the hidden filehandle and its associated operations:

```
open : String → ∀α:⊤. (α, File(α))
open = λx:.String. unsafeOpenFile (unsafeRead h, h), write = λx:.Char. unsafeWrite h; h,
close = unsafeClose
```

Note that, while the type `UnsafeFH` is unrestricted within the scope of `open`, the outside world sees its occurrences at the existentially bound linear type variable `α`. If `open` treats filehandles correctly, then any use of an existential package created by `open` must treat them correctly as well.

We no longer read, write, and close as separate functions, but library functions like `readLine`, `write`, and `close` also serve as fileoperations, not the filehandle itself, and by separating out unpacks and packs from calls to both primitive operations and library functions, our types can reflect the fact that the filehandle returned by `readLine` is the same one that it was given. Writing our functions in this way allows for useful type coercions; for example, suppose we are also able to open files in read-only mode, resulting in restricted existencial packages of the form

```
ROFile(α) = { read : α → (Char, α),
             close : α → Unit }
```

Many functions, including `readLine`, may be defined over this weaker interface, and we can always construct a record of type `ROFile(α)` out of a record of type `File(α)` to allow a read-write filehandle to be treated as though it were read-only.

Of course, such a filesystem interface is of little use if it is too cumbersome for everyday programming. In Section 5.2 we discuss a weaker interface, and we can always construct a record of type `File(α)` out of a record of type `File(α)` to allow a read-write filehandle to be treated as though it were read-only.

2. System F^0 Defined

The syntax of System F^0 is given in Figure 1; Figure 2 shows its call-by-value (and kind-agnostic) operational semantics, which is completely standard. Type variables are annotated with their kinds when bound (by `λ` in expressions or by `∀` in types); kinds also appear in functions (λx:τ. e) and function types (τ1 → τ2). F^0 adopts the value restriction [35], permitting type abstraction only over values, for reasons that will become clear later.

Typing and kinding rules for System F^0, along with auxiliary judgments, are given in Figure 3. Our typing rules take a linear typing context Δ, binding only term variables, in addition to the standard unrestricted context Γ, which binds both type and term variables; following Barber’s DILL [3], this greatly simplifies the mechanization of our soundness proofs. At the kind level, note that the rule K-ARROW gives τ1 → τ2 the kind κ regardless of the kinds of τ1 and τ2. As described in Section 1.2, this decouples the notion of a linear function (i.e., of type τ1 → τ2), which must be used exactly once, from a function that takes a linear argument (i.e., where τ1 has kind ι), which must use its argument exactly once. Rule K-ALL, by contrast, simply gives ∀α:κ. τ the same kind as τ—this is a design choice made in the interests of keeping F^0 simple, as little seems to be gained by allowing these kinds to differ; this choice is also compatible with a type-erasuere interpretation.

Type application must check that the kind of the supplied type argument is compatible with the kind of the variable for which it will be substituted. As mentioned in Section 1, we see linearity as constraining the permitted usage of a variable, which means that it should always be safe to replace a linear type variable with an unrestricted type; we thus allow subbinding via the rule K-SUB. Substitution turns out to be key in many useful examples: for instance, it is what allowed the unrestricted type `UnsafeFH` to be protected by a linear interface.

Weakening—neglecting to use a variable—and contraction—using a variable more than once—should only be possible with variables of unrestricted type. Rather than add explicit contraction and weakening rules, we have built these properties into the rules that require them. The separation of linear and unrestricted typing contexts makes this fairly straightforward; rules T-LVAR and T-UVAR permit weakening by allowing an arbitrary Γ at the leaves of typing derivations, while rule T-APP duplicates Γ but splits Δ via the ⊣ relation. We can thus prove:

4 The lemmas in this section of the paper have all been proved in Coq; the source scripts of our proofs are available from the last author’s web pages.

```
... typing context Δ, binding only term variables, in addition to the standard unrestricted context Γ, which binds both type and term variables; following Barber’s DILL [3], this greatly simplifies the mechanization of our soundness proofs. At the kind level, note that the rule K-ARROW gives τ1 → τ2 the kind κ regardless of the kinds of τ1 and τ2. As described in Section 1.2, this decouples the notion of a linear function (i.e., of type τ1 → τ2), which must be used exactly once, from a function that takes a linear argument (i.e., where τ1 has kind ι), which must use its argument exactly once. Rule K-ALL, by contrast, simply gives ∀α:κ. τ the same kind as τ—this is a design choice made in the interests of keeping F^0 simple, as little seems to be gained by allowing these kinds to differ; this choice is also compatible with a type-erasuere interpretation.
```

```
2. System F^0 Defined

The syntax of System F^0 is given in Figure 1; Figure 2 shows its call-by-value (and kind-agnostic) operational semantics, which is completely standard. Type variables are annotated with their kinds when bound (by `λ` in expressions or by `∀` in types); kinds also appear in functions (λx:τ. e) and function types (τ1 → τ2). F^0 adopts the value restriction [35], permitting type abstraction only over values, for reasons that will become clear later.

Typing and kinding rules for System F^0, along with auxiliary judgments, are given in Figure 3. Our typing rules take a linear typing context Δ, binding only term variables, in addition to the standard unrestricted context Γ, which binds both type and term variables; following Barber’s DILL [3], this greatly simplifies the mechanization of our soundness proofs. At the kind level, note that the rule K-ARROW gives τ1 → τ2 the kind κ regardless of the kinds of τ1 and τ2. As described in Section 1.2, this decouples the notion of a linear function (i.e., of type τ1 → τ2), which must be used exactly once, from a function that takes a linear argument (i.e., where τ1 has kind ι), which must use its argument exactly once. Rule K-ALL, by contrast, simply gives ∀α:κ. τ the same kind as τ—this is a design choice made in the interests of keeping F^0 simple, as little seems to be gained by allowing these kinds to differ; this choice is also compatible with a type-erasuere interpretation.

Type application must check that the kind of the supplied type argument is compatible with the kind of the variable for which it will be substituted. As mentioned in Section 1, we see linearity as constraining the permitted usage of a variable, which means that it should always be safe to replace a linear type variable with an unrestricted type; we thus allow subbinding via the rule K-SUB. Substitution turns out to be key in many useful examples: for instance, it is what allowed the unrestricted type `UnsafeFH` to be protected by a linear interface.

Weakening—neglecting to use a variable—and contraction—using a variable more than once—should only be possible with variables of unrestricted type. Rather than add explicit contraction and weakening rules, we have built these properties into the rules that require them. The separation of linear and unrestricted typing contexts makes this fairly straightforward; rules T-LVAR and T-UVAR permit weakening by allowing an arbitrary Γ at the leaves of typing derivations, while rule T-APP duplicates Γ but splits Δ via the ⊣ relation. We can thus prove:

4 The lemmas in this section of the paper have all been proved in Coq; the source scripts of our proofs are available from the last author’s web pages.

```
2. System F^0 Defined

The syntax of System F^0 is given in Figure 1; Figure 2 shows its call-by-value (and kind-agnostic) operational semantics, which is completely standard. Type variables are annotated with their kinds when bound (by `λ` in expressions or by `∀` in types); kinds also appear in functions (λx:τ. e) and function types (τ1 → τ2). F^0 adopts the value restriction [35], permitting type abstraction only over values, for reasons that will become clear later.

Typing and kinding rules for System F^0, along with auxiliary judgments, are given in Figure 3. Our typing rules take a linear typing context Δ, binding only term variables, in addition to the standard unrestricted context Γ, which binds both type and term variables; following Barber’s DILL [3], this greatly simplifies the mechanization of our soundness proofs. At the kind level, note that the rule K-ARROW gives τ1 → τ2 the kind κ regardless of the kinds of τ1 and τ2. As described in Section 1.2, this decouples the notion of a linear function (i.e., of type τ1 → τ2), which must be used exactly once, from a function that takes a linear argument (i.e., where τ1 has kind ι), which must use its argument exactly once. Rule K-ALL, by contrast, simply gives ∀α:κ. τ the same kind as τ—this is a design choice made in the interests of keeping F^0 simple, as little seems to be gained by allowing these kinds to differ; this choice is also compatible with a type-erasuere interpretation.

Type application must check that the kind of the supplied type argument is compatible with the kind of the variable for which it will be substituted. As mentioned in Section 1, we see linearity as constraining the permitted usage of a variable, which means that it should always be safe to replace a linear type variable with an unrestricted type; we thus allow subbinding via the rule K-SUB. Substitution turns out to be key in many useful examples: for instance, it is what allowed the unrestricted type `UnsafeFH` to be protected by a linear interface.

Weakening—neglecting to use a variable—and contraction—using a variable more than once—should only be possible with variables of unrestricted type. Rather than add explicit contraction and weakening rules, we have built these properties into the rules that require them. The separation of linear and unrestricted typing contexts makes this fairly straightforward; rules T-LVAR and T-UVAR permit weakening by allowing an arbitrary Γ at the leaves of typing derivations, while rule T-APP duplicates Γ but splits Δ via the ⊣ relation. We can thus prove:

4 The lemmas in this section of the paper have all been proved in Coq; the source scripts of our proofs are available from the last author’s web pages.

```
It is easy to see that, modulo the value restriction, System F₀ is an extension of System F. With this in mind, Figure 4 gives several well-known System F encodings that we make use of. Aside from kind annotations, these are all standard; the linear annotations on type variables are for generality—the pairs so encoded are linear, for instance—while those that are on arrows account for captured linear variables in the arguments.

2.1 Metatheory of System F₀

Type Soundness We have verified in Coq that System F₀ enjoys type safety—a crucial but unsurprising result, given its similarity to System F. As is standard, we define soundness in terms of two properties: progress and preservation. Progress, which states that a closed, well-typed non-value can always take an evaluation step, is no different than in ordinary System F:

![Figure 4. System F encodings in F₀](image)

**Lemma 3** (Progress). If \( \vdash e : \tau \) and either \( e \) is a value or there exists some \( e' \) such that \( e \leadsto e' \).

**Proof.** Induction on typing derivations, completely standard. \( \square \)

Preservation, on the other hand, requires a bit of care. As usual, it depends on various substitution lemmas, and we must keep linearity in mind when formulating them.

**Lemma 4** (Substitution).

1. If \( \Gamma_1, \alpha \vdash e : \tau \) and \( \Gamma_1, \tau' \vdash \Gamma_2 \vdash e : \tau' \) and \( \Gamma_1 \vdash e : \tau' \) then \( \Gamma_1, \alpha \vdash \Gamma_2 \vdash e : \tau' \) and \( \Gamma_1, \alpha \vdash e : \tau' \).
2. If \( \Gamma_1, \alpha \vdash e : \tau \) and \( \Gamma_1, \tau' \vdash \Gamma_2 \vdash e : \tau' \) and \( \Gamma_1, \alpha \vdash e : \tau' \) then \( \Gamma_1, \alpha \vdash e : \tau' \) and \( \Gamma_1, \alpha \vdash e : \tau' \).
3. If \( \Gamma_1, \tau' \vdash \Gamma_2 \vdash e : \tau' \) and \( \Gamma_1, \alpha \vdash e : \tau' \) then \( \Gamma_1, \alpha \vdash e : \tau' \) and \( \Gamma_1, \alpha \vdash e : \tau' \).
4. If \( \Gamma_1, \alpha \vdash e : \tau' \) and \( \Gamma_1, \alpha \vdash e : \tau' \) then \( \Gamma_1, \alpha \vdash e : \tau' \) and \( \Gamma_1, \alpha \vdash e : \tau' \).

**Proof.** Each case by induction on the first derivation. The result relies heavily on various strengthening, weakening, and permutation lemmas (with respect to typing, context well-formedness, and the \( \vdash \) relation) regarding the handling of typing contexts. \( \square \)

Note that Substitution (3) does not hold if \( e' \) is permitted to contain free linear variables. Call-by-value reduction allows us to consider only values, however, and—because we have adopted the value restriction—we can prove that unrestricted values contain no free linear variables:

**Lemma 5.** If \( \Gamma, \Delta \vdash \Gamma \vdash e : \tau \) then \( \Delta \vdash \Gamma \vdash e : \tau \).

**Lemma 6** (Preservation). If \( \Gamma, \Delta \vdash e : \tau \) and \( e \leadsto e' \), then \( \Gamma, \Delta \vdash e' : \tau \).

From preservation and progress, soundness follows naturally:

**Theorem 7** (Type soundness). If \( \vdash e : \tau \), then it is never the case that \( e \leadsto e' \) where \( e' \) is not a value but cannot step further.

**Strong normalization and parametricity.** System F also has other properties of interest: it is strongly normalizing—evaluation always eventually reaches a value—and it enjoys relational parametricity [24]. We are able to cheat somewhat in proving the former for F₀ by observing that a well-typed System F expression becomes a well-typed System F expression upon erasure of kind annotations.
As the two systems have identical operational behavior, strong normalization follows immediately.

Parametricity for F° cannot be proved by such an erasure, but, as one might hope from the similarity to System F, it is possible to directly adapt the standard logical relations proof with only minor syntactic differences due to our separation of unrestricted and linear contexts. We have thus proved (in Coq), for the standard relation between type substitutions \( \rho_1 \approx \rho_2 : \Gamma \), relation between term substitutions \( \rho \vdash \gamma_1 \approx \gamma_2 : \Gamma; \Delta \), and computation closure \( C[\tau] \), of relations induced by a type \( \tau \), where \( \rho \) maps type variables to term relations and pairs of types and \( \Gamma \) and \( \Delta \) bind the variables in the domain of the various substitutions:

**Lemma 8 (Parametricity).** If \( \Gamma; \Delta \vdash e : \tau \), \( \rho_1 \approx \rho_2 : \Gamma \), and \( \rho \vdash \gamma_1 \approx \gamma_2 : \Gamma; \Delta \), then \( \rho_1(\gamma_1(e)), (\rho_2(\gamma_2(e))) \in C[\tau]_{\rho} \).

Succinctly, this means that an expression \( e \) with type \( \tau \), under appropriate closing substitutions, is related to itself by (the computation closure of) the relation induced logically by \( \tau \).

Of course, this is only the simplest parametricity result we could provide; it does not take into account the extensions we propose in Section 5, nor does it take advantage of linearity in any way. The interactions between linearity and relational parametricity have been explored by Birkedal et al. [8] and Biernacki et al. [7, 6] have explored program equivalences in the presence of \(!\), both of which suggest avenues for future investigation in the context of System F°. Our appeals to parametricity in Section 3.2, however, require nothing beyond what we have proved.

We view the ease with which we can adapt standard results from System F to System F° as a significant benefit of this design. A direct correspondence between these metamathematical properties and intuitions about what linearity provides is not immediately obvious, however. In Section 4 we will show, by means of elaborated operational semantics, that the restrictions required for our soundness proofs are exactly those needed to satisfy our intuitions.

### 2.2 Comparison to traditional formulations

In contrast to our approach, linear type systems are more traditionally presented without kinds, assuming linearity by default and using the modality \(!\) to allow unrestricted variables (see, for example [3, 6]). For the polymorphic \(\lambda\)-calculus, this gives us:

\[
\sigma ::= \alpha \mid \sigma \rightarrow \sigma \mid \forall \alpha. \sigma \mid !\sigma
\]

\[
t ::= \alpha \mid x \mid \lambda x. \sigma \mid t \, t \mid \Lambda \alpha. \, t \mid t [\sigma]
\]

\[
\Phi ::= \cdot \mid \Phi, \alpha \mid \Phi, x : \sigma
\]

\[
\Psi ::= \cdot \mid \Psi, \alpha : \sigma
\]

For clarity, we distinguish between linear variables \(\alpha\), bound by \(!\) terms, and non-linear variables \(x\), bound by \(\text{let} !\). The interesting typing rules concern the \(!\) modality:

\[
\frac{\Phi_1 \vdash t : \sigma}{\Phi_1 \vdash !t : !\sigma}
\]

\[
\frac{\Phi_1, \alpha \vdash t_1 : !\sigma_1, \Phi_2, x : \sigma_1; \Psi_2 \vdash t_2 : !\sigma_2}{\Phi_1, \Psi_1 \vdash \text{let} x = t_1 \text{ in } t_2 : !\sigma_2}
\]

In other words, the type \(!\sigma\) indicates a term of type \(\sigma\) which uses no linear variables—the same constraint we place on unrestricted functions—and such a term can be captured by the \(\text{let} !\) operation and subsequently used in an unrestricted fashion. Note, however, that the type \(!\sigma\) itself is still linear, even though terms of that type allow for the introduction of unrestricted assumptions; while it is possible to formulate systems where terms of \(!\) types can be duplicated or discarded directly, naïve attempts to do so are unsound—precisely the same reasons that, as discussed in Section 2.1 and Section 4, we require call-by-value reduction and the value restriction in System F°—and sound formulations end up being heavier than those that make this distinction [31, 4].

We can encode the above system in ours easily enough: we first define a translation on types \([\sigma]\) as:

\[
\begin{align*}
[\sigma_1 \rightarrow \sigma_2] & = [\sigma_1] \downarrow [\sigma_2] \\
[\forall \alpha. \sigma] & = \forall \alpha : [\sigma] \\
[!\alpha] & = \text{Unit} \downarrow [\sigma]
\end{align*}
\]

The corresponding translation on terms is straightforward. The only interesting cases involve unrestricted variables and the \(!\) modality:

\[
\begin{align*}
[\text{let } !x = t_1 \text{ in } t_2] & = (\lambda^x. \text{Unit}; [\sigma_1], [\sigma_2]) [t_1]
\end{align*}
\]

Here we treat terms \(!t\) as suspended computations which may be evaluated more than once—or not at all—which is standard.

Translating in the other direction is much less straightforward and space constraints preclude us from including the translation here. The translation on types is kind-directed and, at the term level, the insertion of \(!\) and \(\text{let} !\) operations is not trivial: a function that takes an unrestricted type must now take a \(!\) type and bind it so that the variable need not appear linearly, but such arguments must be repackaged under \(!\) in order to be passed to any subsequent functions. Polymorphic types and expressions also need care—for example, there are four cases needed to translate type instantiation, one for each combination of linear/unrestricted for the polymorphic term and its type argument.

We see this asymmetry in the translations as evidence of System F°’s expressive power and its ability to handle unrestricted terms (the bulk of any most programs) in a concise way, justifying our claim that it provides linearity in a lightweight fashion.

### 3. Examples

To demonstrate System F°’s applicability, we now turn to two categories of examples. First, we demonstrate that, while System F° does not build in notions of references and aliasing, protocols defining correct memory management can indeed be enforced by System F° types. Second, we prove that any protocol expressible as a finite automaton has a corresponding F° type; in addition to establishing that a rather large class of protocols can be encoded in our type system, this proof also highlights intuitions about linearity that still need formalization, which we will tackle in Section 4. (It is easy to see, however, that the regular languages are not an upper limit on System F°’s expressivity; the classic non-regular parenthesis matching example has an obvious protocol type.)

#### 3.1 Reference cells

The filesystem interface in Section 1.2 can, under an appropriate renaming and abstraction over the content types, also be seen as an interface for linear reference cells:

\[
\text{Ref}[\tau](\alpha) = \{ \text{set: } \tau \rightarrow \alpha \rightarrow \alpha, \text{ get: } \alpha \rightarrow (\tau, \alpha), \text{ free: } \alpha \rightarrow \text{Unit} \}
\]

\[
\text{mkRef} : \forall \beta : \alpha. \beta \Rightarrow \exists \alpha : \alpha. (\alpha, \text{Ref}[\beta](\alpha))
\]

\(^6\) Here and elsewhere, we separate the “type parameters” like \(\tau\) from the linear “state parameters” like \(\alpha\) using the notation \([\tau](\alpha)\).
On its own this is not particularly interesting, as a linear reference cell simply encodes the practice of threading a value through a program. Indeed, as with Haskell’s State monad, we could instantiate mkRef such that \( \alpha \) is \( \tau \). Instead, however, let us consider variations on Ref that make more sense as safe interfaces wrapping true, potentially unsafe reference cells, much as File in Section 1.2 protected the primitive, unchecked filehandle calls.

While the above Ref could be such a safe interface, a more obvious one—which itself requires no linearity—is the type GCRef:

\[
\text{GCRef}[\tau](\alpha) = \{ \text{set} : \tau \rightarrow \alpha \rightarrow \alpha, \\
\text{get} : \alpha \rightarrow (\tau, \alpha) \}
\]

\[
\text{mkGCRef} : \forall \beta \times \beta \rightarrow 3\alpha : \beta \rightarrow \text{GCRef}\beta(\alpha)
\]

The operations given by GCRef[\tau](\alpha) are, of course, those of a garbage-collected reference cell of type \( \tau \). By hiding the reference type behind \( \alpha \) we ensure that such garbage-collected references cannot be freed, and, in this case, \( \alpha \) can be unrestricted.

In System F\(^2\), however, we can also define references that begin their lives as linear (and manually managed) but later are put under the garbage collector’s control. Ref[\tau](\alpha) simply needs an additional function to serve as the appropriate coercion:

\[
\text{gc} : \alpha \rightarrow 3\beta : \beta \rightarrow \text{GCRef}\beta(\beta)
\]

By consuming and not returning \( \alpha \), gc prevents free from being called on the now garbage-collected (but unrestricted) reference. We thus have a coercion from alias-free to potentially aliased pointers, a fact that, had we conflated linearity with alias-freedom, would run counter to our subtyping relation of \( \star \subseteq \circ \). (Whether gc needs to do any work at run time depends on the implementation of the memory management system.)

We can take other approaches to memory management as well. For instance, an intermediate point between a strictly linear and a garbage collected reference is a reference that must be explicitly aliased, and where aliases must be explicitly discarded. We can define this easily enough:

\[
\text{RCRef}[\tau](\alpha) = \{ \text{set} : \tau \rightarrow \alpha \rightarrow \alpha, \\
\text{get} : \alpha \rightarrow (\tau, \alpha), \\
\text{alias} : \alpha \rightarrow (\alpha, \alpha), \\
\text{drop} : \alpha \rightarrow \text{Unit} \}
\]

\[
\text{mkRCRef} : \forall \beta \times \beta \rightarrow 3\alpha : \alpha \rightarrow \text{RCRef}\beta(\alpha)
\]

A straightforward implementation of mkRCRef could return a pair of the desired reference cell and an additional cell to act as a counter, along with alias and drop operations that adjust this counter. Both the primary cell and this counter could safely be freed when the count reaches zero.

However, while our access capability is still linear with RCRef, we can never be certain that we possess the only reference to cell in question. To remedy this, as is often done in capability calculi [2, 9], we can give our cells both an exclusive capability \( \alpha \) and a shared capability \( \beta \), with possession of the exclusive capability implying that no outstanding copies of the shared capability remain. If, for example, the cell contents should not be altered if any aliases exist, we can use

\[
\text{ShareRef}[\gamma](\alpha, \beta) = \{ \text{set} : \tau \rightarrow \alpha \rightarrow \alpha, \\
\text{get} : \alpha \rightarrow (\tau, \alpha), \\
\text{claim} : \beta \rightarrow \alpha \rightarrow \beta, \\
\text{alias} : \beta \rightarrow (\beta, \beta), \\
\text{drop} : \beta \rightarrow \text{Unit}, \\
\text{free} : \alpha \rightarrow \text{Unit} \}
\]

\[
\text{mkShareRef} : \forall \gamma \times \gamma \rightarrow 3\beta : \beta \rightarrow \text{ShareRef}\gamma(\alpha, \beta)
\]

Here \( \alpha \oplus \beta \) is a standard sum type; as mentioned in Section 2, these are not encodable in System F\(^2\) as presented so far, but they are an easy extension and are discussed in Section 5.1. The claim function exchanges a shared \( \beta \) for an exclusive \( \alpha \) when the underlying counter indicates that only one alias exists; in all other cases it simply returns the supplied shared capability.

We can do still more if we extend System F\(^2\) with quantification over higher kinds, an extension which meshes well with Section 5.1’s datatypes. For instance, we can define linear references that support strong updates—that is, updates that change the type contained in the reference cell—by abstracting the type of the access capability over the type of the contents, giving it kind \( \star \Rightarrow \circ \):

\[
\text{SURRef}[\tau](\alpha) = \{ \text{set} : \forall \beta \times \beta \rightarrow 3\alpha : \alpha \rightarrow \alpha, \\
\text{get} : \exists \beta \times \beta \rightarrow (\beta, \alpha), \\
\text{free} : \forall \beta \times \beta \rightarrow \text{Unit} \}
\]

\[
\text{mkSURRef} : \forall \beta \times \beta \rightarrow 3\alpha : \alpha \rightarrow \text{SURRef}\beta(\alpha)
\]

Operation records of type SUREf[\tau](\alpha) can easily be coerced to type Ref[\tau](\alpha \rightarrow \tau) by partial application of member functions. Augmenting SUREf along the lines of ShareRef could further allow for sharable reference cells that support strong updates only when they are not shared, a fairly sophisticated feature.

### 3.2 Regular protocols

Apart from memory cells and file operations, what other protocols can System F\(^2\) enforce? Rather than present more individual examples, we will show how any protocol expressible as a regular language can be written in F\(^2\).

We take the standard definition of a DFA as a tuple \( M = (Q, \Sigma, \delta, q_0, F) \), where \( Q \) is a finite set of states, \( \Sigma \) is a finite set of alphabet symbols (in our context, protocol actions), \( \delta \) is a subset of \( \Sigma \times Q \times Q \) (that is, a ternary relation among actions, current states, and next states), \( q_0 \) is a distinguished initial state, and \( F \) is a set of final states. The file access protocol from Section 1.2, for instance, can be thought of as a very simple automaton with states Open and Closed and an alphabet consisting of read, write, and close, with open as the creator of such an automaton.

For ease of notation, we extend our metavariables conventions to allow \( q \) and \( r \) as type variables. Taking \( Q = \{q_0, \ldots, q_n\} \), we can now define the automaton type for the DFA \( M \) as

\[
\tau_M = 3\eta_0 : \circ, \ldots, \eta_n : \circ, \{ \text{a}_{\text{takes}}.q_0.q : q \rightarrow q', \\
\text{for every } (a, q, q') \in \delta, \\
\ldots, \\
\text{done}_{\text{at}}.q : q \rightarrow \text{Unit}, \\
\text{for every } q \in F \}
\]

Such type is trivially inhabited: one can supply an existential package of this type where all type variables are bound to Unit.

We say that the evaluation to a value of an expression \( e \) containing a subexpression of type \( \tau_M \) reflects a word \( w = a_0 \ldots a_k \) if the sequence of record fields used is exactly

\[
a_0 \text{.takes} .q_0.q_0.a_k .\text{takes} .r_k \ldots .a_0 .\text{takes} .r_0.r_0 \]

for some states \( r_0 \) through \( r_k \). To prove that \( \tau_M \) is an accurate representation of \( M \), we need to show, first, that each \( w = a_0 \ldots a_k \) accepted by \( M \) corresponds to an expression of type \( \tau_M \rightarrow \text{Unit} \) that reflects \( w \) and, second, that each \( f \) of type \( \tau_M \rightarrow \text{Unit} \) reflects some \( w \) accepted by \( M \). The former is fairly straightforward:

**Lemma 9.** For any \( w = a_0 \ldots a_k \) accepted by \( M \), there exists \( f \) such that \( \vdash f : \tau_M \rightarrow \text{Unit} \) and for any \( n : \tau_M, f \) reflects \( w \).
Proof. Recall that, if \( w = a_0 \ldots a_n \) is accepted by \( M \), we have a trace of \( M \) on \( w \) of the form \( q_0 a_0 r_0 \ldots r_{k-1} a_k r_k \), where \( (a_0, q_0, r_0) \in \delta, (a_i, r_{i-1}, r_i) \in \delta, \) and \( r_k \in F \). Knowing this and the definition of \( \tau_M \), we can construct

\[
f = \lambda^\tau n: \tau_M. \text{unpack} (q_0, \ldots, q_n, p) = n \text{ in } \lambda \text{start, } op \mapsto p \text{ in } \lambda \text{let } s_0 = \text{ops.a$_0$ takes$q_0$ to$r_0$ start} \text{ in } x. \text{done at } r_k s_k \text{ in } \lambda \text{let } s_k = \text{ops.a$_k$ takes$r_{k-1}$ to$r_k$ s$_{k-1}$ in } x. \text{done at } r_k s_k
\]

This clearly reflects \( w \), and, if \( w \) is indeed accepted by \( M \), it will typecheck successfully.

The other direction depends on arguments from parametricity (which we have already proved) and the nature of linearity (which we will formalize and prove in Section 4).

**Lemma 10.** If \( \vdash f : \tau_M \rightarrow \text{Unit} \), then there exists some \( s \) such that \( M \) accepts \( w \) and \( f s \) reflects \( w \) for any \( n \) where \( \vdash n : \tau_M \).

Proof. Because \( \tau_M \) is linear, \( f \) must eliminate \( n \) before it can return anything of type \( \text{Unit} \). This will require unpacking the existential and then pattern-matching against the pair, leaving the linear start and unrestricted ops.

If \( q_0 \in F \), then \( f \) might immediately apply \( \text{ops.done at } q_0 \). This reflects the empty word \( e \), and, indeed, if \( q_0 \in F \), then \( M \) accepts \( e \). Alternatively, regardless of whether \( q_0 \in F \), \( f \) start can become some other state type (even \( q_0 \) again) by repeated applications of the \( \text{ops.a$_0$ takes$q_0$ to$q_0$} \) functions. If, after such applications, the result can be eliminated by some \( \text{ops.done at } q_j \), then it must be that \( q_j \in F \). Moreover, because of the construction of \( \tau_M \), each \( \text{ops.a$_k$ takes$r_{k-1}$ to$r_k$ s$_{k-1}$} \) application must represent a valid transition of \( \delta \). Thus we are reflecting some \( w \) accepted by \( M \).

In the above proof, familiar intuitions about parametricity justify the arguments that \( f \) behaves the same way for any argument of type \( \tau_M \) and that expressions whose types are variables can only be used in certain ways. The assurance what \( w \) is properly reflected, however, also depends on the connection between the static property of linear typing and the behavior of expressions at runtime. We would like type soundness to guarantee that our intuitions about the behavior of linear expressions are valid; in the next section we prove that this is indeed the case.

### 4. Linear Semantics

The preceding section makes clear that we have yet to formalize all of our intuitions about what linearity means for run-time behavior. This is not as straightforward as it might appear, because some of our intuitions about linearity turn out to be misleading. In essence, this is because linearity restricts variables of linear type, while we want to reason, at runtime, about the behavior of expressions.

For instance, linearity does not guarantee that subexpressions—or even values—of linear type will be used exactly once. Nor would we want it to: recall that the encoding of let expressions in Figure 4 involves a linear function, and we certainly want to allow for let expressions of unrestricted type.

However, there are examples which, in a call-by-name or call-by-need setting, do behave in ways we’d like to prevent. For example, if we have unpacked a reference cell from Section 3.1 as \((r, \text{ops})\), we can dispose of it and continue with the expression \( e \) via

\[
(\lambda^\tau \cdot \text{Unit} \cdot e) (\text{ops.force } r)
\]

But this call to free will only be evaluated in call-by-value—since its result is not used, other evaluation strategies would simply discard the subexpression, leaking memory, even though this is precisely the sort of error we hoped to rule out. Similar concerns for uniqueness types are known in the Clean community [28].

Call-by-value, with its assurance that function arguments are evaluated exactly once, seems well-suited to avoiding these problems, and indeed, in a call-by-value setting—with the addition of the value restriction—we are not able to construct such problematic examples. In other words, the restrictions already in place to ensure soundness in Section 2.1 are everything we need to support our intuitions about linearity. To prove that this is so, the remainder of this section provides an extended operational semantics with which we can reason about linearity at runtime.

#### 4.1 Annotated F\(^0\)

What does it mean to “use” a subexpression? We might mean by that an expression is evaluated to a value, that it is passed to a function, or that it is applied to an argument. In order to understand the run-time effects of linearity, however, we focus on the use of a value, which we divide into two phases:

1. A value is conscripted when it is substituted into a function body in place of a variable.
2. A previously conscripted value is discharged when it is used as a value to allow evaluation to continue—in a context where, were it replaced by a free variable, evaluation would be stuck. \(^7\)

From this linearity’s contribution is clear: in the course of evaluation every conscripted linear value not contained within the final result will be discharged exactly once. To formalize this, we define a linearity-aware operational semantics on expressions extended with tagged values (representing the arguments passed to functions) and tagged types (representing type arguments):

\[
\begin{align*}
\tau &::= \ldots \mid (\tau; \kappa) \\
e &::= \ldots \mid (\upsilon; \tau) \\
C, D &::= \epsilon \mid (\upsilon; \tau) \mid C, C
\end{align*}
\]

Our new semantics is given in Figure 5: we write \( e \overset{\kappa}{\rightarrow} e' \) if \( e \) steps to \( e' \) while conscripting the sequence of tagged values \( C \)

\(^7\) As variables are not values, a free variable on either side of an application results in a stuck expression by the semantics in Figure 2.

---

**Figure 5.** Linearity-aware semantics for linear System F
and discharging the sequence \( D \). Values are conscripted in rule \( \text{L-APPLAM} \), and, in a slight but critical departure from the usual operational semantics, it is the argument value tagged with the type expected by the function that is substituted into the function body. Similarly, while \( \text{L-TAPPLAM} \) does not conscript, it does tag its type argument with its expected kind, thus keeping a record of the kind at which said argument will be considered within its body.

Since a tagged value \( (v:\tau) \) is not itself a value, we have the rule \( \text{L-TAG} \), which both discharges the value and removes its tag. Intuitively this means that the argument to a previous function application is either about to be used again—regardless of which side of an application it is on—or that it is being returned as the final result of the computation. Of course, discharging may be delayed for some time, as tagged values may linger beneath function closures. As with the presentation in Section 2, evaluation does not depend on typing—types and kinds are simply recorded—and, in particular, our new operational semantics does not treat expressions of non-linear type differently from those of linear type.

The kinding and typing rules in Figure 5 are straightforward; the ability to conclude \( e:\tau \) from \( e:(\tau:\kappa) \) regardless of \( \kappa \) makes kind annotations completely transparent, but the same is not true of type annotations. The soundness results in Section 2.1 carry over to the extended system without complication.

As an example, consider an application of the linear polymorphic identity function \( \Lambda\alpha:\cdot.\lambda^x:\alpha.x \) to the natural number 42 of type \( \text{Nat} \). This expression steps as follows:

\[
\begin{align*}
(\Lambda\alpha:\cdot.\lambda^x:\alpha.x) [\text{Nat}] & \xrightarrow{42} (\lambda^x:\alpha.x) [\text{Nat}] & \xrightarrow{42} & \text{L-TAPPLAM} \\
(\lambda^x:\alpha.x) (\text{Nat}[\alpha]) & \xrightarrow{42} & (\text{Nat}[\alpha]) [\text{T}] & \text{L-TAPPLAM} \\
42 & \xrightarrow{42} & 42 & \text{L-TAG}
\end{align*}
\]

Or, taking the obvious tag-concatenating closure \( C \):

\[
(\Lambda\alpha:\cdot.\lambda^x:\alpha.x) [\text{Nat}] \xrightarrow{C} 42
\]

In other words, over the course of evaluation, the value 42 was passed to a function and used (in this case, returned as the final result) exactly once, and it was considered at type \( (\text{Nat}[\alpha]) \). Were it to appear more than once in the lower annotation, we would know that linearity had failed us—although 42 is of type \( \text{Nat} \), here it is being considered as a linear type because \( \alpha \) was declared to be linear, so after being conscripted at this type it should be discharged exactly once. Were it to instead appear as \( (42:\text{Nat}) \), however, there should be no such restrictions on its use, as we can indeed write functions from \( \text{Nat} \) to \( \text{Nat} \) which use their arguments more than once or ignore them altogether.

This semantics is similar in scope to heap semantics for linear functional languages [26], and indeed we could have taken this approach by tracking, instead of the traces \( C \) and \( D \), a single heap \( H \).

Such a semantics would perform substitution at application only if the argument type is unrestricted; for linear types, we would simply associate the variable with its argument in the heap, removing it when the variable is used. But, while we could still prove Theorem 12 with such a semantics, we need the ability to track unrestricted arguments—typically not placed in the heap—and to reason about the order of function calls in order to support the reasoning we used in Section 3, which we will return to in Section 4.3.

### 4.2 Linearity at run time

Before we can formalize our intuitions about linearity, we need some auxiliary definitions. We write \( \Sigma(e) \) for the multiset of tagged values appearing within \( e \), defined simply as:

\[
\begin{align*}
\Sigma(x) & = \emptyset \\
\Sigma((v:\tau)) & = \{(v:\tau)\} \cup \Sigma(v) \\
\Sigma((\Lambda\alpha:\cdot.e_1)) & = \Sigma(e_1) \\
\Sigma(e_1 \cdot e_2) & = \Sigma(e_1) \cup \Sigma(e_2) \\
\Sigma(e_1 \mapsto e_2) & = \Sigma(e_1) \cup \Sigma(e_2)
\end{align*}
\]

We write \( \{C\} \) to denote the treatment of a sequence as a multiset, and, if \( S \) is a multiset of tagged values, we write \( S\kappa \) for the subset of \( S \) that omits any \( (v:\tau) \) where \( \vdash :\tau \kappa \).

We define proper expressions as those in which no unrestricted value contains a value as a subexpression tagged with linear type—that is, a well-typed expression \( e \) is proper iff, for every value subexpression \( v \) in \( e \), if \( e \) has some type \( \tau \) and \( \tau \) can be given kind \( \kappa \) then \( \Sigma(v)\kappa = \emptyset \). This property is preserved by evaluation:

**Lemma 11 (Proper expressions).** If \( e \) is proper and \( e \xrightarrow{S} e' \), then \( e' \) is also proper.

\( \square \)

Thus, as long as we write our source expressions in the language described in Section 2, we will never evaluate to an improper result—this is essentially the runtime version of Lemma 5. If we add new constructs, this property will continue to hold as long as

1. constructs under which evaluation can proceed (like conventional tuples) are given unrestricted types only if all of their components are also unrestricted, and
2. constructs that suspend computation either contain only values (like \( \Lambda \)) or only typecheck at an unrestricted type given an empty linear context (like \( \Lambda \)).

(Or course, it also suffices to require that a new construct always be typed linearly, as we do for additive conjunction in Section 5.1.)

We are now equipped to prove our main result, that linearity guarantees a correspondence between values conscripted and discharged at linear type:

**Theorem 12 (Run time linearity).** If \( e \) is proper and \( e \xrightarrow{D} e' \), then \( \Sigma(e)\kappa \subseteq \Sigma(e')\kappa \) and \( \Sigma(D)\kappa = \emptyset \).

\( \square \)

**Proof.** By straightforward induction over the annotated evaluation relation; \( \text{L-APPLAM} \) is the only interesting case. From \( T-\text{UVAR} \) and Lemma 11 we know that, unless the argument occurs exactly once in the function body, it will be of unrestricted type and contain no linear tags. We thus preserve our tag balance.

In other words, at runtime, arguments treated linearly are never duplicated or discarded, exactly as we would hope. We can also immediately prove the following corollary, summarizing everything we know about expressions with unrestricted type:

**Corollary 13 (Unrestricted results).** If \( e \) is proper, \( \vdash :\tau \), and \( \vdash :\kappa \), then there exists some \( v \) such that \( e \xrightarrow{D} v \), \( \Sigma(v)\kappa \subseteq \Sigma(e)\kappa \), and \( \Sigma(D)\kappa = \emptyset \).

**Proof.** Follows directly from soundness, strong normalization, Lemma 11, and Theorem 12.

\( \square \)

### 4.3 Applications

To demonstrate how the above applies to the DFA encoding given in Section 3.2, we first examine the simpler encodings of products and existential types from Section 2.

**Products** Our traces treat pairs exactly as we would expect. Given \( e_1, e_2 \xrightarrow{D_1} v_1 \) and \( e_2, e_2 \xrightarrow{D_2} v_2 \), where \( e_1 \) has type \( \tau_1 \) and \( e_2 \) has type \( \tau_2 \), recall that

\[
(\tau_1, \tau_2) \xrightarrow{D} (e_1, e_2)
\]

We are now equipped to prove our main result, that linearity guarantees a correspondence between values conscripted and discharged at linear type:

**Theorem 12 (Run time linearity).** If \( e \) is proper and \( e \xrightarrow{D} e' \), then \( \Sigma(e)\kappa \subseteq \Sigma(e')\kappa \) and \( \Sigma(D)\kappa = \emptyset \).

**Proof.** By straightforward induction over the annotated evaluation relation; \( \text{L-APPLAM} \) is the only interesting case. From \( T-\text{UVAR} \) and Lemma 11 we know that, unless the argument occurs exactly once in the function body, it will be of unrestricted type and contain no linear tags. We thus preserve our tag balance.

In other words, at runtime, arguments treated linearly are never duplicated or discarded, exactly as we would hope. We can also immediately prove the following corollary, summarizing everything we know about expressions with unrestricted type:

**Corollary 13 (Unrestricted results).** If \( e \) is proper, \( \vdash :\tau \), and \( \vdash :\kappa \), then there exists some \( v \) such that \( e \xrightarrow{D} v \), \( \Sigma(v)\kappa \subseteq \Sigma(e)\kappa \), and \( \Sigma(D)\kappa = \emptyset \).

**Proof.** Follows directly from soundness, strong normalization, Lemma 11, and Theorem 12.

\( \square \)

Our traces treat pairs exactly as we would expect. Given \( e_1, e_2 \xrightarrow{D_1} v_1 \) and \( e_2, e_2 \xrightarrow{D_2} v_2 \), where \( e_1 \) has type \( \tau_1 \) and \( e_2 \) has type \( \tau_2 \), recall that

\[
(\tau_1, \tau_2) \xrightarrow{D} (e_1, e_2)
\]
Clearly, then,
\[(e_1, e_2) \quad C_{1, 2}, (v_1 : (\tau_1 : \alpha)) \rightarrow (v_2 : (\tau_2 : \epsilon)) \rightarrow^* (v_1, v_2)\]

Given our representation of pairs as closures, this is reasonable; though the conscription trace tags \(v_1\) and \(v_2\) with linear types, they will be discharged at these types as soon as \((v_1, v_2)\) is destructed and, assuming the provided function \(f\) is of declared type \(\tau_1 \rightarrow \tau_2 \rightarrow \gamma\), recognized at their original, ontaged types \(\tau_1\) and \(\tau_2\).

**Existentials** Existential types are a bit more complicated. Given \(e_1 \rightarrow^* v_1\) where \(e_1\) has type \(\{\alpha \rightarrow \tau\}\), \(\tau_1\) has kind \(\kappa_1\), and \(\tau\) has kind \(\kappa\), let
\[
e_2 = \text{pack } \alpha : \kappa = \tau \text{ in } e_1 : \tau_1
\]

\[
= \text{let } x = e_1 \in \Delta : \beta : \kappa \rightarrow \forall \alpha \beta. \lambda^e x : (\forall \alpha x : (\forall \alpha x : \tau_1) \rightarrow \beta), f \in \tau_1 \rightarrow 0
\]

\[
e_2 = (\lambda^e x : (\forall \alpha x : \tau) \rightarrow \tau_1). \Delta : \beta : \kappa \rightarrow \forall \alpha \beta. \lambda^e x : (\forall \alpha x : \tau_1) \rightarrow \beta, f \in \tau_1 \rightarrow 0
\]

Letting \(v_2\) refer to the result of fully evaluating \(e_2\), we have \(e_2 \in \Delta : \alpha \rightarrow \tau_2 \rightarrow \gamma\); the details of this reduction can be seen in Figure 6. This seems sensible for the encoding of an existential package, as it reflects exactly that the result of evaluating \(e_1\) has been captured in a function closure, just as with the capture that occurs when applying the constructor.

Now, to use \(v_2\), take \(e'\) of type \(\tau'\) and let
\[
e_3 = \text{unpack } \alpha : \kappa = \tau \text{ in } e' \quad v_2 \rightarrow^* \quad v_2 \quad \tau'
\]

The evaluation of this expression, as shown in Figure 6, is a bit lengthier; the interesting thing to note, though, is that, before the existentially wrapped \(v_1\) is passed to the function wrapping \(e'\), it is discharged at type \((\alpha \rightarrow \tau)\) \(\tau_1\) and reconscripated at the type \((\alpha \rightarrow (\tau : \kappa))\) \(\tau_1\). This stricter reconscription stands in direct contrast to what occurs in the elimination of pairs, and, if \(\kappa\) is \(\omega\), we know that any component of \(v_1\) with type \((\tau : \kappa)\) must be discharged exactly once in \(e'\) for each time it is conscripated, regardless of whether or not \(\tau\) can also be given kind \(\kappa\).

Subsequently, of course, it could be reconscripated simply at type \(\tau\), but thanks to parametricity we know that \(e'\) cannot do this on its own—as in the filesystem we know that \(e'\) can only be done without it. The resulting evaluation reflecting a function already present in the existential package.

**DFAs** Finally, to show how the arguments made in proving Lemma 10 are made rigorous by this machinery, consider making two more minor syntactic changes to System F\(^{\omega}\): first, rather than writing \(\lambda^e x : \tau, e\) for functions, write \(fn^e\) name \(x : \tau, e\), and second, rather than the simple L-TAG, add rules for the left and right sides of applications which further add L and R tags to discharged values, along with a final result rule which adds no further tag. It doesn’t matter what names we give to functions, but if we choose to use our record labels as function names when possible—and require that all other names are drawn from some set disjoint from said record labels—then our definition of an evaluation reflecting a word becomes very simple: the DFA trace must be visible in \(D\) by examining the values tagged with \(L\).

We can now return to the proof given in Section 3 and consider the problem in terms of value traces. After unpacking the existential, we will have the linear initial state conscripted but not yet discharged, and because our entire expression has the unrestricted type \(\Lambda\), we know from Corollary 13 that it must be discharged before the final result is returned. Parametricity further restricts states to being discharged on the right side of applications. Then, as before, it is simply a matter of following the state as functions are applied to it, noting the labels that can become attached to the value trace; in the end, when some done at \(q\) is applied, our value trace will contain a clear accepting DFA trace.

5. Extensions: Towards Practicality

Although System F\(^{\omega}\) as presented so far can serve as an expressive core calculus, to be usable in practice it is necessary to extend it with features suitable for a surface language. This section shows how to add a variety of useful constructs familiar from functional programming. None of these extensions are particularly difficult, though we point out a few places where linearity must be minded.

5.1 Language additions

**Polikind products** System F\(^{\omega}\) as it stands can already encode both unrestricted and linear (multiplicative) products. In practice, however, it is useful to build in support for n-ary tuples. As shown in Figure 7, the typing rules are mostly standard; evaluation rules are completely standard and have been omitted. The kinding rule K-PROD, in combination with subkinding, allows the same syntax
recursive, polymorphic datatypes—as with multiplicative products, Algebraic datatypes note that this is compatible with Lemma 4 (3). These operational semantics might duplicate than pattern matching. In keeping with our syntactic conventions, that immediately unwinds itself to a value as shown in E-F.

Adding support for fixpoint computation turns out to |

[54x68]...\langle τ_1, ..., τ_n \rangle \circ Δ \vdash e : \langle τ_1, ..., τ_n \rangle\]

[54x114]new syntax and rules, where we write |

[D-D] \Gamma; \Delta \vdash e_1 : \tau_1, ..., e_n : \tau_n\]

Additive products As mentioned in Section 2, we cannot encode linear sums in System F° as presented so far—we have no means of sharing the linear context among different subexpressions, all but one of which will be dropped. Connectives that enable this are referred to as additive in the terminology of linear logic; traditionally they are written as \( \tau_1 \oplus \tau_2 \) for sums, and \( \tau_1 \& \tau_2 \) for additive products—lazy linear products eliminated by projection rather than pattern matching. In keeping with our syntactic conventions, however, we introduce n-ary additive products \( \langle \tau_1, ..., \tau_n \rangle \); sums could be encoded in terms of these products but are also subsumed by the algebraic datatypes discussed below.

Figure 8 gives the new syntax, definitions, and static and dynamic rules for n-ary additive products. Note that additive products necessarily denote suspended computations, as further evaluation progress cannot be made until a branch is chosen. Unlike multiplicative products, additive products always have kind \( \exists \) regardless of the kinds of the component types, as seen in K-ADD.

Fixpoints Adding support for fixpoint computation turns out to be remarkably straightforward: recursion implies the ability to repeatedly invoke a function within its own body, which naturally puts us in the unrestricted portion of System F°. Figure 9 shows the new syntax and rules, where we write \( \text{fix } f. v \) for the computation that immediately unwinds itself to a value as shown in E-F. Since these operational semantics might duplicate \( v \), T-Fix requires that fixpoint expressions be closed with respect to the linear context; note that this is compatible with Lemma 4 (3).

Algebraic datatypes Figure 10 shows the static rules for general recursive, polymorphic datatypes—as with multiplicative products, the dynamic rules are standard. We assume a signature \( \Sigma_T \) that maps a datatype constructor \( T \) to its (higher) kind \( \kappa_T \), which describes \( T \)'s type parameters; \( T \) has arity \( k_T \). Term-level constructors for type \( T \) are written \( \text{ctr}_T \), and their types are given by the map \( \Sigma_T \). The type of a term-level constructor is polymorphic over the parameters of the datatype, but it may also expect additional type parameters, which act as through they are existentially bound—this design is similar to that found in Haskell. Importantly, if the result kind of the type constructor \( T \) is \( \exists \), then the value parameters to the term-level constructors must also be of kind \( \exists \). This constraint, written as part of the signature well formedness checks in Figure 10, is necessary to ensure that an unrestricted datatype can’t hide a linear value and thus allow it to be duplicated or discarded.

The additive nature of datatypes is shown in T-CASE, which uses the same linear context \( \Delta_2 \) when checking each branch. The branches themselves are polymorphic linear functions—additionally abstracted over any existential type parameters—one of which will be applied to create a result of the appropriate type. As with additive products, each branch will capture the same linear free variables, but only one of the branches will fire at run time.

5.2 Syntactic support for linearity

In addition to making it possible for types to express safe protocols over stateful resources, it must also be convenient for programmers to write client code that interacts with these interfaces. For instance, a file copy program, using the interface given in Section 1.2, should look like simple imperative code and not require the programmer to think explicitly about existentially bound linear type variables and the threading of linear filehandles. In the rest of this section we sketch out some possible syntactic support—partly inspired by but simpler than Haskell’s type classes—for programming with linearity, which allows us to copy files as follows: \(^6\)

\[
\begin{align*}
\text{let } &\text{copy = fix } f. \\
&\begin{align*}
&\text{action } \{ &\text{[File]} h_1, &\text{[File]} h_2, \\
&y \leftarrow h_1 \cdot \text{read}; \\
&\text{if } y = \text{EOF} &\text{then return } () \\
&\text{else } h_2 \cdot \text{write } y; \\
&\left( h_1, h_2 \right) . f \\
&\} \\
&\text{in } \\
&\text{start } \{ &\text{[File]} h_1 = \text{open } \text{"InFile.txt"}; \\
&\text{start } \{ &\text{[File]} h_2 = \text{open } \text{"OutFile.txt"}; \\
&\left( h_1, h_2 \right) . \text{copy}; \\
&h_1 . \text{close}; &h_2 . \text{close} \\
&\}
\end{align*}
\end{align*}
\]

This program would fail to typecheck if either of the calls to close were omitted, if one of the file handles were closed within the body of copy, or if a read or write operation were called after close. In the rest of this section we will show how to translate from this syntax into System F° as described up to this point.

First, we make more explicit the pattern common to the protocol examples we have seen thus far. Let \( \Theta = \alpha_1, \ldots, \alpha_n \) be a list of type variables representing the states of the protocol.\(^7\) The possible actions of the protocol can be given by a record of operations, each of which causes a state transition from \( \sigma_{in} \) to \( \sigma_{out} \).

\[
\begin{align*}
\text{Ops}(\Theta) = \{ &\text{op}_1 : \forall i, k_1 . \tau_{in} \cdot \sigma_{in} ; \tau_{out} \; \tau_{out} (\tau_{out}, \sigma_{out}) \\
&\ldots \\
&\text{op}_m : \forall i, k_m . \tau_{in} \cdot \sigma_{in} ; \tau_{out} \; \tau_{out} (\tau_{out}, \sigma_{out}) \}
\end{align*}
\]

Here, \( \sigma_{in}, \sigma_{out} \in \Theta \) is the start state of the operation and \( \sigma_{out} \in \Theta \cup \{ \text{Unit} \} \) is either the end state of the operation or Unit, indicating

\(^6\) Technically, to match up with what follows, the return type of write must be \((\text{Unit}, \alpha)\) rather than just \(\alpha\); we also assume that EOF is of type Char.

\(^7\) In this section, we write \(\bigvee_{\alpha_1, \ldots, \alpha_n} \Theta \cdot \tau\) as short hand for \(\forall \alpha_1 \cdot \ldots \forall \alpha_n \cdot \Theta \cdot \tau\), and similarly for existentials.
that the protocol is complete. We assume that each operation might be polymorphic and, for simplicity, that each takes one input of type \(\tau\) and produces an output of type \(\tau\), both of which might contain occurrences of variables in \(\Theta\). Almost all of the protocol examples we have seen can be written in this form.

If \(\alpha_{out} \in \Theta\) is the start state, then an initial instance \(e_1\) of the protocol is represented by a value \(e_{\alpha_{out}}\). \(\Theta\) thanks to \(\alpha_{out}\), this value has linear kind, so it must be unpacked and the resulting tuple deconstructed. We suggest a convenient notation to simplify this process:

\[
\text{start } [\text{Ops}] \; h = e_1 \in e_2 \triangleq \text{unpack } \Theta_{h_0}, x = e_1 \in \text{let } (h, \text{ops}_h) = x \in e_2
\]

Here, \(\Theta_{h_0}\) is a list of fresh type variables, \(h\) is bound to the linear state, and \(\text{ops}_h\) is the record of operations associated with \(h\). The syntax includes the annotation \(\text{Ops}\) to help with type checking and to determine what syntactic sugar applies with respect to \(h\) in the body \(e_2\). Note that \(\alpha_{out}\) has type \(\text{Ops}(\Theta_h)\), which is unrestricted and hence may be duplicated at will; this invariant will be maintained throughout the interpretation of our syntax.

The intuition behind our scheme is simple: we use the single name \(h\) for the linear handle associated with the thread of a protocol—this implicit reuse can never be an issue, since the variable is linear. This handle (which must be typed by a variable in \(\Theta_h\)) then determines an appropriate record of operations, \(\text{ops}_h\), associated with the protocol. The programmer never mentions \(\alpha_{out}\) explicitly; rather it is threaded through the computation automatically, much like a classdef.

The basic notation treats the protocol operations as “methods” of the linear handle \(h\) with implicit argument \(\text{ops}_h\), binding the result to the variable \(y\) and conflating the initial and result handles:

\[
y \leftarrow h.\text{op}_1 \; [\tau] \; e_1; 
\]

\[
\text{let } (y, h) = \text{ops}_h.\text{op}_1 \; [\tau] \; e_1 \; h \in e_2
\]

If the programmer wishes to write a client function that takes one or more handles, each following its own protocol, the function must be polymorphic over the appropriate \(\text{ops}\) records. We call such functions “actions”; it makes sense to give them a type very similar to the operations of a protocol, since they must take a vector of resource handles each in some protocol state and return a new vector of protocol states.

\[
\text{action } [\text{Ops}_1 \ldots \text{Ops}_n] \; h_1, \ldots, h_n \rightarrow \text{Ops}_1(\Theta_{h_1}) \ldots \text{Ops}_n(\Theta_{h_n}) \rightarrow \text{out}(\Theta_{h_1} \ldots \Theta_{h_n})
\]

\[
\Delta \vdash \text{case } e \; \text{of } \text{out} \rightarrow \text{let } (y, s) = f(\Theta_{h_1}) \ldots \Theta_{h_n} \; \text{ops}_{h_1} \ldots \text{ops}_{h_n} \; [\tau] \; e_1 \; h_1, \ldots, h_n \in e_2
\]

Here \(e\) must have type \(\text{outs}(\Theta_{out}, \ldots, \text{outs}_n)\), where \(\alpha_{out} \in \Theta_h\) and the type of \(h_j\) must be some \(\alpha_{out}\) in \(\Theta_{out}\), representing the state that handle must be in when this function is called. Our syntactic sugar requires that type inference be able to determine the \(\alpha_{h_1}\)’s by inspecting \(e\). We expect this to be rather straightforward, since the type is uniquely determined by the first operation invoked on the handle; alternatively we could add a type annotation to \(h_1\). Inside the body of such an action, we allow a convenient means of packaging the state values to be returned to the caller:

\[
\text{return } e \triangleq (e_1, \ldots, e_n)
\]

Taken together, these constraints ensure that the type of an action is a polymorphic operation over a vector of states; i.e., has the type

\[
\forall \Theta_{h_1} : \alpha_{out} \ldots \text{outs}_n \rightarrow \forall \Theta_{h_1} : \alpha_{out} \ldots \text{outs}_n \rightarrow \forall \Theta_{h_1} : \alpha_{out} \ldots \text{outs}_n \rightarrow \forall \Theta_{h_1} : \alpha_{out} \ldots \text{outs}_n
\]

Inside the body of an action, the “method call” syntax defined earlier can be used to invoke protocol operations on the handles \(h_j\). Since an action is just another sort of operation over handles—albeit a higher level one—we would like similar syntax for invoking such an operation on a list of handles. Assuming \(f\) has the type of an action as given above, we define:

\[
y \leftarrow (h_1, \ldots, h_n); f \; [\tau] \; e_1; e_2 \triangleq \text{let } (y, s) = f(\Theta_{h_1}) \ldots \Theta_{h_n} \; \text{ops}_{h_1} \ldots \text{ops}_{h_n} \; [\tau] \; e_1 \; h_1, \ldots, h_n \in e_2
\]

One can easily define variants of both the above and the earlier protocol operation notation that omit the argument \(e_1\) or the binding for \(y\) in the case that one or both of \(\tau_{a_1}\) and \(\tau_{av}\) are Unit, or that do not rebind the handle variable in cases where the handle is consumed, as with close. Also, if the operation invocation is last in a sequence, \(e_2\) and the corresponding the let binding may be omitted, since the result of the operation is the result of the sequence—this is useful, e.g., for making tail calls.

Of course, there are many ways of using linear types, and the above does not account for all of them. We believe, however, that it does show the feasibility of making System F\(^\omega\) protocols palatable to the programmer, and that this provides further evidence that linearity as implemented in System F\(^\omega\) deserves to be considered for inclusion in more mainstream functional programming languages.

6. Conclusion

We have presented System F\(^\omega\), a simple variant of the linear polymorphic \(\lambda\)-calculus that nevertheless can enforce a rich variety of
protocols. System F₂ is sound and enjoys parametricity, and we have proved that protocol enforcement is faithful—that is, linear resources are never misused—thanks to these properties. We have demonstrated the applicability of System F₂ through a variety of examples and by showing how to extend it with features geared towards practical programming with linear types.

References


[29] Philip Wadler. Linear types can change the world! In M. Broy and C. Jones, editors, Programming Concepts and Methods, Sea of Galilee, Israel, April 1990. North Holland. IFIP TC 2 Working Conference.


