Inférence de flots d'information pour ML

Vincent Simonet
INRIA Rocquencourt – Projet Cristal

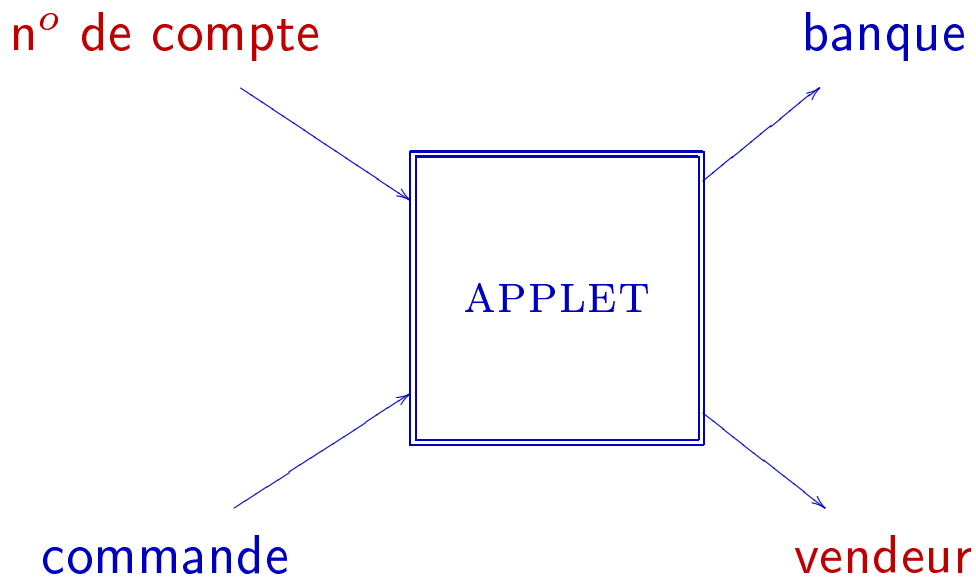
Séminaire DÉMONS (L.R.I.)

Vendredi 30 novembre 2001

Vincent.Simonet@inria.fr http://cristal.inria.fr/~simonet/

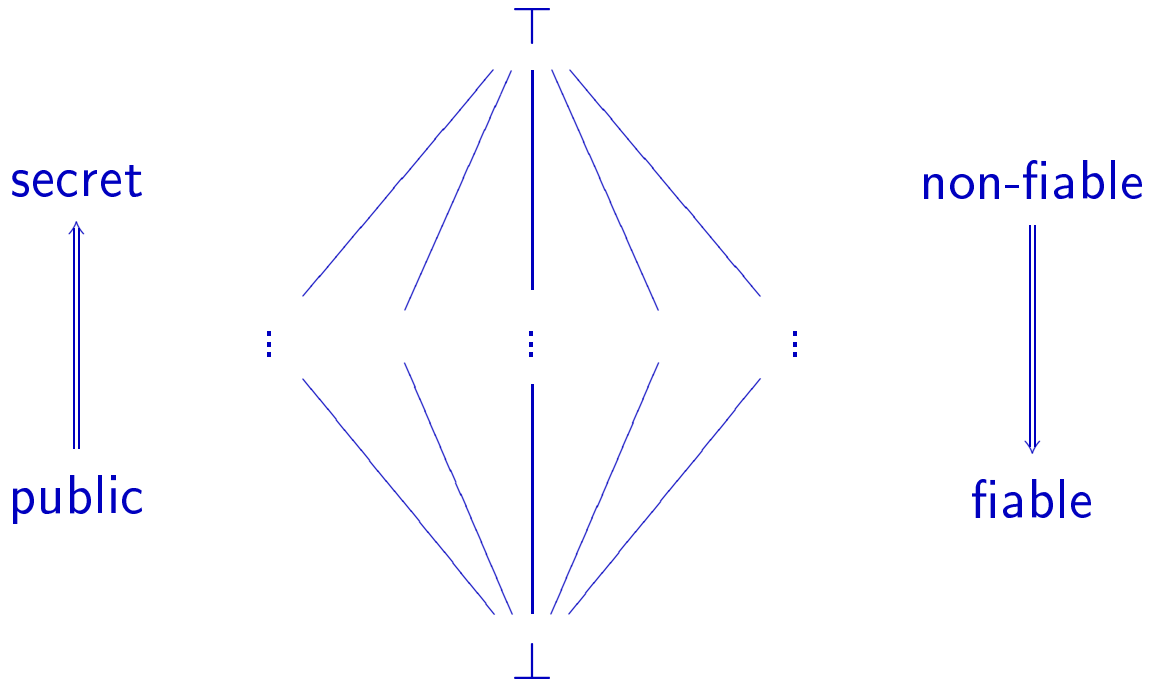
Analyse de flots d'information

L'analyse de flots d'information consite à analyser les dépendances entre les entrées et les sorties d'un programme de manière à vérifier qu'il satisfait certaines propriétés de confidentialité ou d'intégrité vis-à-vis des données qu'il manipule.



Niveaux d'information

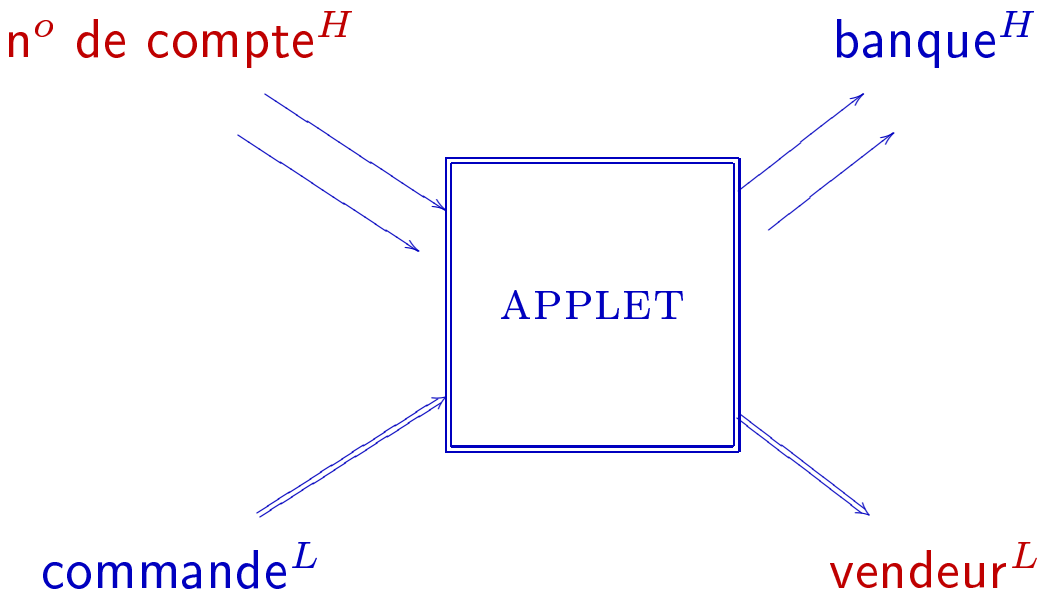
On associe à chaque donnée un niveau d'information, dans un treillis $\mathcal{L}$. Ces niveaux peuvent représenter différentes propriétés : sécurité, intégrité...



Pour la suite, on fixe $\mathcal{L} = \{L \leq H\}$.

Analyse de flots d'information

Non-interférence



L'indépendance entre une entrée et une sortie est exprimée sous forme d'une propriété de non-interférence.

Typage

Les flots d'informations peuvent être représentés par des types annotés :

$$\text{compte}^{H} \times \text{commande}^{L} \to \text{banque}^{H} \times \text{vendeur}^{L}$$

ou plus généralement par des schémas polymorphes contraints :

$$(\forall \alpha\beta\gamma\delta)\ [\alpha \sqcup \beta \leq \gamma, \beta \leq \delta]\ \text{compte}^{\alpha} \times \text{commande}^{\beta} \to \text{banque}^{\gamma} \times \text{vendeur}^{\delta}$$

Systèmes existants

Dennis Volpano et Geoffrey Smith (1997)
Système de type sur un langage impératif simple. Limité au premier ordre et à un nombre fini de références globales.

Nevin Heintze et Jon G. Riecke SLam Calculus (1997)
Lambda-calcul avec références et processus. Le typage des références n'est pas fin. Il n'est pas énoncé ni prouvé de propriété de sûreté.

Andrew C. Myers JFlow (1999)
Analyse de flots d'information sur Java. Le système est complexe et n'est pas prouvé.

Steve Zdancewic et Andrew C. Myers (2001)
Analyse sur un langage de bas niveau avec continuations linéaires.

Le langage ML

λ-calcul en appel par valeur avec let-polymorphe

x $\qquad$ k $\qquad$ $\lambda x.e$

$e_1\, e_2$ $\qquad$ let $x = v$ in e $\qquad$ bind $x = e_1$ in e_2

avec références

ref e $\qquad$ $e_1 := e_2$ $\qquad$ $!\, e$

et exceptions

$\varepsilon\, e$ $\qquad$ raise e $\qquad$ e_1 handle $\varepsilon\, x \succ e_2$ $\qquad$ e_1 handle $x \succ e_2$

Formes v-normales

$$
\begin{array}{lcl}
v & ::= & x \mid k \mid \lambda x.e \mid \varepsilon\, v \\
e & ::= & v\, v \mid \mathsf{ref}\ v \mid v := v \mid !\, v \mid \mathsf{raise}\ v \mid \mathsf{let}\ x = v\ \mathsf{in}\ e \mid E[v] \\
E & ::= & \mathsf{bind}\ x = [\,]\ \mathsf{in}\ e \mid [\,]\ \mathsf{handle}\ \varepsilon\, x \succ e \mid [\,]\ \mathsf{handle}\ x \succ e
\end{array}
$$

On peut mettre toute expression habituelle sous cette forme, à condition de choisir un ordre d'évaluation. Par exemple :

$$
e_1\, e_2 \Rightarrow \begin{cases} \mathsf{bind}\ x_1 = e_1\ \mathsf{in}\ (\mathsf{bind}\ x_2 = e_2\ \mathsf{in}\ x_1\, x_2) & \textit{éval. de gauche à droite} \\ \mathsf{bind}\ x_2 = e_2\ \mathsf{in}\ (\mathsf{bind}\ x_1 = e_1\ \mathsf{in}\ x_1\, x_2) & \textit{éval. de droite à gauche} \end{cases}
$$

Exemples de flot d'information

Flots directs

$$x := \text{not } y$$

$$x := (\text{if } y \text{ then } \text{false} \text{ else } \text{true})$$

Flots indirects

$$\text{if } y \text{ then } x := \text{false} \text{ else } x := \text{true}$$

$$x := \text{true};\ \text{if } y \text{ then } x := \text{false} \text{ else } ()$$

$$x := \text{true};\ (\text{if } y \text{ then raise } A \text{ else } ()) \text{ handle } _ \succ x := \text{false}$$

Exemples de flot d'information

Niveau associé au « compteur de programme »

Supposons que y représente une donnée « secrète » (H) :

$$\text{if } y \text{ then } \underbrace{x := \text{false}}_{pc=H} \text{ else } \underbrace{x := \text{true}}_{pc=H}$$

$$\underbrace{x := \text{true};}_{pc=L} \text{ if } y \text{ then } \underbrace{x := \text{false}}_{pc=H} \text{ else } ()$$

$$\text{let } f = \lambda b.(x := b) \text{ in } \underbrace{f\,\text{true};}_{pc=L} \text{ if } y \text{ then } \underbrace{f\,\text{false}}_{pc=H} \text{ else } ()$$

À chaque point du programme correspond un niveau pc qui représente les informations que l'on peut apprendre en sachant que ce point a été exécuté.

Système de types

Approche semi-syntaxique

(exemples dans le cas du typage ML)

Système logique	**Système syntaxique**
Types bruts e.g. int, int $\rightarrow$ int...	Expressions de type e.g. int, α, $\alpha \rightarrow \alpha$...
Polytypes e.g. $\{t \rightarrow t \mid t$ type brut$\}$	Schémas de type e.g. $\forall\alpha.\alpha \rightarrow \alpha$

On raisonne sur le système logique. Le système syntaxique est interprété dans le système logique.

Système de types

Algèbre de types

Les niveaux d'information ℓ, pc appartiennent à un treillis $\mathcal{L}$.

Les exceptions sont décrites par des rangées d'alternatives r :

$$\begin{array}{lcl} a & ::= & \mathsf{Abs} \mid \mathsf{Pre}\ pc \\ r & ::= & \{\varepsilon \mapsto a\}_{\varepsilon \in \mathcal{E}} \end{array}$$

Les types sont annotées par des niveaux et des rangées :

$$t \quad ::= \quad \mathsf{int}^{\ell} \mid \mathsf{unit} \mid (t \xrightarrow{pc\ [r]} t)^{\ell} \mid t\ \mathsf{ref}^{\ell} \mid r\ \mathsf{exn}^{\ell}$$

Jugements

On distingue deux formes de jugements.

Jugements sur des valeurs

$$\Gamma \vdash v : t$$

Jugements sur des expressions

$$pc, \Gamma \vdash e : t \ \ [r]$$

Contraintes

Contraintes de sous-typage $t_1 \leq t_2$

La relation de sous-typage étend l'ordre sur les niveaux aux types. Par exemple :

$$\mathsf{int}^{\ell_1} \leq \mathsf{int}^{\ell_2} \Leftrightarrow \ell_1 \leq \ell_2 \qquad t_1\ \mathsf{ref}^{\ell_1} \leq t_2\ \mathsf{ref}^{\ell_2} \Leftrightarrow t_1 = t_2 \text{ et } \ell_1 \leq \ell_2$$

Gardes $\ell \lhd t$

Les gardes permettent de « marquer » un type par un niveau d'information :

$$pc \lhd \mathsf{int}^{\ell} \Leftrightarrow pc \leq \ell \qquad pc \lhd t\ \mathsf{ref}^{\ell} \Leftrightarrow pc \leq \ell$$

Système de types

Sous-typage et polymorphisme

Sous-typage et polymorphisme interviennent de manière orthogonale :

Sous-typage Permet d'augmenter librement le niveau d'une donnée (e.g. considérer *secrète* une donnée *publique* ou *non-fiable* une donnée *fiable*) :

$$\frac{\Gamma \vdash v : t \qquad t \leq t'}{\Gamma \vdash v : t'}$$

Polymorphisme Nécessaire pour pouvoir utiliser la même fonction avec des entrées de niveaux différents :

$$\text{let } succ \;=\; \lambda x.(x + 1)$$

Système de types

Références

$$\textsc{Ref}\quad \frac{\Gamma \vdash v : t \qquad pc \lhd t}{pc, \Gamma \vdash \mathsf{ref}\ v : t\ \mathsf{ref}^{\ell}\ [r]}$$

$$\textsc{Deref}\quad \frac{\Gamma \vdash v : t'\ \mathsf{ref}^{\ell} \qquad t' \leq t \qquad \ell \lhd t}{pc, \Gamma \vdash\ !\, v : t\ [r]}$$

$$\textsc{Assign}\quad \frac{\Gamma \vdash v_1 : t\ \mathsf{ref}^{\ell} \qquad \Gamma \vdash v_2 : t \qquad pc \lhd t \qquad \ell \lhd t}{pc, \Gamma \vdash v_1 := v_2 : \mathsf{unit}\ [r]}$$

Le contenu d'une référence a un niveau supérieur à égal

- au pc du point où la référence est créée,
- au pc de chaque point du programme où son contenu est susceptible d'être modifié.

Système de types

Exceptions

RAISE

$$\frac{\Gamma \vdash v : typexn(\varepsilon)}{pc, \Gamma \vdash \mathsf{raise}\ (\varepsilon\, v) : *\ \ [\varepsilon : \mathsf{Pre}\ pc; \partial\mathsf{Abs}]}$$

HANDLE

$$\frac{pc, \Gamma \vdash e_1 : t\ \ [\varepsilon : \mathsf{Pre}\ pc'; r_1] \qquad pc \sqcup pc', \Gamma[x \mapsto typexn(\varepsilon)] \vdash e_2 : t\ \ [\varepsilon : a_2; r_2] \qquad pc' \lhd t}{pc, \Gamma \vdash e_1\ \mathsf{handle}\ \varepsilon\, x \succ e_2 : t\ \ [\varepsilon : a_2; r_1 \sqcup r_2]}$$

Non-interférence

On considère une expression e de type int^L avec un « trou » x marqué H :

$$(x \mapsto t) \vdash e : \mathsf{int}^L \qquad\qquad H \lhd t$$

Théorème de non-interférence

$$\text{Si } \begin{cases} \vdash v_1 : t \\ \vdash v_2 : t \end{cases} \text{ et } \begin{cases} e[x \Leftarrow v_1] \to^* v_1' \\ e[x \Leftarrow v_2] \to^* v_2' \end{cases} \text{ alors } v_1' = v_2'$$

Le résultat de l'évaluation de e ne dépend pas de la valeur d'entrée placée dans le trou.

Preuve de non-interférence

1. On définit une extension *ad hoc* du langage qui permet de raisonner sur les points communs et les différences entre deux programmes.
2. On vérifie que le système de types pour le langage étendu possède *subject reduction*.
3. On montre que la non-interférence pour le langage initial est une conséquence de *subject reduction*.

Calcul partagé

Le calcul partagé permet de considérer simultanément les exécutions indépendantes de deux expressions en gardant trace des sous-expressions qu'elles partagent.

Syntaxe

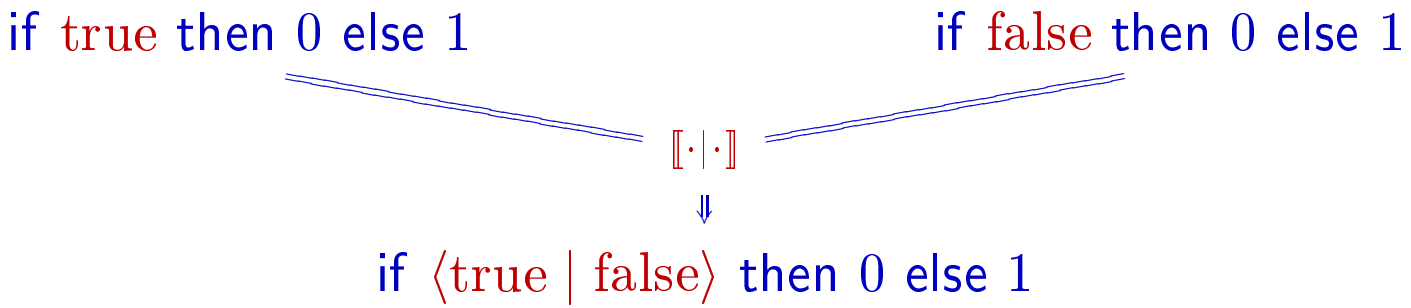
$$v ::= \ldots \mid \langle v \mid v \rangle \qquad\qquad e ::= \ldots \mid \langle e \mid e \rangle$$

On ne considère que les expressions où les $\langle \ldots \mid \ldots \rangle$ ne sont pas imbriqués.

Preuve de non-interférence

Codage

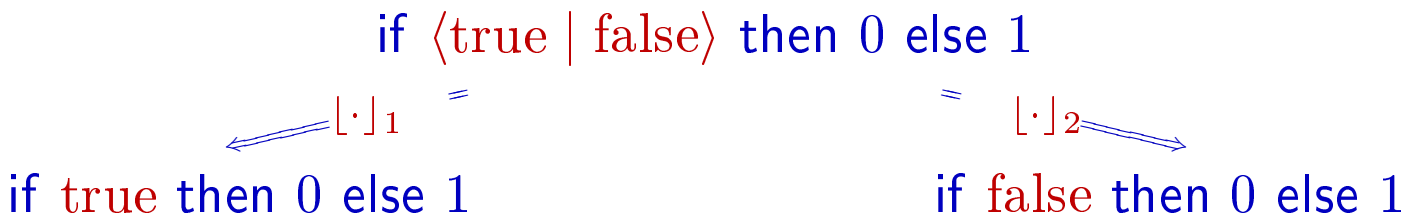
Une expression du calcul partagé permet de représenter une paire d'expressions du calcul simple en partageant certaines sous-expressions.



⟨if true then 0 else 1 | if false then 0 else 1⟩

Projections

Les projections $\lfloor\cdot\rfloor_1$ et $\lfloor\cdot\rfloor_2$ permettent de retrouver les deux expressions simples codées par une expression du calcul partagé.



Preuve de non-interférence

Réduction du calcul partagé

On dérive la sémantique du calcul partagé de celle du calcul simple : à chaque fois que des crochets bloquent la réduction, il faut les déplacer.

$$(\lambda x.e)\, v \to e[x \Leftarrow v] \qquad (\beta)$$

$$\langle v_1 \mid v_2 \rangle\, v \to \langle v_1 \lfloor v \rfloor_1 \mid v_2 \lfloor v \rfloor_2 \rangle \qquad \text{(lift-app)}$$

Exemples

$$\begin{aligned} \langle \lambda x.x \mid \lambda x.x+1 \rangle\, 3 &\to \langle (\lambda x.x)\, 3 \mid (\lambda x.x+1)\, 3 \rangle \\ &\to \langle 3 \mid (\lambda x.x+1)\, 3 \rangle \to \langle 3 \mid 4 \rangle \\ \langle \lambda x.x \mid \lambda x.x+1 \rangle\, \langle 3 \mid 2 \rangle &\to \langle (\lambda x.x)\, 3 \mid (\lambda x.x+1)\, 2 \rangle \\ &\to \langle 3 \mid (\lambda x.x+1)\, 3 \rangle \to \langle 3 \mid 3 \rangle \end{aligned}$$

Preuve de non-interférence

Simulation

Correction

Si $e \to e'$ alors $\begin{cases} \lfloor e \rfloor_1 \to^= \lfloor e' \rfloor_1 \\ \lfloor e \rfloor_2 \to^= \lfloor e' \rfloor_2 \end{cases}$

(calcul partagé) (calcul simple)

Complétude

Si $\begin{cases} e_1 \to^* v_1 \\ e_2 \to^* v_2 \end{cases}$ alors $[\![e_1 \mid e_2]\!] \to^* [\![v_1 \mid v_2]\!]$

(calcul simple) (calcul partagé)

Preuve de non-interférence

Typage de $\langle \ldots \mid \ldots \rangle$

Pour que $\langle v_1 \mid v_2 \rangle$ ait le type t, on impose que $H \lhd t$:

$$\frac{\Gamma \vdash v_1 : t \qquad \Gamma \vdash v_2 : t \qquad H \lhd t}{\Gamma \vdash \langle v_1 \mid v_2 \rangle : t}\ \textsc{Bracket}$$

Ainsi :

- Une valeur de type int^H peut être en entier k ou un crochet $\langle k_1 \mid k_2 \rangle$.
- Une valeur de type int^L ne peut être qu'un entier k.

Preuve de non-interférence

Subject reduction et non-interférence

Soit $(x \mapsto t) \vdash e : \mathsf{int}^L$ avec $H \lhd t$.

Subject-reduction

Si $\vdash e' : \mathsf{int}^L$ et $e' \to^* v'$ alors $\vdash v' : \mathsf{int}^L$

$$e' = e[x \Leftarrow v] \qquad\qquad v' = k$$

Non-interférence pour le calcul partagé

Si $\vdash v : t$ et $e[x \Leftarrow v] \to^* v'$ alors $\lfloor v' \rfloor_1 = \lfloor v' \rfloor_2$

Non-interférence

Soit $(x \mapsto t) \vdash e : \mathsf{int}^L$ avec $H \lhd t$.

Non-interférence pour le calcul partagé

Si $\quad \vdash v : t \quad$ et $\quad e[x \Leftarrow v] \to^* v' \quad$ alors $\quad \lfloor v' \rfloor_1 = \lfloor v' \rfloor_2$

$$v = \langle v_1 \mid v_2 \rangle \qquad\qquad v' = [\![v'_1 \mid v'_2]\!]$$

Non-interférence pour le calcul simple

$$\text{Si} \quad \begin{cases} \vdash v_1 : t \\ \vdash v_2 : t \end{cases} \quad \text{et} \quad \begin{cases} e[x \Leftarrow v_1] \to^* v'_1 \\ e[x \Leftarrow v_2] \to^* v'_2 \end{cases} \quad \text{alors} \quad v'_1 = v'_2$$

Extension du langage

On souhaite généralement étendre le langage étudié :

Pour accroître la richesse du langage Ajout de sommes, de paires. Étude d'un cas général pour des « primitives » des langages réels (opérations arithmétiques, de comparaison).

Pour mieux typer certains idiomes

$$e_1 \text{ finally } e_2 \hookrightarrow \text{bind } x = (e_1 \text{ handle } y \succ e_2;\ \text{raise } y) \text{ in } e_2;\ x$$

$$e_1 \text{ handle } x \succ e_2 \text{ reraise} \hookrightarrow e_1 \text{ handle } x \succ (e_2;\ \text{raise } x)$$

Extension du langage

Sommes et paires

$$t \quad ::= \quad \ldots \mid (t_1 + t_2)^\ell \mid t_1 \times t_2$$

Il n'est pas nécessaire de munir le constructeur de types $\times$ d'un niveau d'information propre. Les niveaux présents dans les types des composantes suffisent :

$$\begin{array}{lcl} \ell \lhd t_1 \times t_2 & \Leftrightarrow & \ell \lhd t_1 \wedge \ell \lhd t_2 \\ t_1 \times t_2 \blacktriangleleft \ell & \Leftrightarrow & t_1 \blacktriangleleft \ell \wedge t_2 \blacktriangleleft \ell \end{array}$$

Extension du langage

Primitives

Les primitives monomorphes telles que les opérations arithmétiques n'entrainent pas de difficulté :

$$\frac{\Gamma \vdash v_1 : \mathsf{int}^{\ell} \qquad \Gamma \vdash v_2 : \mathsf{int}^{\ell}}{pc, \Gamma \vdash v_1 + v_2 : \mathsf{int}^{\ell} \ [\partial\mathsf{Abs}]}$$

Le typage de primitives polymorphes qui « filtrent » la structure de leurs entrées nécessite une nouvelle forme de contraintes :

$$\frac{\Gamma \vdash v_1 : t \qquad \Gamma \vdash v_2 : t \qquad t \blacktriangleleft \ell}{pc, \Gamma \vdash v_1 = v_2 : \mathsf{bool}^{\ell} \ [\partial\mathsf{Abs}]} \qquad \frac{\Gamma \vdash v : t \qquad t \blacktriangleleft \ell}{pc, \Gamma \vdash \mathsf{hash}\ v : \mathsf{int}^{\ell} \ [\partial\mathsf{Abs}]}$$

$t \blacktriangleleft \ell$ contraint *tous* les niveaux d'information apparaissant dans t et ses sous termes à être inférieurs à ℓ.

Vers une extension du compilateur Caml

Le langage étudié permet de prendre en compte la quasi-totalité du langage Caml (à l'exception de la bibliothèque `threads`).

L'implantation d'un prototype est en cours. Elle nécessite de résoudre plusieurs problèmes liés à l'utilisation d'un système de types doté de sous-typage :

- Efficacité de l'algorithme d'inférence
- Lisibilité des types inférés
- Clarté des messages d'erreur
- ...

Inférence de types

Un algorithme d'inférence comporte deux parties relativement distinctes.

Un jeu de règles d'inférence On peut dériver de manière quasi-systématique un jeu de règles d'inférence d'un système tel que le nôtre.

$$\frac{\textsc{Ref}\quad \Gamma \vdash v : t \qquad pc \lhd t}{pc, \Gamma \vdash \mathsf{ref}\ v : t\ \mathsf{ref}^{\ell}\ [r]} \rightsquigarrow \frac{\textsc{Inf-Ref}\quad \Gamma, C \vdash v : \alpha}{\pi, \Gamma, C \cup \{\beta = \alpha\ \mathsf{ref}^{\lambda}, \pi \lhd \alpha\} \vdash \mathsf{ref}\ v : \beta\ [\rho]}$$

Un solveur Les schémas de types comportent des ensembles de contraintes. Il faut tester leur satisfiabilité et les simplifier.

Déclarations de types

```
type ('a, 'b) list =
    []
  | (::) of 'a * ('a, 'b) list
  level 'b


type ('a, 'b, 'c) queue = {
    mutable in: ('a, 'b) list;
    mutable out: ('a, 'b) list
  }
  level 'c
```

Exemple : listes

```
let rec length = function
    [] -> 0
  | _ :: l -> 1 + length l
```

val length : $\forall[\alpha \leq \beta].\ * \ \mathsf{list}^{\alpha} \rightarrow \mathsf{int}^{\beta}$

```
let rec iter f = function
    [] -> ()
  | x :: l -> f x; iter f l
```

val iter : $\forall[\sqcup\delta \leq \gamma].(\alpha \xrightarrow{\gamma\ [\delta]} *)^{\gamma} \rightarrow \alpha\,\mathsf{list}^{\gamma} \xrightarrow{\gamma\ [\delta]} \mathsf{unit}$

Exemple : queues

```
let push p elt =
  p.in <- elt :: p.in
```

val push : $\forall[\gamma \leq \beta].(\alpha, \beta)\,\mathsf{queue}^{\gamma} \to \alpha \xrightarrow{\gamma\ [*]} \mathsf{unit}$

```
let rec pop p = match p.out with
  hd :: tl -> p.out <- tl; hd
| [] -> match p.in with
          [] -> raise Empty
        | _ -> balance p; pop p
```

val pop : $\forall[\alpha \leq \alpha', \beta \lhd \alpha', \gamma \sqcup \pi \leq \beta].(\alpha, \beta)\,\mathsf{queue}^{\gamma} \xrightarrow{\pi\ [\mathsf{Empty}:\beta;*]} \alpha'$

Référence

[1] François Pottier and Vincent Simonet. Information flow inference for ML. In *Proceedings of the 29th ACM Symposium on Principles of Programming Languages (POPL'02)*. `http://cristal.inria.fr/~simonet/publis/fpottier-simonet-popl02.ps.gz`.