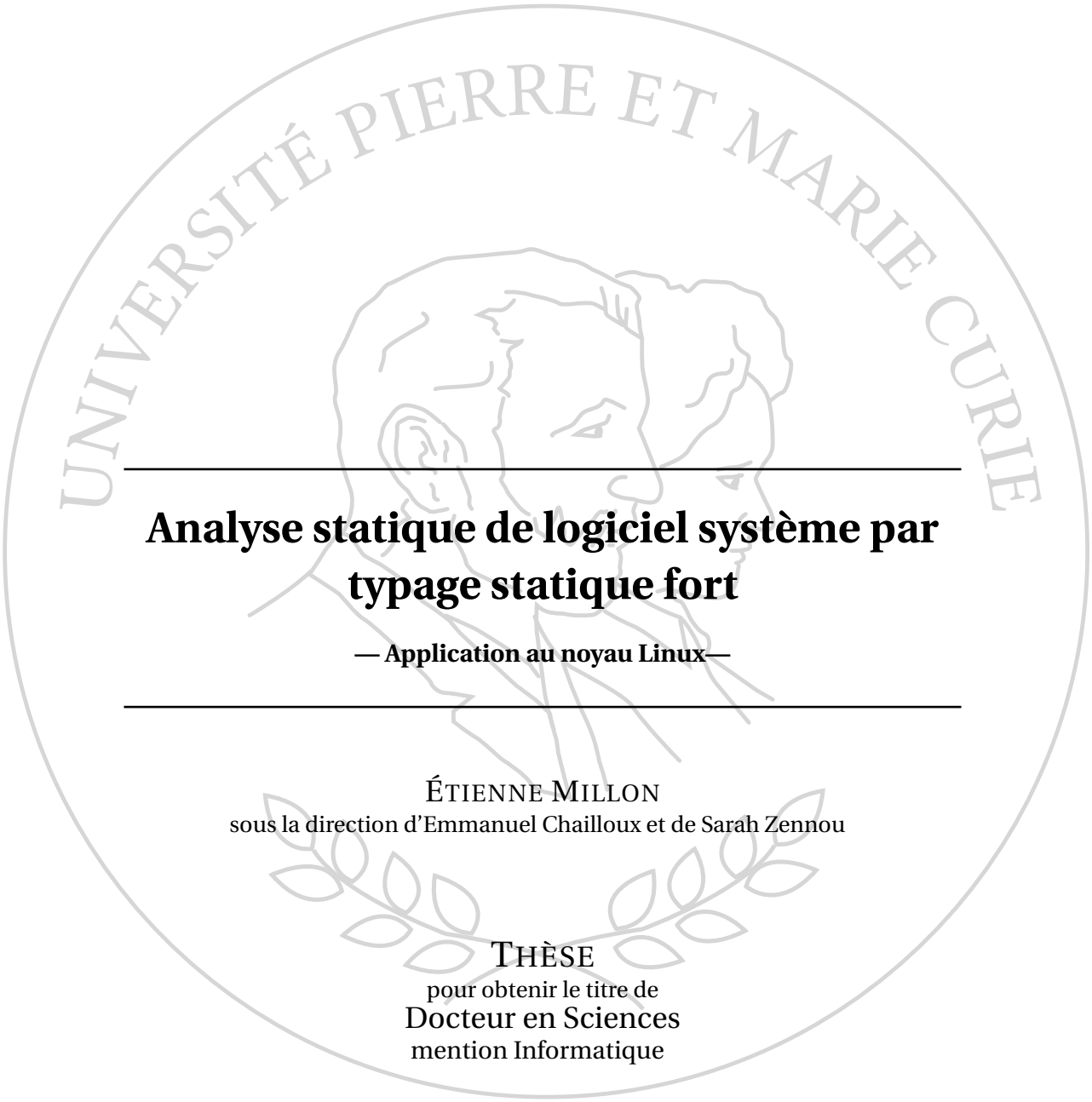
UNIVERSITÉ PIERRE ET MARIE CURIE
ÉCOLE DOCTORALE INFORMATIQUE, TÉLÉCOMMUNICATIONS ET ÉLECTRONIQUE



Analyse statique de logiciel système par typage statique fort

— Application au noyau Linux—

ÉTIENNE MILLON
sous la direction d'Emmanuel Chailloux et de Sarah Zennou

THÈSE
pour obtenir le titre de
Docteur en Sciences
mention Informatique

Version du 6 juin 2013

TABLE DES MATIÈRES

CHAPITRE 1

INTRODUCTION

Communication, audiovisuel, transports, médecine : tout ces domaines se sont transformés dans les dernières décennies, en particulier grâce à la révolution numérique. En effet le plus petit appareil électrique contient maintenant des composants matériels programmables.

En 2013, on pense bien sûr aux téléphones portables dont la fonctionnalité et la complexité les rapprochent des ordinateurs de bureau. Par exemple, le système d'exploitation Android de Google est basé sur le noyau Linux, destiné à la base aux micro-ordinateurs.

Le noyau d'un système d'exploitation est chargé de faire l'intermédiaire entre le matériel (processeur, mémoire, périphériques, ...) et les applications exécutées sur celui-ci (par exemple un navigateur web, une calculatrice ou un carnet d'adresses). Il doit aussi garantir la sécurité de celles-ci.

En tant qu'intermédiaire de confiance, le noyau a un certain nombre de responsabilités et est le seul à avoir accès à certaines informations sensibles. Il est capital de s'assurer qu'il est bien le seul à pouvoir y accéder. En particulier, il faut pouvoir vérifier que les requêtes faites par l'utilisateur au noyau ne peuvent pas volontairement ou involontairement détourner ce dernier et lui faire fuiter des informations confidentielles.

Le problème est que comme tous les logiciels, les noyaux de système d'exploitation sont écrits par des humains, qui ne sont pas parfaits. Loin de là : on estime qu'avant relecture, 1000 lignes de code contiennent entre 5 et 100 erreurs de programmation en moyenne. Les activités de relecture et de débogage ont beau prendre la majeure partie du temps de développement, il est facile de laisser passer des défauts de programmation.

Une technique efficace est de réaliser des tests, c'est-à-dire exécuter le programme sous un environnement contrôlé. On peut alors détecter des comportements non désirés. Mais même avec des tests les plus exhaustifs il n'est pas possible de couvrir tous les cas d'utilisation.

Une autre approche est d'analyser le code source du programme avant de l'exécuter et de refuser de lancer les programmes qui contiennent des erreurs, quitte à limiter la possibilité d'en écrire certains. C'est l'analyse statique de programmes.

Une des techniques d'analyse statique les plus répandues est le typage statique, qui consiste à associer à chaque morceau de programme, une étiquette décrivant quel genre de valeur sera produite par son évaluation. Par exemple, si n est le nom d'une variable entière, alors $n+2$ produira toujours une valeur entière.

Pour garantir l'isolation d'un noyau de système d'exploitation, un des points cruciaux est de restreindre la manière dont sont traitées les informations provenant des programmes

utilisateur. Le but de cet thèse est de montrer que le typage statique peut être utilisé pour détecter et interdire ces manipulations dangereuses.

1.1 Rôle d'un système d'exploitation

Un ordinateur est constitué de nombreux composants matériels : microprocesseur, mémoire, et divers périphériques. Et au niveau de l'utilisateur, des dizaines de logiciels permettent d'effectuer toutes sortes de calculs et de communications. Le système d'exploitation permet de faire l'interface entre ces niveaux d'abstraction.

Au cours de l'histoire des systèmes informatiques, la manière de les programmer a beaucoup évolué. Au départ, les programmeurs avaient accès au matériel dans son intégralité : toute la mémoire pouvait être accédée, toutes les instructions pouvaient être utilisées.

Néanmoins c'est un peu restrictif, puisque cela ne permet qu'à une personne d'interagir avec le système. Dans la seconde moitié des années 60, sont apparus les premiers systèmes "à temps partagé", permettant à plusieurs utilisateurs de travailler en même temps.

Permettre l'exécution de plusieurs programmes en même temps est une idée révolutionnaire, mais elle n'est pas sans difficultés techniques : en effet les ressources de la machine doivent être aussi partagées entre les utilisateurs et les programmes. Par exemple, plusieurs programmes vont utiliser le processeur les uns à la suite des autres ; et chaque programme aura à sa disposition une partie de la mémoire principale, ou du disque dur.

Si plusieurs programmes s'exécutent de manière concurrente sur le même matériel, il faut s'assurer que l'un ne puisse pas écrire dans la mémoire de l'autre, ou que les deux n'utilisent pas la carte réseau en même temps. Ce sont des rôles du système d'exploitation.

Ainsi, au lieu d'accéder directement au matériel via des instructions de bas niveau, les programmes communiquent avec le noyau, qui centralise donc les appels au matériel, et abstrait certaines opérations.

Par exemple, comparons ce qui se passe concrètement lors de la copie de données depuis un cédérom ou une clef USB.

- Dans le cas du cédérom, il faut interroger le bus SATA, interroger le lecteur sur la présence d'un disque dans le lecteur, activer le moteur, calculer le numéro de trame des données sur le disque, demander la lecture, puis déclencher une copie de la mémoire.
- Avec une clef, il faut interroger le bus USB, rechercher le bon numéro de périphérique, le bon numéro de canal dans celui-ci, lui appliquer une commande de lecture au bon numéro de bloc, puis copier la mémoire.

Ces deux opérations, bien qu'elles aient la même intention (copier de la mémoire depuis un périphérique amovible), ne sont pas effectuées en extension de la même manière. C'est pourquoi le système d'exploitation fournit les notions de fichier, lecteur, etc : le programmeur n'a plus qu'à utiliser des commandes de haut niveau ("monter un lecteur", "ouvrir un fichier", "lire dans un fichier") et selon le type de lecteur, le système d'exploitation effectuera les actions appropriées.

En résumé, un système d'exploitation est l'intermédiaire entre le logiciel et le matériel, et en particulier est responsable de la gestion de la mémoire, des périphériques et des processus. Les détails d'implantation ne sont pas présentés à l'utilisateur ; à la place il manipule des abstractions, comme la notion de fichier. Pour une explication détaillée du concept de système d'exploitation ainsi que des cas d'étude, on pourra se référer à [Tan07].

Mode du processeur	Privilège (code)	Privilège (données)	Accès possible
Noyau	Noyau	Noyau	☑
Noyau	Noyau	Utilisateur	☑
Noyau	Utilisateur	Noyau	☑
Noyau	Utilisateur	Utilisateur	☑
Utilisateur	Noyau	Noyau	□
Utilisateur	Noyau	Utilisateur	□
Utilisateur	Utilisateur	Noyau	□
Utilisateur	Utilisateur	Utilisateur	☑

FIGURE 1.1: Règles d'exécution de code et d'accès à la mémoire.

1.2 Séparation entre noyau et espace utilisateur

Puisque le noyau est garant du bon fonctionnement du système, il ne doit pas pouvoir être manipulé directement par l'utilisateur ou les programmes exécutés. Ainsi, il est nécessaire de mettre en place des protections entre les espaces noyau et utilisateur.

Au niveau matériel, on utilise la notion de *niveaux de privilèges* pour déterminer s'il est possible d'exécuter une instruction.

D'une part, le processeur contient un niveau de privilège intrinsèque. D'autre part, chaque zone mémoire contenant du code ou des données possède également un niveau de privilège minimum nécessaire. L'exécution d'une instruction est alors possible si et seulement si le niveau de privilège du processeur est supérieur à celui de l'instruction et des opérandes mémoires qui y sont présentes [1].

Par exemple, supposons qu'un programme utilisateur contienne l'instruction `mov %eax, 0x54c3b2f7`. Celle-ci consiste à déplacer le contenu du registre EAX vers l'adresse mémoire `0x54c3b2f7`. Puisque cette adresse est dans l'espace utilisateur (la raison est décrite dans le chapitre 2), aucune erreur de protection mémoire n'est déclenchée.

Ainsi, pour une instruction manipulant des données en mémoire, les accès possibles sont décrits dans la figure 1.1. En cas d'impossibilité (signalée par un □), une erreur se produit et l'exécution s'arrête.

En plus de cette vérification, certains types d'instructions sont explicitement réservés au mode le plus privilégié : par exemple les lectures ou écritures sur des ports matériels, ou celles qui permettent de définir les niveaux de privilèges des différentes zones mémoire.

Les programmes utilisateur ne pouvant pas accéder à ces instructions de bas niveau, ils sont très limités dans ce qu'ils peuvent faire. Puisque l'interaction avec le matériel (comme l'écriture sur un disque dur) se fait uniquement via des instructions privilégiées, ils sont limités à l'utilisation du processeur et de la mémoire, permettant uniquement de réaliser des calculs.

Pour utiliser le matériel ou accéder à des abstractions de haut niveau (comme créer un nouveau processus), ils doivent donc passer par l'intermédiaire du noyau. La communication entre le noyau et les programmes utilisateur est constituée par le mécanisme des *appels système*.

Lors d'un appel système, une fonction du noyau est invoquée (en mode noyau) avec des paramètres provenant de l'utilisateur. Il faut donc être particulièrement précautionneux dans le traitement de ces données.

1. Ici "supérieur" est synonyme de "plus privilégié". Dans l'implantation d'Intel présentée dans le chapitre 2, les niveaux sont numérotés de 0 à 3 où le niveau 0 est le plus privilégié.

Par exemple, considérons un appel système de lecture depuis un disque : on passe au noyau les arguments (d, o, n, a) où d est le nom du disque, o (pour *offset*) l'adresse sur le disque où commencer la lecture, n le nombre d'octets à lire et a l'adresse en mémoire où commencer à stocker les résultats.

Dans le cas d'utilisation prévu, le noyau va copier la mémoire lue dans a. Le processeur est en mode noyau, en train d'exécuter une instruction du noyau manipulant des données utilisateur. D'après la figure 1.1 aucune erreur ne se produit.

Mais même si ce cas ne produit pas d'erreur à l'exécution, il est tout de même erroné. En effet, si on passe à l'appel système une adresse a faisant partie de l'espace noyau, que se passe-t-il ?

L'exécution est presque identique : au moment de la copie on est en mode noyau, en train d'exécuter une instruction du noyau manipulant des données noyau. Encore une fois il n'y a pas d'erreur à l'exécution.

On peut donc écrire n'importe où en mémoire. De même, une fonction d'écriture sur un disque (et lisant en mémoire) permettrait de lire de la mémoire du noyau. À partir de ces primitives, on peut accéder aux autres processus exécutés, ou détourner l'exécution vers du code arbitraire. L'isolation est totalement brisée à cause de ces appels système.

La cause de ceci est qu'on a accédé à la mémoire en testant les privilèges du noyau au lieu de tester les privilèges de celui qui a fait la requête (l'utilisateur). Ce problème est connu sous le nom de *confused deputy problem* [Har88].

La bonne manière d'implanter un appel système est donc d'interdire le déréférencement direct des pointeurs dont la valeur peut être contrôlée par l'utilisateur. Dans le cas du passage par adresse d'un argument, il aurait fallu vérifier à l'exécution que celui-ci a bien les mêmes privilèges que l'appelant.

Il est facile d'oublier d'ajouter cette vérification, puisque le cas "normal" fonctionne. Avec ce genre d'exemple on voit comment les bugs peuvent arriver si fréquemment et pourquoi il est aussi capital de les détecter avant l'exécution.

1.3 Systèmes de types

La plupart des langages de programmation incorporent la notion de type, dont un des buts est d'empêcher de manipuler des données incompatibles entre elles.

Au niveau du langage machine, les seules données qu'un ordinateur manipule sont des nombres. Selon les opérations effectuées, ils seront interprétés comme des entiers, des adresses mémoires, ou des caractères. Pourtant il est clair que certaines opérations n'ont pas de sens : par exemple, multiplier un nombre par une adresse, ou déréférencer le résultat d'une division sont des comportements qu'on voudrait pouvoir empêcher.

En un mot, le but du typage est de classifier les objets et de restreindre les opérations possibles selon la classe d'un objet : en somme, "ne pas ajouter des pommes et des oranges". Le modèle qui permet cette classification est appelé *système de types* et est en général constitué d'un ensemble de *règles de typage*, comme "un entier plus un entier égale un entier".

Typage dynamique Dans ce cas, chaque valeur manipulée par le programme est décorée d'une étiquette définissant comment interpréter la valeur en question. Les règles de typage sont alors réalisées à l'exécution. Par exemple, l'opérateur "+" vérifie que ces deux opérandes ont une étiquette "entier", et construit alors une valeur obtenue en faisant l'addition des deux valeurs, avec une étiquette "entier". Par exemple, le langage Python [3] utilise cette stratégie.

Typage statique Dans ce cas on fait les vérifications à la compilation. En quelque sorte, l'approche dynamique est pessimiste, puisqu'elle demande de traiter à chaque étape le cas où les types ne sont pas corrects. Intuitivement, dans le cas où toutes les fonctions se comportent bien, faire la vérification est inutile. Pour vérifier ceci, on donne à chaque fonction un contrat comme "si deux entiers sont passés, et que la fonction renvoie une valeur, alors cette valeur sera un entier". Cet ensemble de contrats peut être vérifié statiquement par le compilateur, à l'aide d'un système de types statique.

Par exemple, on peut dire que l Ainsi l'opérateur "+" a pour type (INT, INT) → INT. Cela veut dire que si on lui passe deux entiers (INT INT), alors la valeur obtenue est également un entier.

Typage fort ou faible Contrairement à la distinction claire entre typage statique ou dynamique, la séparation entre typage fort et faible est moins nette. En poussant à l'extrême, les systèmes de types forts permettent d'éliminer totalement la nécessité de réaliser des tests de typage. Mais souvent ce n'est pas le cas, car il peut y avoir des constructions au sein du langage qui permettent de contourner le système de types, comme un opérateur de transtypage.

Polymorphisme Parfois, il est trop restrictif de donner un unique type à une fonction. Si on considère une fonction ajoutant un élément à une liste, ou une autre triant un tableau en place, leur type doit-il faire intervenir le type des éléments manipulés ?

En première approximation, on peut imaginer fournir une version du code par type de données à manipuler. C'est la solution retenue par les premières versions du langage Pascal, ce qui rendait très difficile l'écriture de bibliothèques [Ker81]. On parle alors de monomorphisme.

Une autre manière de procéder est d'autoriser plusieurs fonctions à avoir le même nom, mais avec des types d'arguments différents. Par exemple, on peut définir séparément l'addition entre deux entiers, entre deux flottants, ou entre un entier et un flottant. Selon les informations connues à la compilation, la bonne version sera choisie. C'est ainsi que fonctionnent les opérateurs en C++. On parle de polymorphisme *ad hoc*, ou de surcharge.

La dernière possibilité est le polymorphisme paramétrique, qui consiste à utiliser le même code quelque soit le type des arguments. Dans ce cas, on utilise une seule fonction pour traiter une liste d'entiers ou une liste de flottants, par exemple. Au lieu d'associer à chaque fonction un type, dans certains cas on lui associe un type paramétré, instanciable en un type concret. Dans le cas des fonctions de traitement de liste, l'idée est que lorsqu'on ne touche pas aux éléments, alors le traitement est valable quelque soit leur type. Cette technique a été décrite en premier dans [Mil78].

Pour un tour d'horizon de différents systèmes types de statiques, avec en particulier du polymorphisme, on pourra se référer à [Pie02].

1.4 Langages d'implantation

Puisque notre but est d'analyser du code provenant de systèmes d'exploitation, il est nécessaire de s'intéresser aux langages de programmation dans lesquels ils sont écrits.

Assembleur Le noyau d'un système d'exploitation nécessite d'accéder au matériel, donc il est naturellement bas niveau. Pour accéder aux fonctionnalités spécifiques de chaque processeur, il est donc nécessaire d'en implanter une partie dans le langage d'assemblage natif

de chaque architecture. Historiquement, les premiers système d'exploitations étaient entièrement écrits en assembleur. Cela est de plus en plus rare, sauf dans les cas où les ressources sont trop limitées pour exécuter du code compilé, comme dans les systèmes embarqués ou temps-réel.

C Le système Unix, développé à partir de 1969, a tout d'abord été développé en assembleur sur un mini-ordinateur PDP-7, puis a été porté sur d'autres architectures matérielles. Pour aider ce portage, il a été nécessaire de créer un "assembleur portable", le langage C [KR88, ISO99]. Son but est de fournir des abstractions au dessus du langage d'assemblage. Les structures de contrôle (`if`, `while`, `for`) permettent d'utiliser la programmation structurée, c'est-à-dire en limitant l'utilisation de l'instruction `goto`. Les types de données sont également abstraits de la machine : ainsi, `int` désigne un entier machine, indépendamment de sa taille concrète. Son système de types est assez rudimentaire : toutes les formes de transtypage sont acceptées, certaines conversions sont insérées automatiquement par le compilateur, et la plupart des abstractions fournies par le langage sont perméables. Le noyau Linux est écrit dans un dialecte du langage C.

C++ Le manque de modularité de C a conduit à la création d'un nouveau langage plus riche autour de celui-ci. Dans les années 1980, C++ est apparu comme un C amélioré, avec de nombreuses fonctionnalités supplémentaires comme un modèle objet avec des classes, du polymorphisme *ad hoc* et un système de *templates* puissant permettant d'implanter une forme de polymorphisme universel. Microsoft Windows, initialement écrit en C, s'est enrichi de plus en plus de code C++, en faisant un des principaux systèmes d'exploitation écrits dans ce langage.

Ada Le langage C est bas niveau, mais son système de types n'est pas suffisamment évolué pour le rendre plus sûr que le langage d'assemblage vers lequel il est compilé. Dans les années 1970, le département de la défense américain (*DoD*) a créé un comité dans le but de créer un nouveau langage pour ses systèmes temps-réels embarqués. Le résultat est Ada, un langage impératif avec un système de types statique plus contraignant mais plus sûr que celui de C. Par exemple, deux types entiers différents (comme les entiers de 16 bits et ceux de 32 bits) ne sont pas compatibles. Il faut ajouter une conversion explicite dans le code, qui se traduira par un test à l'exécution et éventuellement par une erreur `Constraint_Error`. Bien qu'il ne se soit pas beaucoup développé dans les applications grand public, Ada garde toujours une place de choix dans l'implantation de systèmes embarqués, notamment dans les applications spatiales et de défense.

Divers Il est bien sûr possible d'implanter un système d'exploitation dans n'importe quel langage, si celui-ci peut être compilé vers du langage machine et qu'il permet facilement d'inclure des fragments d'assembleur pour les parties de bas niveau. Néanmoins il s'agit en général de travaux de recherche. On peut ainsi trouver des systèmes d'exploitation écrits en C# [HLA+05], Haskell [HJLT05] ou Ocaml [MMR+13].

1.5 Langages d'analyse

Les langages décrits précédemments sont faits pour être facilement écrits par des programmeurs humains. En général ils sont ambigüs ou ont des comportements implicites. Pour analyser du code source, il est plus pratique d'avoir une représentation intermédiaire plus

simple afin d'avoir moins de traitements dupliqués. Dans de nombreux projets, des sous-ensembles de C ont été définis pour aller dans ce sens.

Middle-ends

Les premiers candidats sont bien entendu les représentations intermédiaires utilisées dans les compilateurs C. Elles ont l'avantage d'accepter en plus du C standard, les diverses extensions (GNU, Microsoft, Plan9) utilisées par la plupart des logiciels. En particulier, le noyau Linux repose fortement sur les extensions GNU.

GCC utilise une représentation interne nommée GIMPLE[Mer03]. Il s'agit d'une structure d'arbre écrite en C, reposant sur de nombreuses macros afin de cacher les détails d'implantation interne pouvant varier entre deux versions. Cette représentation étant réputée difficile à manipuler, le projet MELT[Sta11] permet de générer un greffon de compilateur à partir d'un dialecte de Lisp.

LLVM [LA04] est un compilateur développé par la communauté puis sponsorisé Apple. À la différence de GCC, sa base de code est écrite en C++. Il utilise une représentation intermédiaire qui peut être manipulée soit sous forme d'une structure de données C++, soit d'un fichier de code-octet compact, soit sous forme textuelle.

Cmm est une représentation interne utilisée pour la génération de code lors de la compilation d'Objective Caml [1] , et disponible dans les sources du compilateur sous le chemin `asmcomp/cmm.mli` (il s'agit donc d'une structure de données OCaml). Ce langage a l'avantage d'être très restreint, mais malheureusement il n'existe pas directement de traducteur permettant de compiler C vers Cmm.

C- - [PJNO97] [7], dont le nom est inspiré du précédent, est un projet qui visait à unifier les langages intermédiaires utilisés par les compilateurs. L'idée est que si un front-end peut émettre du C- - (sous forme de texte), il est possible d'obtenir du code machine efficace. Le compilateur Haskell GHC, par exemple, utilise une représentation intermédiaire très similaire à C- -.

Langages intermédiaires ad hoc

Comme le problème de construire une représentation intermédiaire adaptée à une analyse statique n'est pas nouveau, plusieurs projets ont déjà essayé d'y apporter une solution. Puisque qu'ils sont développés en parallèle des compilateurs, le support des extensions est en général moins important dans ces langages.

CIL [NMRW02] [6] est une représentation en OCaml d'un programme C, développée depuis 2002. Grâce à un mécanisme de greffons, elle permet de prototyper rapidement des analyses statiques de programmes.

Compcert est un projet qui vise à produire un compilateur certifié pour C. C'est-à-dire que le fait que les transformations conservent la sémantique est prouvé. Il utilise de nombreux langages intermédiaires, dont CIL. Pour le front-end, le langage se nomme Clight[BDL06]. Les passes de middle-end, quant à elles, utilisent Cminor[AB07].

Newspeak [HL08] est un langage intermédiaire développé par EADS Innovation Works, et qui est spécialisé dans l'analyse de valeurs par interprétation abstraite. Il sera décrit plus en détails dans la section 7.1.

La solution retenue : SAFESPEAK

Dans la suite de cette thèse nous nous appuierons sur SAFESPEAK, un langage intermédiaire minimal, pour exprimer les règles de typage de la manière la plus simple possible.

1.6 Plan de la thèse

Cette thèse comporte trois parties.

La partie I présente le contexte de ces travaux. Le fonctionnement général d'un système d'exploitation y est détaillé, et les problèmes de manipulation de pointeurs contrôlés par l'utilisateur y sont introduits. On fait ensuite un tour d'horizon des techniques existantes permettant de traiter ce problème. Cette partie se conclut par la proposition au cœur de cette thèse.

La partie II décrit notre solution : SAFESPEAK, un langage impératif. On y décrit sa syntaxe, sa sémantique ainsi qu'un système de types statiques. On l'étend ensuite pour capturer les problèmes d'adressage mémoire présents dans les systèmes d'exploitation en ajoutant des pointeurs contrôlés par l'utilisateur. Le système de types est également étendu. Pour chacune de ces variantes, on établit la propriété de sûreté de typage reliant la sémantique d'exécution aux types statiques.

La partie III documente la démarche expérimentale associée à ces travaux. L'implantation du système de types est décrite, afin que la manière de transformer automatiquement du code C en SAFESPEAK. Un cas d'étude est déroulé, consistant d'un bug ayant touché le noyau Linux. Il est démontré que le système de type capture précisément ce genre d'erreur de programmation. Enfin, les possibilités d'extension tant théoriques qu'expérimentales sont présentées.

Première partie

Méthodes formelles pour la sécurité

Après avoir décrit le contexte général de ces travaux, nous on décrivons les enjeux.

Le chapitre 2 explore plus en détail le fonctionnement d'un système d'exploitation, y compris la séparation du code en plusieurs niveaux de privilèges. L'architecture Intel 32 bits est prise comme support. En particulier, le mécanisme des appels système est décrit et on montre qu'une implantation naïve de la communication entre espaces utilisateur et noyau casse toute isolation.

Le chapitre 3 consiste en un tour d'horizon des techniques existantes en analyses de programmes. Ces analyses se centrent autour des problèmes liés à la vérification de code système ou embarqué, y compris le problème de manipulation mémoire évoqué dans le chapitre 2.

CHAPITRE 2

SYSTÈMES D'EXPLOITATION

Le système d'exploitation est le programme qui permet à un système informatique d'exécuter d'autre programmes. Son rôle est donc capital et ses responsabilités multiples. Dans ce chapitre, nous allons voir quel est son rôle, et comment il peut être implanté. Pour ce faire, nous étudierons l'exemple d'une architecture Intel 32 bits, et d'un noyau Linux 2.6.

2.1 Architecture et assembleur Intel

L'implantation d'un système d'exploitation est très proche du matériel sur lequel il s'exécute. Pour étudier une implantation en particulier, voyons ce que permet le matériel lui-même.

Dans cette section nous décrivons le fonctionnement d'un processeur utilisant une architecture Intel 32 bits. Les exemples de code seront écrits en syntaxe AT&T, celle que comprend l'assembleur GNU.

La référence pour la description de l'assembleur Intel est la documentation du constructeur [Int10] ; une bonne explication de l'agencement dans la pile peut aussi être trouvée dans [One96].

Pour faire des calculs, le processeur est composé de registres, qui sont des petites zones de mémoire interne, et peut accéder à la mémoire principale.

La mémoire principale contient divers types des données :

- le code des programmes à exécuter
- les données à disposition des programmes
- la pile d'appels

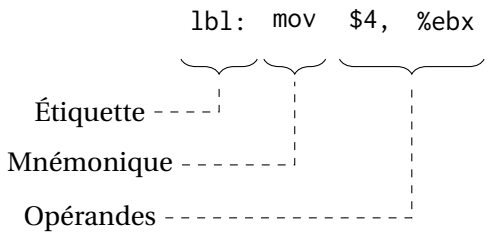
La pile d'appels est une zone de mémoire qui est notamment utilisée pour tenir une trace des calculs en cours. Par exemple, c'est ici que seront stockées les données propres à chaque fonction appelée : paramètres, adresse de retour et variables locales. La pile est manipulée par un pointeur de pile (*stack pointer*), qui est l'adresse du "haut de la pile". On peut la manipuler en empilant des données (les placer au niveau du pointeur de pile et déplacer celui si) ou dépilant des données (déplacer le pointeur de pile dans l'autre sens et retourner la valeur présente à cet endroit).

L'état du processeur est défini par la valeur de ses registres, qui sont des petites zones de mémoire interne (quelques bits chacun). Par exemple, la valeur du pointeur de pile est stockée dans ESP. Le registre EBP, couplé à ESP sert à adresser les variables locales et paramètres d'une fonction, comme ce sera expliqué dans la section 2.2.

L'adresse de l'instruction courante est accessible dans le registre EIP.

En plus de ces registres spéciaux, le processeur possède de nombreux registres génériques, qui peuvent être utilisés pour réaliser des calculs intermédiaires. Ils sont nommés EAX, EBX, ECX, EDX, ESI et EDI. Ils peuvent être utilisés pour n'importe quel type d'opération, mais certains sont spécialisés : par exemple il est plus efficace d'utiliser EAX en accumulateur, ou ECX en compteur.

Les calculs sont décrits sous forme d'une suite d'instructions. Chaque instruction est composée d'un mnémonique et d'une liste d'opérandes. Les mnémoniques (`mov`, `call`, `sub`, etc) définissent un type d'opération à appliquer sur les opérandes. L'instruction peut aussi être précédée d'une étiquette, qui correspondra à son adresse.



Ces opérandes peuvent être de plusieurs types :

- un nombre, noté $4
- le nom d'un registre, noté `%eax`
- une opérande mémoire, c'est-à-dire le contenu de la mémoire à une adresse effective. Cette adresse effective peut être exprimée de plusieurs manières :
  - directement : `addr`
  - indirectement : `(%ecx)`. L'adresse effective est le contenu du registre.
  - "base + déplacement" : `4(%ecx)`. L'adresse effective est le contenu du registre plus le déplacement (4 ici).

En pratique il y a des modes d'adressage plus complexes, et toutes les combinaisons ne sont pas possibles, mais ceux-ci suffiront à décrire les exemples suivants :

- `mov src, dst` copie le contenu de `src` dans `dst`.
- `add src, dst` calcule la somme des contenus de `src` et `dst` et place ce résultat dans `dst`.
- `push src` place `src` sur la pile, c'est-à-dire que cette instruction enlève au pointeur de pile ESP la taille de `src`, puis place `src` à l'adresse mémoire de la nouvelle valeur ESP.
- `pop src` réalise l'opération inverse : elle charge le contenu de la mémoire à l'adresse ESP dans `src` puis incrémente ESP de la taille correspondante.
- `jmp addr` saute à l'adresse `addr` : c'est l'équivalent de `mov addr, %eip`.
- `call addr` sert aux appels de fonction : cela revient à `push %eip` puis `jmp addr`.
- `ret` sert à revenir d'une fonction : c'est l'équivalent de `pop %eip`.

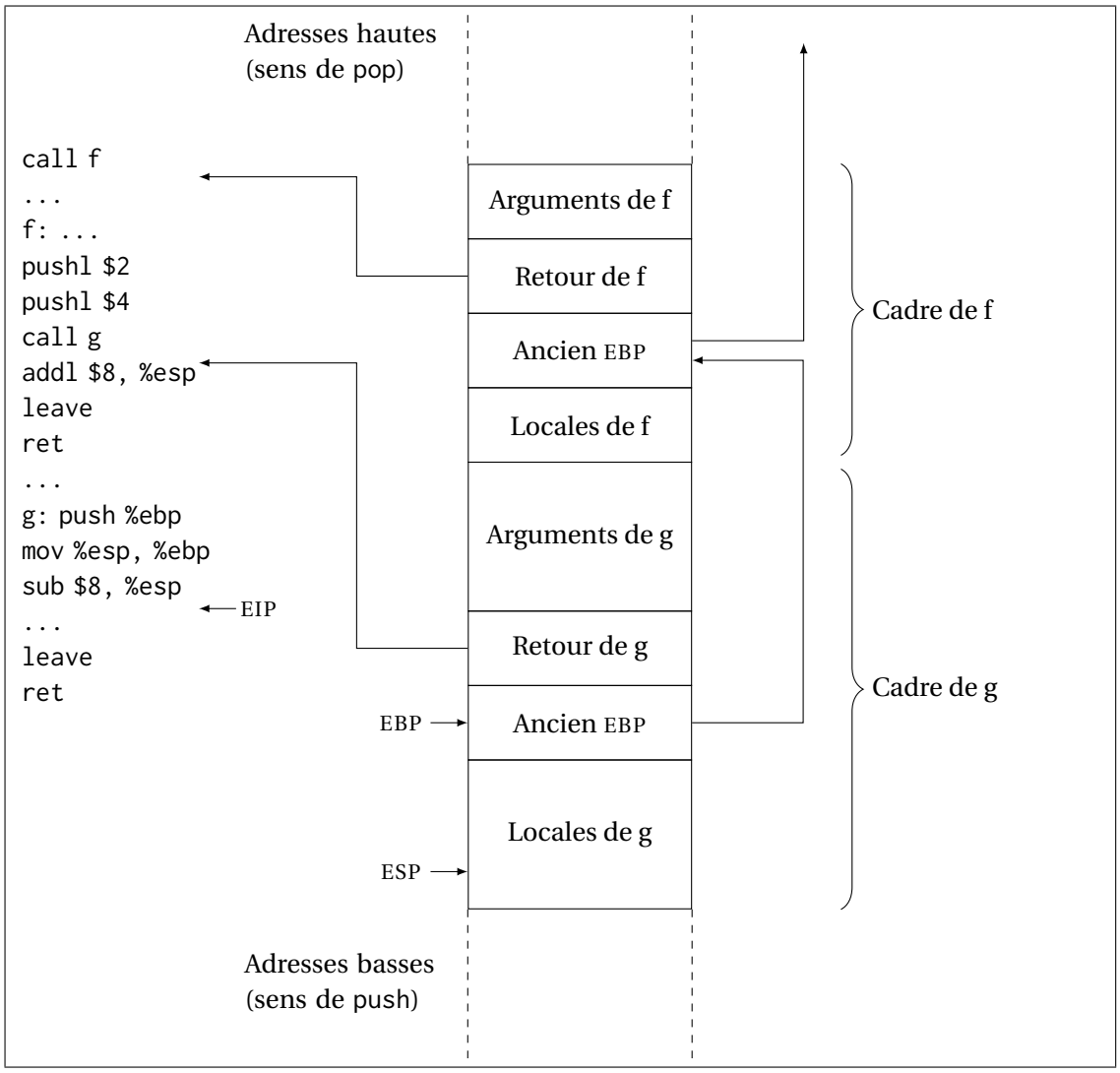


FIGURE 2.1: Cadres de pile.

2.2 Fonctions et conventions d'appel

Dans le langage d'assemblage, il n'y a pas de notion de fonction ; mais `call` et `ret` permettent de sauvegarder et de restaurer une adresse de retour, ce qui permet de faire un saut et revenir à l'adresse initiale. Ce système permet déjà de créer des procédures, c'est-à-dire des fonctions sans arguments ni valeur de retour.

Pour gérer ceux-ci, il faut que les deux morceaux (appelant et appelé) se mettent d'accord sur une convention d'appel commune. La convention utilisée sous GNU/Linux est appelée *cdecl* et possède les caractéristiques suivantes :

- la valeur de retour d'une fonction est stockée dans EAX
- EAX, ECX et EDX peuvent être écrasés sans avoir à les sauvegarder
- les arguments sont placés sur la pile (et enlevés) par l'appelant. Les paramètres sont empilés de droite à gauche.

Pour accéder à ses paramètres, une fonction peut donc utiliser un adressage relatif à ESP. Cela peut fonctionner, mais cela complique les choses si elle contient aussi des variables lo-

cales. En effet, les variables locales sont placées sur la pile, au dessus des (c'est-à-dire, empilées après) paramètres, augmentant le décalage.

Pour simplifier, la pile est organisée en cadres logiques : chaque cadre correspond à un niveau dans la pile d'appels de fonctions. Si f appelle g, qui appelle h, il y aura dans l'ordre sur la pile le cadre de f, celui de g puis celui de h.

Ces cadres sont chainés à l'aide du registre EBP : à tout moment, EBP contient l'adresse du cadre de l'appelant.

Prenons exemple sur la figure 2.1 : pour appeler `g(4,2)`, `f` empile les arguments de droite à gauche. L'instruction `call g` empile l'adresse de l'instruction suivante sur la pile puis saute au début de g.

Au début de la fonction, les trois instructions permettent de sauvegarder l'ancienne valeur de EBP, faire pointer EBP à une endroit fixe dans le cadre de pile, puis allouer 8 octets de mémoire pour les variables locales.

Dans le corps de la fonction g, on peut donc se référer aux variables locales par `-4(%ebp)`, `-8(%ebp)`, etc, et aux arguments par `8(%ebp)`, `12(%ebp)`, etc.

À la fin de la fonction, l'instruction `leave` est équivalente à `mov %ebp, %esp` puis `pop %ebp` et permet de défaire le cadre de pile, laissant l'adresse de retour en haut de pile. Le `ret` final la dépile et y saute.

2.3 Tâches, niveaux de privilèges

Sans mécanisme particulier, le processeur exécuterait uniquement une suite d'instruction à la fois. Pour lui permettre d'exécuter plusieurs tâches, un système de partage du temps existe.

À des intervalles de temps réguliers, le système est programmé pour recevoir une interruption. C'est une condition exceptionnelle (au même titre qu'une division par zéro) qui fait sauter automatiquement le processeur dans une routine de traitement d'interruption. À cet endroit le code peut sauvegarder les registres et restaurer un autre ensemble de registres, ce qui permet d'exécuter plusieurs tâches de manière entrelacée. Si l'alternance est assez rapide, cela peut donner l'illusion que les programmes s'exécutent en parallèle. Comme l'interruption peut survenir à tout moment, on parle de multitâche préemptif.

En plus de cet ordonnancement de processus, l'architecture Intel permet d'affecter des niveaux de privilège à ces tâches, en restreignant le type d'instructions exécutables, ou en donnant un accès limité à la mémoire aux tâches de niveaux moins élevés.

Il y a 4 niveaux de privilèges (nommés aussi *rings*) : le *ring* 0 est le plus privilégié, le *ring* 3 le moins privilégié. Dans l'exemple précédent, on pourrait isoler l'ordonnanceur de processus en le faisant s'exécuter en *ring* 0 alors que les autres tâches seraient en *ring* 3.

2.4 Mémoire virtuelle

À partir du moment où plusieurs processus s'exécutent de manière concurrente, un problème d'isolation se pose : si un processus peut lire dans la mémoire d'un autre, des informations peuvent fuiter ; et s'il peut y écrire, il peut en détourner l'exécution.

Le mécanisme de mémoire virtuelle permet de donner à deux tâches une vue différente de la mémoire : c'est-à-dire que vue de tâches différentes, une adresse contiendra une valeur différente.

Ce mécanisme est contrôlé par valeur du registre CR3 : les 10 premiers bits d'une adresse virtuelle sont un index dans le répertoire de pages qui commence à l'adresse contenue dans

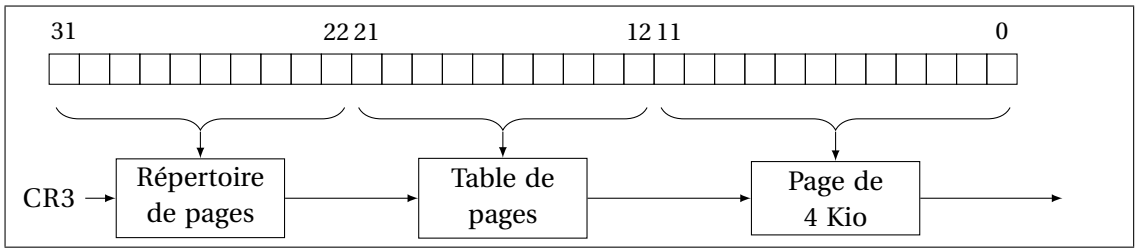


FIGURE 2.2: Implantation de la mémoire virtuelle

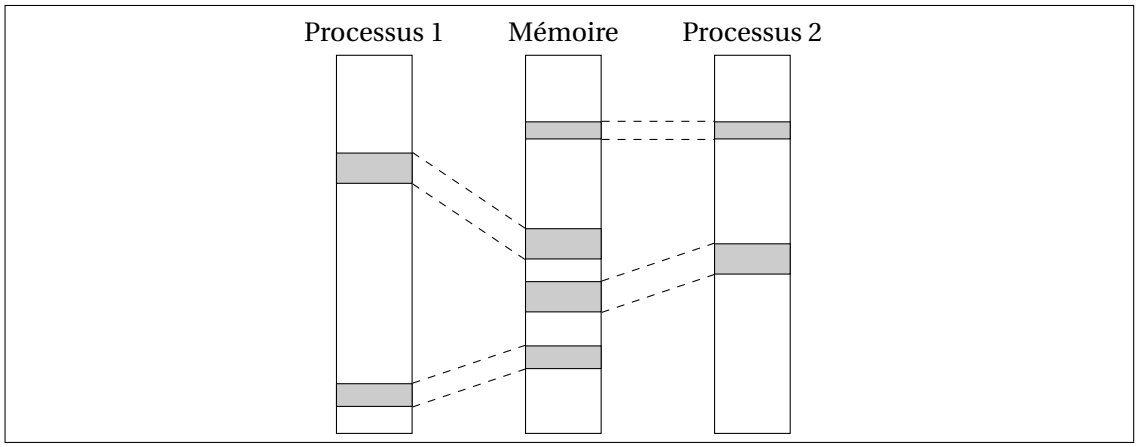


FIGURE 2.3: Mécanisme de mémoire virtuelle.

CR3. À cet index, se trouve l'adresse d'une table de pages. Les 10 bits suivants de l'adresse sont un index dans cette page, donnant l'adresse d'une page de 4 kibioctets (figure 2.2).

En ce qui concerne la mémoire, les différentes tâches ont une vision différente de la mémoire physique : c'est-à-dire que deux tâches lisant à une même adresse peuvent avoir un résultat différent. C'est le concept de mémoire virtuelle (fig 2.3).

2.5 Sécurité des appels système

On a vu que les appels systèmes permettent aux programmes utilisateur d'accéder au services du noyau. Ils forment donc une interface particulièrement sensible aux problèmes de sécurité.

Comme pour toutes les interfaces, on peut être plus ou moins fin. D'un côté, une interface pas assez fine serait trop restrictive et ne permettrait pas d'implémenter tout type de logiciel. De l'autre, une interface trop laxiste ("écrire dans tel registre matériel") empêche toute isolation. Il faut donc trouver la bonne granularité.

Nous allons présenter ici une difficulté liée à la manipulation de mémoire au sein de certains types d'appels système.

Il y a deux grands types d'appels systèmes : d'une part, ceux qui renvoient un simple nombre, comme `getpid` qui renvoie le numéro du processus appelant.

```
pid_t pid = getpid();
printf("%d\n", pid);
```

Ici, pas de difficulté particulière : la communication entre le *ring* 0 et le *ring* 3 est faite uniquement à travers les registres, comme décrit dans la section 8.4.

```c
struct timeval tv;
struct timezone tz;
int z = gettimeofday(&tv, &tz);
if (z == 0) {
    printf( "tv.tv_sec = %ld\ntv.tv_usec = %ld\n"
            "tz.tz_minuteswest = %d\ntz.tz_dsttime = %d\n",
             tv.tv_sec, tv.tv_usec,
             tz.tz_minuteswest, tz.tz_dsttime
          );
}
```

FIGURE 2.4: Appel de gettimeofday

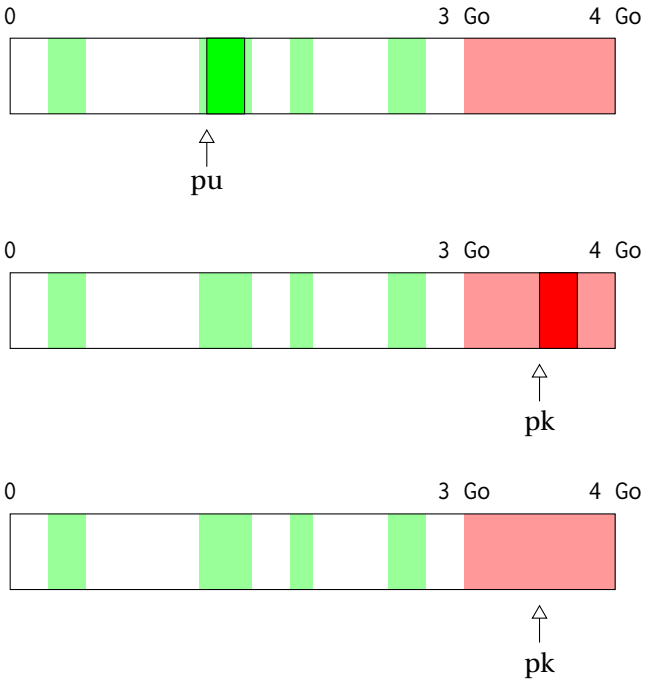


FIGURE 2.5: Zones mémoire

Mais la plupart des appels systèmes communiquent de l'information de manière indirecte, à travers un pointeur. L'appellant alloue une zone mémoire dans son espace d'adressage et passe un pointeur à l'appel système. Ce mécanisme est utilisé par exemple par la fonction `gettimeofday` (figure 2.4).

Considérons une implémentation naïve de cet appel système qui écrirait directement à l'adresse pointée. La figure 2.5(a) présente ce qui se passe lorsque le pointeur fourni est dans l'espace d'adressage du processus : c'est le cas d'utilisation normal et l'écriture est donc possible.

Si l'utilisateur passe un pointeur dont la valeur est supérieure à `0xc0000000` (figure 2.5(b)), que se passe-t-il ? Comme le déréférencement est fait dans le code du noyau, il est également fait en *ring* 0, et va pouvoir être réalisé sans erreur : l'écriture se fait et potentiellement une structure importante du noyau est écrasée.

Un utilisateur malicieux peut donc utiliser cet appel système pour écrire à n'importe quelle adresse dans l'espace d'adressage du noyau. Ce problème vient du fait que l'appel

```
SYSCALL_DEFINE2(gettimeofday, struct timeval __user *, tv,
                struct timezone __user *, tz)
{
        if (likely(tv != NULL)) {
                struct timeval ktv;
                do_gettimeofday(&ktv);
                if (copy_to_user(tv, &ktv, sizeof(ktv)))
                        return -EFAULT;
        }
        if (unlikely(tz != NULL)) {
                if (copy_to_user(tz, &sys_tz, sizeof(sys_tz)))
                        return -EFAULT;
        }
        return 0;
}
```

FIGURE 2.6: Implantation de l'appel système `gettimeofday`

système utilise les privilèges du noyau au lieu de celui qui contrôle la valeur des paramètres sensibles. Celà s'appelle le *Confused Deputy Problem*[Har88].

La bonne solution est de tester dynamiquement la valeur du pointeur : si la valeur du pointeur est supérieure à `0xc0000000`, il faut indiquer une erreur avant d'écrire (figure 2.5(c). Sinon, cela ne veut pas dire que le déréférencement se fera sans erreur, mais au moins le noyau est protégé.

Dans le noyau, un ensemble de fonctions permet d'effectuer des copies sûres. La fonction `access_ok` réalise le test décrit précédemment. Les fonctions `copy_from_user` et `copy_to_user` réalisent une copie de la mémoire après avoir fait ce test. Ainsi, l'implantation correcte de l'appel système `gettimeofday` fait appel à celle-ci (figure 2.6).

Pour préserver la sécurité du noyau, il est donc nécessaire de vérifier la valeur de tous les pointeurs dont la valeur est contrôlée par l'utilisateur. Cette conclusion est assez contraignante, puisqu'il existe de nombreux endroits dans le noyau où des données proviennent de l'utilisateur. Il est donc raisonnable de vouloir vérifier automatiquement et statiquement l'absence de tels défauts.

CHAPITRE 3

ÉTAT DE L'ART

Dans ce chapitre, nous présentons un tour d'horizon des techniques existantes permettant d'analyser des programmes. Un accent est mis sur la propriété de sécurité décrite dans le chapitre 2, mais on ne se limite pas à celle-ci.

L'analyse statique de programmes est un sujet de recherche actif depuis l'apparition de la science informatique. On commence par en présenter une classification, puis on montrera des exemples pertinents permettant d'analyser du code système ou embarqué.

3.1 Taxonomie

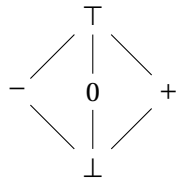
Techniques statiques et dynamiques L'analyse peut être faite au moment de la compilation, ou au moment de l'exécution. En général on peut obtenir des informations plus précises de manière dynamique, mais cela ne prend en compte que les parties du programme qui seront vraiment exécutées. Un autre problème des techniques dynamiques est qu'il est souvent nécessaire d'instrumenter l'environnement d'exécution (ce qui — dans le cas où cela est possible — peut se traduire par un impact en performances). L'approche statique, en revanche, nécessite de construire à l'arrêt une carte mentale du programme, ce qui n'est pas toujours possible dans certains langages.

Pour s'assurer de la correction d'un programme, on ne peut pas s'appuyer uniquement sur des tests — ou de manière générale sur des analyses dynamiques — car même avec des tests exhaustifs, il est impossible d'étudier l'ensemble complet de tous les comportements possibles. Par exemple, si un bug se présente lors d'une interaction entre deux composants qui n'a pas été testée, il passera inaperçu. Pour cette raison, la plupart des analyses présentées ici sont statiques.

Cohérence et complétude Le but d'une analyse statique est de catégoriser les programmes selon s'ils satisfont ou non un ensemble de propriétés fixées à l'avance. Malheureusement, cela n'est que rarement possible car l'ensemble des valeurs possibles lors de l'exécution d'un programme n'est pas un ensemble calculable (théorème de Rice [Ric53]). Autrement dit, il ne peut exister une procédure de décision prenant un programme et le déclarant correct ou incorrect.

Il n'est donc pas possible d'écrire un analyseur statique parfait, ne se trompant jamais. Toute technique statique va donc de se retrouver dans au moins un des cas suivants :

- un programme valide (pour une propriété donnée) est rejeté : on parle de *faux positif*.
- un programme invalide n'est pas détecté : on parle de *faux négatif*.



$$\gamma(-) = \mathbb{R}^-$$
$$\gamma(0) = \{0\}$$
$$\gamma(+) = \mathbb{R}^+$$

FIGURE 3.1: Domaine des signes

En général, et dans notre cas, on préfère s'assurer que les programmes acceptés possèdent la propriété recherchée, quitte à en rejeter certains. C'est l'approche que nous retiendrons.

Par ailleurs la plupart des techniques ne concernent que les programmes qui terminent. On étudie donc la correction, ou les propriétés des termes convergents. Prouver automatiquement que l'exécution ne boucle pas est une propriété toute autre qui n'est pas ici considérée.

3.2 Méthodes syntaxiques

L'analyse la plus simple consiste à traiter un programme comme du texte, et à y rechercher des motifs dangereux. Ainsi, utiliser des outils comme `grep` permet parfois de trouver un grand nombre de vulnérabilités [Spe05].

On peut continuer cette approche en recherchant des motifs mais en étant sensible à la syntaxe et au flot de contrôle du programme. Cette notion de *semantic grep* est présente dans l'outil Coccinelle [BDH+09, PTS+11] : on peut définir des *patches sémantiques* pour détecter ou modifier des constructions particulières.

Ces techniques sont utiles parce qu'elles permettent de plonger rapidement dans le code, en identifiant par exemple des appels à des fonctions dangereuses. En revanche, cela n'est utile que lorsque les propriétés que l'ont recherche sont très locales. Elles offrent également peu de garantie (ce n'est pas parce qu'une construction dangereuse n'est pas présente que le code est correct).

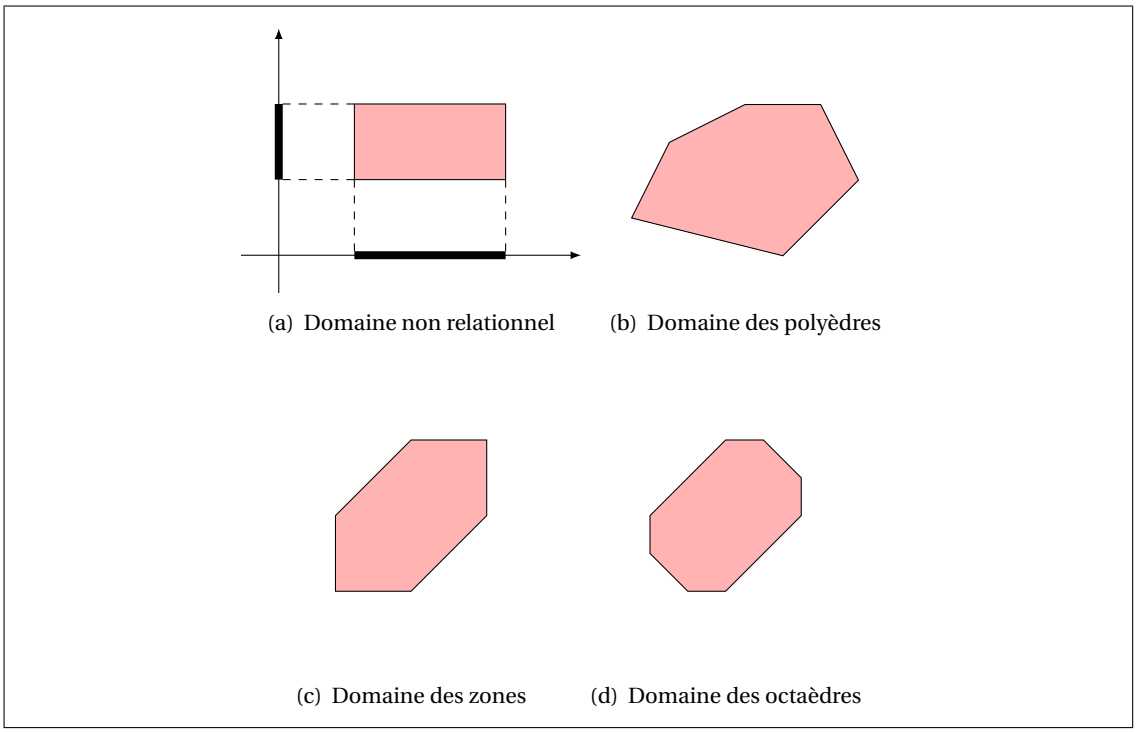
3.3 Interprétation abstraite

L'interprétation abstraite est une technique d'analyse générique qui permet de simuler statiquement tous les comportements d'un programme [CC77, CC92]. Un exemple d'application est de calculer les bornes de variation des variables pour s'assurer qu'aucun débordement de tableau n'est possible [AH07].

L'idée est d'associer à chaque ensemble concret de valeurs, une représentation abstraite. Sur celle-ci, on peut définir des opérations indépendantes de la valeur exacte des données, mais préservant l'abstraction (figure 3.1). Par exemple, les règles comme " $-$ " $\times$ " $-$ " = " $+$ " définissent le domaine abstrait des signes. Les domaines ont une structure de treillis, c'est-à-dire qu'ils possèdent les notions d'ordre partiel et d'union de valeurs. En calculant les valeurs extremales d'une variable, on obtient le domaine des intervalles.

De tels domaines ne capturent aucune relation entre variables. ils sont dits non relationnels. Lorsque plusieurs variables sont analysées en même temps, utiliser de tels domaines consiste à considérer un produit cartésien d'ensembles abstraits (figure 3.2(a)).

Des domaines abstraits plus précis permettent de retenir celles-ci. Pour ce faire, il faut modéliser l'ensemble des valeurs des variables comme un tout. Parmi les domaines rela-



(a) Domaine non relationnel (b) Domaine des polyèdres

(c) Domaine des zones (d) Domaine des octaèdres

FIGURE 3.2: Quelques domaines abstraits

tionnels courants on peut citer : le domaine des polyèdres, permettant de retenir tous les invariants affines entre fonctions (figure 3.2(b)) ; le domaine des zones, permettant de représenter des relations affines de la forme $v_i - v_j \leq c$ (figure 3.2(c)) ; ou encore le domaine des octogones qui est un compromis entre les polyèdres et les zones. Il permet de représenter les relations $\pm v_i \pm v_j \leq c$ (figure 3.2(d)).

En plus des domaines numériques, il est nécessaire d'employer des domaines spécialisés dans la modélisation de la mémoire. Cela est nécessaire pour pouvoir prendre en compte les pointeurs. Par exemple, on peut représenter un pointeur par un ensemble de variables possiblement pointées et une valeur abstraite représentant le décalage (*offset*) du pointeur par rapport au début de la zone mémoire. Cette valeur peut elle-même être abstraite par un domaine numérique.

Au delà des domaines eux-mêmes, l'analyse se fait sous forme d'un calcul de point fixe. La manière la plus simple est d'utiliser un algorithme de *liste de travail*, décrit par exemple dans [SRH95]. Les raffinement en revanche sont nombreux.

Dès [CC77] il est remarqué que la terminaison de l'analyse n'est assurée que si le treillis des valeurs abstraites est de hauteur finie, ou qu'un opérateur d'élargissement (*widening*) ∇ est employé. L'idée est qu'une fois qu'on a calculé quelques termes d'une suite croissante, on peut réaliser une projection de celle ci. Par exemple, dans le domaine des intervalles, $[0;2] \nabla [0;3] = [0;+\infty[$. On atteint alors un point fixe mais qui est plus grand que celui qu'on aurait obtenu sans cette accélération : on perd en précision. Pour en gagner, on peut redescendre sur le treillis des points fixe avec une suite d'itérations décroissantes [Gra92]. Dans l'itération de point fixe, il est possible d'obtenir les résultats de manière plus efficace en choisissant un ordre particulier dans les calculs des sous-itérations[GGTZ07].

En termes d'ingéniérie logicielle, implanter un analyseur statique est un défi en soi. En plus des domaines abstraits, d'un itérateur, il faut traduire le code source à analyser dans un langage, et traduire les résultats de l'analyse en un ensemble d'alarmes à présenter à l'utilisateur.

Cette technique est très puissante : si un interpréteur abstrait *sound* (c'est-à-dire réalisant une surapproximation) analyse un programme et ne renvoie pas d'erreur, alors on a prouvé que le programme est correct (par rapport aux propriétés que vérifient les domaines abstraits). Cela a été appliqué avec succès avec les analyseurs Astrée [Mau04, CCF+05, CCF+09] chez Airbus ou CGS [VB04] à la NASA par exemple.

Cependant, ces analyses sont difficiles à mettre en œuvre. Avec des domaines abstraits classiques comme ceux présentés ci-dessus, les premières analyses peuvent remonter un nombre prohibitif de fausses alarmes. Pour "aider" l'analyse, il faut soit annoter le code soit développer des domaines abstraits *ad hoc* au programme à analyser.

Il existe également des analyseurs statiques par interprétation abstraite qui ne sont pas *sound*, c'est-à-dire qui peuvent manquer des comportements erronés. Leur approche est plus d'aider le programmeur à détecter certains types de bugs pendant le développement. On citer l'exemple de Coverity [BBC+10], qui publie régulièrement des rapports de qualité sur certains logiciels open source. Néanmoins, de part leur aspect non *sound*, les analyses réalisées ne peuvent pas être assimilées à de la vérification formelle en tant que telle.

3.4 Typage

Le typage, introduit dans la section 1.3, peut aussi être utilisé pour la vérification de programmes. On peut le voir comme une manière de catégoriser les types de données manipulés par la machine, mais également à plus au niveau comme une manière d'articuler les différents composants d'un programme. Mais on peut aussi programmer avec les types, c'est-à-dire utiliser le compilateur (dans le cas statique) ou l'environnement d'exécution (dans le cas dynamique) pour vérifier des propriétés écrites par le programmeur.

Systèmes ad hoc Les systèmes de types les plus simples expriment des contrats esssentiellement liés à la sûreté d'exécution, pour ne pas utiliser des valeurs de types incompatibles entre eux. Mais il est possible d'étendre le langage avec des annotations plus riches : par exemple en vérifiant statiquement que des listes ne sont pas vides[KcS07], ou dans le domaine de la sécurité, d'empêcher des fuites d'information [LZ06].

Qualificateurs de types Dans le cas particulier des vulnérabilités liées à une mauvaise utilisation de la mémoire, les développeurs du noyau Linux ont ajouté un système d'annotations au code source. Un pointeur peut être décoré d'une annotation `__kernel` ou `__user` selon s'il est sûr ou pas. Celle-ci sont ignorées par le compilateur, mais un outil d'analyse statique ad-hoc nommé Sparse [5] [TTL] utilisé pour détecter les cas les plus simples d'erreurs. Il demande aussi au programmeur d'ajouter de nombreuses annotations dans le programme.

Ce système d'annotations sur les types a été formalisé sous le nom de *qualificateurs de types* [FJKA06] : chaque type peut être décoré d'un ensemble de qualificateurs (à la manière de `const`), et des règles de typage permettent d'établir des propriétés sur le programme.

Plus précisément, les jugements de typage de la forme $\Gamma \vdash e : t$ sont remplacés par des jugements de typage qualifiés $\Gamma \vdash e : t\ q$. Les qualificateurs q permettent d'exprimer plusieurs jugements. Par exemple, pointeurs on peut étudier le fait qu'une variable soit constante ou pas, que sa valeur soit connue à la compilation, ou encore qu'elle puisse être nulle ou pas. La spécificité de ce système est que les qualificateurs sont ordonnés, du plus spécifique au moins spécifique, et que l'on forme alors un treillis à partir de ces informations. Partant des deux caractéristiques précédentes, on forme le treillis de la figure 3.3. Le qualificateur `const` désigne les données dont la valeur change au cours de l'exécution ; `dynamic` celles qui

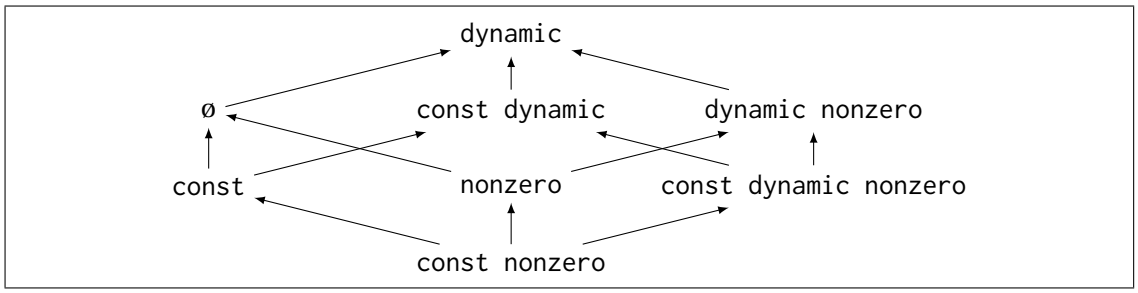


FIGURE 3.3: Treillis de qualificateurs

peuvent être connues à la compilation ; et `nonzero` celles qui ne peuvent jamais être nulles. Le cube sur lequel se trouvent correspond à une relation d'ordre, du plus spécifique (en bas) au plus général (en haut).

Cette relation d'ordre $\preceq$ entre qualificateurs, induit une relation de sous-typage $\sqsubseteq$ entre les types qualifiés : si $q \preceq q'$, alors $t\ q \sqsubseteq t\ q'$.

Ces analyses ont été implantée dans l'outil CQual. Ce système peut servir à inférer les annotations `const` [FFA99], à l'analyse de souillure pour les chaînes de format [STFW01] (pouvant poser des problèmes de sécurité [New00]) et des propriétés dépendantes du flot de contrôle, comme des invariants sur les verrous [FTA02], à rapprocher du concept de *typestates* [SY86]. Il a également été appliqué à la classe de vulnérabilités sur les pointeurs utilisateurs dont il est ici l'objet [JW04].

Ces techniques de typage sont séduisantes parce qu'elles sont en général simples à mettre en place : à l'aide d'un ensemble de règles, on attribue un type à chaque expression. Si le typage se termine sans erreur, alors on est assuré de la correction du programme (par rapport aux propriétés capturées par le système de types).

Le typage statique peut également être implanté de manière efficace. La plupart des systèmes de types peuvent être vérifiés en pratique dans un temps linéaire en la taille du programme considéré.

3.5 Analyse de code système

Les logiciels système demandent des garanties de sécurité et de fiabilité. Pour les assurer, de nombreuses analyses *ad-hoc* ciblent les noyaux de systèmes d'exploitation.

Ajouter un système de types forts à C est l'idée centrale de CCured [NCH+05]. Dans les cas où il n'est pas possible de conclure, des vérifications à l'exécution sont ajoutées. Cependant, cela nécessite une instrumentation dynamique qui se paye en performances.

Saturn [ABD+07] est un système pour analyser du code système écrit en C. Il traite le problème des pointeurs utilisateur en utilisant une analyse de forme "pointe-sur" [BA08]. Comme l'interprétation abstraite, son but est d'être très précis, au détriment d'un temps de calcul important dans certains cas.

3.6 Langages sûrs

Une autre approche est de concevoir un langage à la fois bas niveau et sûr, permettant d'exprimer des programmes proches de la machine tout en interdisant les constructions dangereuses.

Le langage Cyclone [JMG+02] est conçu comme un C "sûr". Afin d'apporter plus de sûreté au modèle mémoire de C, des tests dynamiques sont ajoutés, par exemple aux endroits ou

des conversions implicites peuvent poser problème. Le langage se distingue par le fait qu'il possède plusieurs types de pointeurs : des pointeurs classiques (`int *`), des pointeurs "jamais nuls" (`int @`; un test à l'exécution est alors inséré), et des "pointeurs lourds" (`int ?`; qui contiennent des informations sur la zone mémoire pointée). L'arithmétique des pointeurs n'est autorisée que sur ces derniers, rendant impossibles les débordements de tableaux (ceux-ci étant détectés au pire à l'exécution). Le problème des *dangling pointers* est résolu en utilisant un système de régions [GMJ+02], inspiré des travaux de Jouvelot, Talpin et Tofte [TJ92, TT93, TT94]. Cela permet d'interdire statiquement les constructions où l'on déréférence un pointeur faisant référence à une région de mémoire qui n'est plus allouée.

Le langage Rust [9] développé par Mozilla prend un approche similaire en distinguant plusieurs types de pointeurs pour gérer la mémoire de manière plus fine. Les *managed pointers* (notés `@int`) utilisent un ramasse-miette pour libérer la mémoire allouée lorsqu'ils ne sont plus accessibles. Les *owning pointers* (notés `~int`) décrivent une relation 1 à 1 entre deux objets, comme les `std::unique_ptr` de C++ : la mémoire est libérée lorsque le pointeur l'est. Les *borrowed pointers* (notés `&int`) correspondent aux pointeurs obtenus en prenant l'adresse d'un objet, ou d'un champ d'un objet. Une analyse statique faite lors de la compilation s'assure que la durée de vie de ces pointeurs est plus longue que l'objet pointé, afin d'éviter les *dangling pointers* (pointeurs fous). Cette analyse est également basée sur les régions. Une fonction qui retourne l'adresse d'une variable locale sera donc rejetée par le compilateur. Enfin, le dernier type est celui des *raw pointers* (notés `*int`), pour lesquels le langage n'apporte aucune garantie (il faut d'ailleurs encapsuler chaque utilisation dans un bloc marqué explicitement `unsafe`. Ils sont équivalents aux pointeurs de C.

Ces techniques sont utiles pour créer des nouveaux programmes sûrs, mais on ne peut pas les appliquer pour étudier la correction de logiciels existants. Leurs systèmes de types apportent dans le langage différents types de pointeurs, permettant de manipuler finement la mémoire, à la manière des *smart pointers* de C++. De cette approche on retient surtout l'analyse de régions de Rust qui permet de manipuler de manière sûre les adresses des variables locales, et les pointeurs lourds de Cyclone qui apportent une sûreté à l'arithmétique de pointeurs, au prix d'un test dynamique.

3.7 Logique de Hoare

Une technique pour vérifier statiquement des propriétés sur la sémantique d'un programme a été formalisée par Robert Floyd [Flo67] et Tony Hoare [Hoa69].

Elle consiste à écrire les invariants qui sont maintenus à un point donné du programme. Ces propositions sont écrites dans une logique $\mathscr{L}$. Chaque instruction i est annotée d'une pré-condition P et d'une post-condition Q, ce que l'on note $\{P\}\ i\ \{Q\}$. Cela signifie que si P est vérifiée et que l'exécution de i se termine, alors Q sera vérifiée.

En plus des règles de $\mathscr{L}$, des règles d'inférence traduisent la sémantique du programme ; par exemple la règle de composition est :

$$\frac{\{P\}\ i_1\ \{Q\} \qquad \{Q\}\ i_2\ \{R\}}{\{P\}\ i_1; i_2\ \{R\}}\ (\textsc{Hoare-Seq})$$

Les pré-conditions peuvent être renforcées et les post-conditions relâchées :

$$\frac{\vdash_{\mathscr{L}} P \Rightarrow P' \qquad \{P\}\ i\ \{Q\} \qquad \vdash_{\mathscr{L}} Q' \Rightarrow Q}{\{P'\}\ i\ \{Q'\}}\ (\textsc{Hoare-Consequence})$$

Il est alors possible d'annoter le programme avec ses invariants formalisés de manière explicite dans $\mathscr{L}$. Ceux-ci seront vérifiés à la compilation lorsque c'est possible, sinon à l'exécution.

La règle de conséquence permet de découpler les propriétés du programme lui-même : plusieurs niveaux d'annotations sont possibles, du moins précis au plus précis. En fait, il est même possible d'annoter chaque point de contrôle par l'ensemble d'annotations vide : $\{T\}\ i\ \{T\}$ est toujours vrai.

Augmenter graduellement les pré- et post-conditions est néanmoins assez difficile, puisqu'il peut être nécessaire de modifier l'ensemble des conditions à la fois. Cette difficulté est mentionnée dans [DRS00], où un système de programmation par contrats est utilisé pour vérifier la correction de routines de manipulation de chaînes en C.

Ce type d'annotations a été implanté par exemple pour le langage Java dans le système JML[LBR99] ou pour le langage C# dans Spec#[BLS05].

3.8 Assistants de preuve

Avec un système de types classique, le fait qu'un terme soit bien typé amène quelques propriétés sur son exécution. Par exemple, le fait que seulement un ensemble réduit d'erreurs puisse arriver (comme la division par zéro).

En enrichissant le langage des types, on peut augmenter l'expressivité du typage Par exemple, on peut former des types "entier pair", "vecteur de n entiers", ou encore "liste triée d'entiers".

Habituellement, les termes peuvent dépendre d'autres termes (par composition) ou de types (par des annotations). Les types peuvent également dépendre d'autres types (par composition de types : par exemple, un couple de a et de b a pour type $a * b$). Enrichir l'expressivité du typage revient essentiellement à introduire des termes dans les types, comme n dans l'exemple précédent du vecteurs de n entiers. C'est pourquoi on parle de types dépendants. Parmi les langages proposant ces types on peut citer Coq [The04], Agda [BDN09] ou Isabelle [NPW02].

Dans un langage classique, la plupart des types sont habitables, c'est à dire qu'il existe des termes ayant ces types. En revanche, avec les types dépendants ce n'est pas toujours vrai : par exemple "vecteur de -1 entiers" n'a pas d'habitants. Ainsi, pouvoir construire un terme d'un type donné est une information en soi.

On peut voir ce phénomène sous un autre angle : les termes sont à leur type ce que les preuves sont à leur théorème. C'est la correspondance de Curry-Howard. Prenons un exemple : si on dispose d'un terme de type a, et d'une fonction de type $a \to b$, alors on peut en déduire (par simple application) un terme de type b. Vu sous forme logique : si on a une preuve de A, et une preuve de $A \Rightarrow B$, alors on peut en tirer une preuve de B. À l'aide de cette correspondance, on peut donc voir les systèmes de types dépendants comme des assistants de preuves généralistes où l'on peut faire vérifier par un compilateur des théorèmes mathématiques comme le théorème des 4 couleurs [Gon05].

Cette technique fait se rejoindre les mondes du typage et des annotations par triplets de Hoare : on enrichit le langage des types suffisamment pour encoder une logique, et on vérifie suffisamment les triplets statiquement pour éliminer tous les tests dynamiques. Cependant cette technique est très complexe à mettre en œuvre, puisqu'il faut encoder toutes les propriétés voulues dans un formalisme de très bas niveau (similaire à la théorie des ensembles).

3.9 Conclusion

Il existe de nombreuses techniques pour vérifier du code système ou embarqué. Il y a divers choix à faire entre l'expressivité, l'intégration de tests dynamiques, ou la facilité de mise en œuvre.

Le typage statique est une solution assez pragmatique, puisqu'elle peut s'appliquer à des programmes existants, et est efficace à utiliser. Son expressivité limitée nous empêche de reposer entièrement sur elle pour garantir l'absence d'erreur dans les programmes système. C'est pourquoi nous approchons la sûreté de la manière suivante dans notre solution, SAFESPEAK.

Tout d'abord, on utilise le typage pour manipuler les données de manière compatible : les types des opérations et fonctions sont vérifiés à la compilation.

Ensuite, les accès aux tableaux et aux pointeurs sont vérifiés dynamiquement. Dans le cas où une erreur est déclenchée, l'exécution s'arrête plutôt que de corrompre la mémoire. La pile est également nettoyée à chaque retour de fonction afin d'éviter les pointeurs fous (*dangling pointers*).

Enfin, les pointeurs provenant de l'espace utilisateur sont repérés statiquement afin que leur déréférencement se fasse au sein de fonctions sûres. Cela permet de préserver l'isolation entre le noyau et l'espace utilisateur.

CONCLUSION DE LA PARTIE I

Cette thèse part de deux constats.

D'une part, le problème de manipulation sûre de pointeurs utilisateur est éliminé si on interdit certaines opérations sur ces pointeurs, en en faisant un type à part. En effet, la seule opération dangereuse est le déréférencement. Si on interdit les déréférencements syntaxiques (opérateur `*`) et qu'on restreint les cas nécessaires à des fonctions sûres comme `copy_to_user`, on élimine les comportements dangereux. Bien sûr, un examen manuel exhaustif de tous les déréférencements est possible, mais trop long et source d'erreurs.

D'autre part, le système de types de C est trop primitif pour pouvoir garantir une véritable isolation entre deux types de même représentation : il n'y a pas de types abstraits. Certes, `typedef` permet d'introduire un nouveau nom pour un type, mais ce n'est qu'un raccourci syntaxique. Le compilateur ne peut en effet pas considérer un programme sans avoir la définition quasi-complète des types qui y apparaissent. La seule exception concerne les manipulations de structures par pointeurs : si on ne fait que passer des pointeurs, il n'est pas nécessaire de connaître la taille ni la disposition de la structure, donc il est possible de ne pas connaître ces informations. Cette technique, connue sous le nom de *pointeurs opaques*, n'est pas applicable aux autres types.

Le but de cette thèse est donc de définir un langage intermédiaire proche de C, mais bien typable, et de l'adjoindre d'un système de types tel que les programmes bien typés manipulent les pointeurs utilisateur sans causer de problèmes de sécurité.

Deuxième partie

Un langage pour l'analyse de code système : SAFESPEAK

Dans cette partie, nous allons présenter un langage impératif modélisant le langage C. Le chapitre 4 décrit sa syntaxe, ainsi que sa sémantique. À ce point, de nombreux programmes sont acceptés mais qui provoquent des erreurs à l'exécution.

Afin de rejeter ces programmes incorrects, on définit ensuite dans le chapitre 5 une sémantique statique s'appuyant sur un système de types simples. Des propriétés de sûreté de typage sont ensuite établies, permettant de catégoriser l'ensemble des erreurs à l'exécution possibles.

Le chapitre 6 commence par étendre notre langage avec une nouvelle classe d'erreurs à l'exécution, modélisant les accès à la mémoire utilisateur catégorisé comme dangereux dans le chapitre 2. Une extension au système de types du chapitre 5 est ensuite établie, et on prouve que les programmes ainsi typés ne peuvent pas atteindre ces cas d'erreur.

Trois types d'erreurs à l'exécution sont possibles :

- les erreurs de typage (dynamique), lorsqu'on tente d'appliquer à une opération des valeurs incompatibles (additionner un entier et une fonction par exemple.
- les erreurs de sécurité, qui consistent en le déréférencement d'un pointeur dont la valeur est contrôlée par l'espace utilisateur. Celles-ci sont uniquement possibles en contexte noyau.
- les erreurs mémoire, qui résultent d'un débordement de tableau, du déréférencement d'un pointeur invalide ou d'arithmétique de pointeur invalide.

En résumé, l'introduction des types simples enlève la possibilité de rencontrer des erreurs de typage dynamique, et l'ajout des qualificateurs interdit les erreurs de sécurité.

Langage	Types	Erreurs possibles		
		Typage	Sécurité	Mémoire
SAFESPEAK	sans	☑	N/A	☑
SAFESPEAK	simples	☐	N/A	☑
SAFESPEAK noyau	simples	☐	☑	☑
SAFESPEAK noyau	qualifiés	☐	☐	☑

CHAPITRE 4

SYNTAXE ET SÉMANTIQUE

Dans ce chapitre nous présentons SAFESPEAK, un langage impératif inspiré de C. Sa syntaxe est tout d'abord décrite ; puis une sémantique opérationnelle est explicitée.

Ce langage servira de support au systèmes de types décrit dans le chapitre 5 et enrichi dans le chapitre 6.

La traduction depuis C sera explicitée dans le chapitre 7.

4.1 Notations

Ensembles inductifs

Dans ce chapitre (et les chapitres suivants), on définit de nombreux ensembles inductifs. Plutôt que d'écrire la construction explicite par point fixe, on emploie une notation en grammaire.

Étudions l'exemple des listes chaînées composées d'élements de $\mathbb{N}$.

Notons L cet ensemble ; si $[]$ est la liste vide et $n :: l$ la liste formée d'une "tête" $n \in \mathbb{N}$ et d'une "queue" $l \in L$. Toute liste est donc d'une des formes suivantes :

- $[]$
- $n_1 :: []$
- $n_1 :: n_2 :: []$
- etc.

On peut donc L de la manière inductive suivante :

$$\mathrm{L} = \mathrm{fix}(L')$$
$$L'(E) = \{[]\} \cup \{n :: l / n \in \mathbb{N}, l \in E\}$$

où

$$\mathrm{fix}(f) = \bigcup_{n=0}^{\infty} f^n(\emptyset)$$
$$f^0(x) = x$$
$$\forall n > 0, f^n(x) = f^{n-1}(f(x))$$

(L'itération n de l'union correspond aux listes comprenant au plus n éléments)

Plutôt que d'écrire cette définition précise mais chargée, on écrira à la place une définition en compréhension :

Listes	$l ::= []$	Liste vide
	$\mid \; n :: l$	Construction de liste

Chaque ensemble est identifié de manière unique par les noms de variables métasyntaxiques : n pour les entiers et l pour les listes ici. Si plusieurs métavariables du même ensemble doivent apparaître, elles sont indicées. Par exemple, on peut définir des arbres binaires d'entiers de la manière suivante :

Arbres binaires	$a ::= \mathrm{F}$	Feuille
	$\mid \; \mathrm{N}(a_1, n, a_2)$	Nœud

Cette notation a aussi l'avantage de s'étendre facilement aux définitions mutuellement récursives.

Inférence

La sémantique opérationnelle consiste en la définition d'une relation de transition $\cdot \to \cdot$ entre états de l'interpréteur [1].

Cette relation est définie inductivement sur la syntaxe du programme. Plutôt que de présenter l'induction explicitement, elle est représentée par des jugements logiques et des règles d'inférences, de la forme :

$$\frac{P_1 \quad \dots \quad P_n}{C}\ (\textsc{Nom})$$

Les P_i sont les prémisses, et C la conclusion. Cette règle s'interprète de la manière suivante : si les P_i sont prouvées, alors C est prouvée.

Certaines règles n'ont pas de prémisse, ce sont des axiomes :

$$\frac{}{A}\ (\textsc{Ax})$$

Compte-tenu de la structure des règles, la preuve d'un jugement pourra donc être vue sous la forme d'un arbre :

$$\dfrac{\dfrac{\dfrac{}{A_1}\ (\textsc{r3}) \quad \dfrac{}{A_2}\ (\textsc{r4})}{B_1}\ (\textsc{r2}) \quad \dfrac{\dfrac{}{A_3}\ (\textsc{r6})}{B_2}\ (\textsc{r5})}{C}\ (\textsc{r1})$$

1. Dans le chapitre 5, la relation de typage $\cdot \vdash \cdot : \cdot$ sera définie par la même technique.

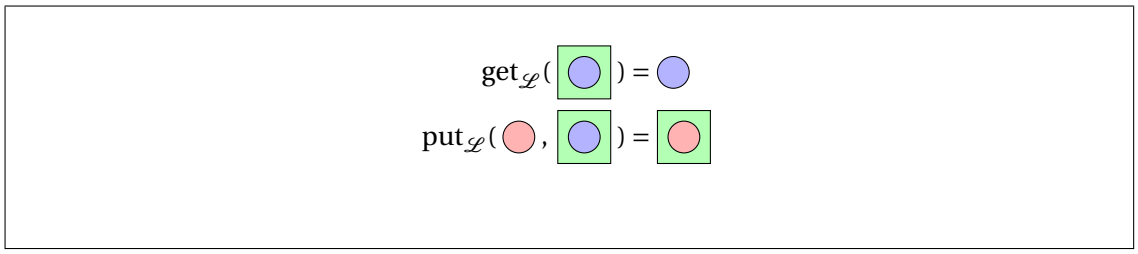


FIGURE 4.1: Fonctionnement d'une lentille

Lentilles

La notion d'accesseur utilisée ici est directement inspirée des *lentilles* utilisées en programmation fonctionnelle, décrite dans [FGM+07] et [vL11].

Définition 4.1 (Lentille)**.** *Étant donnés deux ensembles R et A, une* lentille $\mathscr{L} \in \mathrm{Lens}_{R,A}$ *(ou* accesseur*) est un moyen d'accéder en lecture ou en écriture à sous-valeur de type A au sein d'une valeur de type R (pour* record*). Elle est consistuée des opérations suivantes :*

- *une fonction de lecture* $\mathrm{get}_{\mathscr{L}} : R \to A$
- *une fonction de mise à jour* $\mathrm{put}_{\mathscr{L}} : (A \times R) \to R$

telles que pour tous $a \in A, a' \in A, r \in R$:

$$\mathrm{put}_{\mathscr{L}}(\mathrm{get}_{\mathscr{L}}(r), r) = r \qquad \text{(GetPut)}$$
$$\mathrm{get}_{\mathscr{L}}(\mathrm{put}_{\mathscr{L}}(a, r)) = a \qquad \text{(PutGet)}$$
$$\mathrm{put}_{\mathscr{L}}(a', \mathrm{put}_{\mathscr{L}}(a, r)) = \mathrm{put}_{\mathscr{L}}(a', r) \qquad \text{(PutPut)}$$

On note $\mathscr{L} = \langle \mathrm{get}_{\mathscr{L}} | \mathrm{put}_{\mathscr{L}} \rangle$.

GetPut *signifie que si on lit une valeur puis qu'on la réecrit, l'objet n'est pas modifié ;* PutGet *décrit l'opération inverse : si on écrit une valeur dans le champ, c'est la valeur qui sera lue ; enfin,* PutPut *évoque le fait que chaque écriture est totale : quand deux écritures se suivent, seule la seconde compte.*

Une illustration se trouve dans la figure 4.1.

Exemple 4.1 (Lentilles de tête et de queue de liste)**.** *Soit E un ensemble. On considère $L(E)$, l'ensemble des listes d'éléments de E.*

On définit les fonctions suivantes. Notons qu'elles ne sont pas définies sur la liste vide []*, qui pourra être traité comme un cas d'erreur.*

$$\mathrm{get}_T(t :: q) = t$$
$$\mathrm{put}_T(t', t :: q) = t' :: q$$
$$\mathrm{get}_Q(t :: q) = q$$
$$\mathrm{put}_Q(q', t :: q) = t :: q'$$

Alors $T = \langle \mathrm{get}_T | \mathrm{put}_T \rangle \in \mathrm{Lens}_{L(E),E}$ *et* $Q = \langle \mathrm{get}_Q | \mathrm{put}_Q \rangle \in \mathrm{Lens}_{L(E),L(E)}$.

On a par exemple :

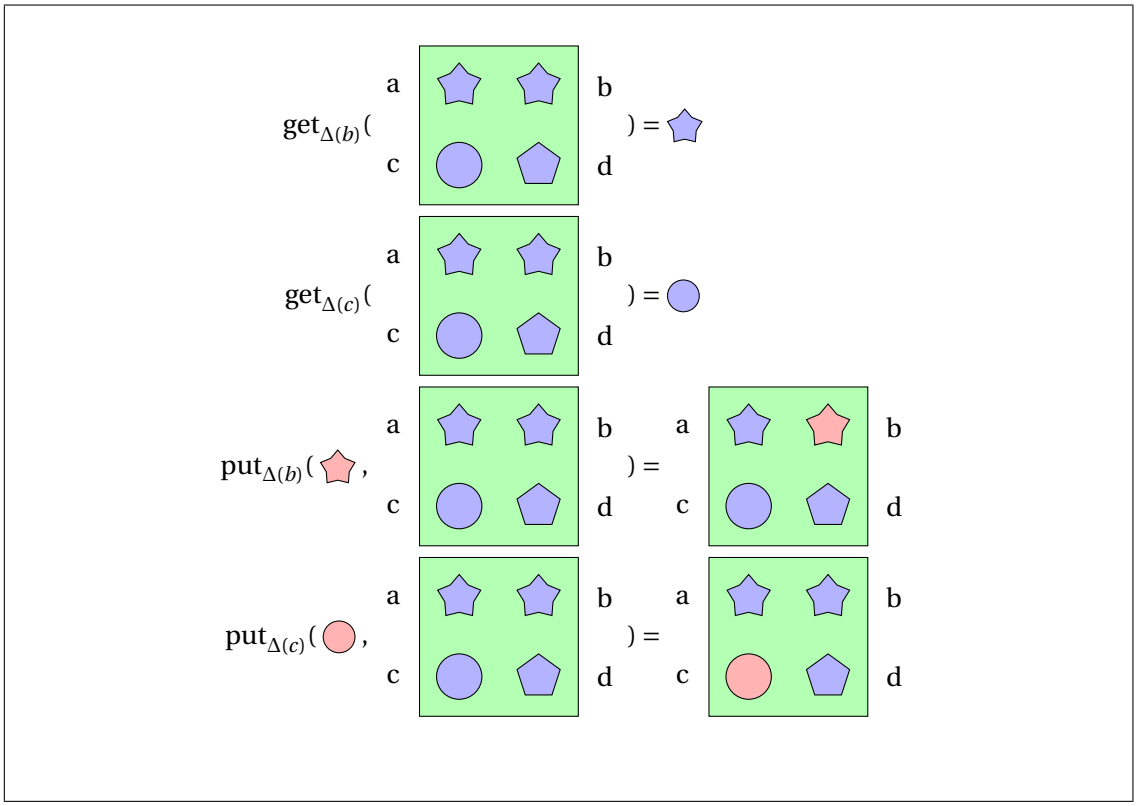


FIGURE 4.2: Fonctionnement d'une lentille indexée

$\text{get}_T(1::6::1::8::[]) = 1$
et :
$\text{put}_Q(7,3::6::1::5::[]) = 7::6::1::5::[]$.

Définition 4.2 (Lentille indexée)**.** *Les objets de certains ensembles R sont composés de plusieurs sous-objets accessibles à travers un indice $i \in I$. Une lentille indexée est une fonction Δ qui associe à un indice i une lentille entre R et un de ses champs A_i :*

$$\forall i \in I, \exists A_i, \Delta(i) \in \text{Lens}_{R,A_i}$$

On note alors :

$$r[i]_\Delta \stackrel{\text{def}}{=} \text{get}_{\Delta(i)}(r)$$
$$r[i \leftarrow a]_\Delta \stackrel{\text{def}}{=} \text{put}_{\Delta(i)}(a,r)$$

Un exemple est illustré dans la figure 4.2.

Exemple 4.2 (Lentille "n^e élément d'un tuple")**.** *Soient $n \in \mathbb{N}$, et n ensembles $E_1,\ldots,E_n$. Pour tout $i \in [1;n]$, on définit :*

$$g_i((x_1,\ldots,x_n)) = x_i$$
$$p_i(y,(x_1,\ldots,x_n)) = (x_1,\ldots,x_{i-1},y,x_{i+1},\ldots,x_n)$$

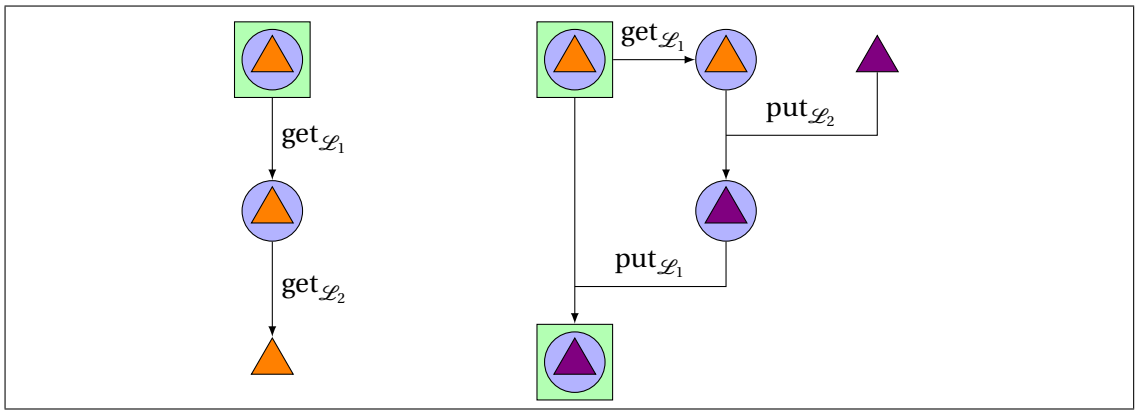


FIGURE 4.3: Composition de lentilles

Définissons $T(i) = \langle g_i | p_i \rangle$. *Alors* $T(i) \in \text{Lens}_{(E_1 \times \ldots \times E_n), E_i}$.
Donc T est une lentille indexée, et on a par exemple :

$$\begin{aligned}(3,1,4,1,5)[2]_T &= \text{get}_{T(2)}((3,1,4,1,5)) \\ &= 1\end{aligned}$$

$$\begin{aligned}(9,2,6,5,3)[3 \leftarrow 1]_T &= \text{put}_{T(3)}(1,(9,2,6,5,3)) \\ &= (9,2,1,5,3)\end{aligned}$$

Définition 4.3 (Composition de lentilles)**.** *Soient* $\mathscr{L}_1 \in \text{Lens}_{A,B}$ *et* $\mathscr{L}_2 \in \text{Lens}_{B,C}$.
La composition de $\mathscr{L}_1$ *et* $\mathscr{L}_2$ *est la lentille* $\mathscr{L} \in \text{Lens}_{A,C}$ *définie de la manière suivante :*

$$\begin{aligned}\text{get}_{\mathscr{L}}(r) &= \text{get}_{\mathscr{L}_2}(\text{get}_{\mathscr{L}_1} r) \\ \text{put}_{\mathscr{L}}(a,r) &= \text{put}_{\mathscr{L}_1}(\text{put}_{\mathscr{L}_2}(a, \text{get}_{\mathscr{L}_1} r), r)\end{aligned}$$

On notera alors $\mathscr{L} = \mathscr{L}_1 \ggg \mathscr{L}_2$.

Cette définition est illustrée dans la figure 4.3. Une preuve que la composition est une lentille est donnée en annexe D.1.

4.2 Fonctionnalités

Fonctions et procédures : Un des problèmes classiques dans les langages impératifs est de distinguer les fonctions (qui retournent une valeur) et les procédures (qui n'en retournent pas). La solution choisie par C est de marquer les procédures comme retournant un "faux" type `void`. Mais c'est uniquement syntaxique : il n'est pas possible de manipuler cette valeur de retour de type `void`.

L'autre possibilité, souvent prise dans les langages fonctionnels, est de ne pas faire de distinction entre ces deux cas et d'interdire les procédures. Les fonctions ne retournant pas de valeur "intéressante" renvoient alors une valeur d'un type à un seul élément appelé $()$ [2], et donc le type sera noté UNIT.

2. Cette notation évoque un n-uplet à 0 composante.

En C, puisqu'il n'y a pas de valeurs de type `void`, la notation `void *` a un sens particulier : elle désigne les pointeurs de type non défini, qui sont compatibles avec n'importe quel autre type de pointeur (c'est la seule forme — rudimentaire — de généricité qu'offre le langage). Ici, la valeur () est une valeur comme les autres, et on peut constuire un pointeur de type UNIT* qui n'aura pas de signification particulière : c'est seulement un pointeur vers une valeur de type UNIT.

Tableaux : ce sont des valeurs composées qui contiennent un certain nombre de valeurs d'un même type. Par exemple, 100 entiers. On accède à ces valeurs par un indice entier, qui dans le cas général n'est pas connu à la compilation. C'est une erreur (Ω_{array}) d'accéder à un tableau en dehors de ses bornes, c'est-à-dire en dehors de $[0; n-1]$ pour un tableau à n éléments.

Les tableaux sont notés $[e_1; \ldots; e_n]$, et le cas dégénéré ($n = 0$) est interdit.

Structures : comme les tableaux, ce sont des valeurs composées mais hétérogènes. Les différents éléments (appelés *champs*) sont désignés par noms l (pour *label*) et de manière statique (il n'y a pas de mécanisme pour faire référence à un nom dans le programme).

Les structures sont notées $\{l_1 : e_1; \ldots; l_n : e_n\}$ et comme dans le cas des tableaux, le cas dégénéré ($n = 0$) est interdit.

Dans le programme, le nom de champ l est complété de la définition complète de la structure S. Celle-ci n'est pas utilisée dans l'évaluation et sera donc décrite au chapitre 5. Bien sûr, écrire la totalité de la structure dans le code du programme serait fastidieux. Nous décrirons dans le chapitre 7 comment obtenir automatiquement ces annotations depuis un langage comme C qui utilise des noms de structures.

4.3 Principes

Nous voulons capturer l'essence de C. Les traits principaux sont les suivants :

Types de données : très simples. Entiers machine, flottants, pointeurs et types composés (structures et tableaux) composés de ceux-ci.

Variables : elles sont mutables, et on peut passer des données par valeur ou par pointeur.

Flôt de contrôle : il repose sur les construction "if" et "while". Les autres types de boucle ("for" et "do/while") peuvent être construits avec ces opérateurs.

Fonctions : le code est organisé en fonctions "simples", c'est-à-dire qui ne sont pas des fermetures. Même si le corps d'une fonction peut être inclus dans le corps d'une autre, il n'est pas possible d'accéder aux variables de la portée entourante depuis la fonction intérieure.

4.4 Syntaxe

Les figures 4.4, 4.5 et 4.6 présentent notre langage intermédiaire. Il contient la plupart des fonctionnalités présentes dans les langages impératifs comme C.

Un programme est organisé en fonctions, qui contiennent des instructions, qui elles-mêmes manipulent des expressions.

Constantes	$c ::= i$	Entier
	$\mid d$	Flottant
	$\mid$ NULL	Pointeur nul
	$\mid ()$	Valeur unité
Expressions	$e ::= c$	Constante
	$\mid lv$	Accès mémoire
	$\mid \boxminus\ e$	Opération unaire
	$\mid e \boxplus e$	Opération binaire
	$\mid \&lv$	Pointeur
	$\mid lv \leftarrow e$	Affectation
	$\mid \{l_1 : e_1; \ldots; l_n : e_n\}$	Structure
	$\mid [e_1; \ldots; e_n]$	Tableau
	$\mid f$	Fonction
	$\mid e(e_1, \ldots, e_n)$	Appel de fonction
Left-values	$lv ::= x$	Variable
	$\mid *lv$	Déréférencement
	$\mid lv.l_S$	Accès à un champ
	$\mid lv[e]$	Accès à un élément
Fonctions	$f ::= \text{fun}(x_1, \ldots, x_n)\{i\}$	Arguments, corps

FIGURE 4.4: Syntaxe – expressions

Le flot de contrôle est simplifié par rapport à C : il ne contient que l'alternative ("if") et la boucle "while". Les autres formes de boucle ("do/while" et "for") peuvent être émulées par une boucle "while".

Les fonctionnalités manquantes, et comment les émuler, seront discutés dans le chapitre 9.

Pour l'alternative, on introduit également la forme courte IF$(e)\{i\}$ = IF$(e)\{i\}$ELSE{PASS}.

Les opérateurs sont donnés dans la figure 4.6.

4.5 Définitions préliminaires

On suppose avoir à notre disposition un ensemble infini dénombrable d'identificateurs ID (par exemple des chaînes de caractères).

X^* est l'ensemble des suites finies de X, indexées à partir de 1. Si $u \in X^*$, on note $|u|$ le nombre d'éléments de u (le cardinal de son ensemble de définition). Pour $1 \leq i \leq |u|$, on note $u_i = u(i)$ le i-ème élément de la suite.

On peut aussi voir les suites comme des listes : on note [] la suite vide, telle que $|[]| = 0$. On définit en outre la construction de suite de la manière suivante : si $x \in X$ et $u \in X^*$, la liste

Instructions	$i ::=$ PASS	Instruction vide
	$\mid\ i; i$	Séquence
	$\mid\ e$	Expression
	$\mid$ DECL $x = e$ IN$\{i\}$	Déclaration de variable
	$\mid$ IF$(e)\{i\}$ELSE$\{i\}$	Alternative
	$\mid$ WHILE$(e)\{i\}$	Boucle
	$\mid$ RETURN(e)	Retour de fonction
Phrases	$p ::= x = e$	Variable globale
	$\mid\ e$	Évaluation d'expression
Programme	$P ::= (p_1, \ldots, p_n)$	Phrases

FIGURE 4.5: Syntaxe – instructions

Opérateurs binaires	$\boxplus ::= +, -, \times, /, \%$	Arithmétique entière
	$\mid\ +., -., \times., /.$	Arithmétique flottante
	$\mid\ +_p, -_p$	Arithmétique de pointeurs
	$\mid\ =, \neq, \leq, \geq, <, >$	Comparaisons
	$\mid$ &, |, ^	Opérateurs bit à bit
	$\mid$ &&, ||	Opérateurs logiques
	$\mid\ \ll, \gg$	Décalages
Opérateurs unaires	$\boxminus ::= +, -$	Arithmétique entière
	$\mid\ +., -.$	Arithmétique flottante
	$\mid\ \sim$	Négation bit à bit
	$\mid$!	Négation logique

FIGURE 4.6: Syntaxe – opérateurs

$x :: u \in X^*$ est la liste v telle que :

$$\begin{aligned} v_1 &= x \\ \forall i \in [1;|u|], v_{i+1} &= u_i \end{aligned}$$

La concaténation des listes u et v est la liste $u@v = w$ telle que :

$$\begin{aligned} |w| &= |u| + |v| \\ \forall i \in [1;|u|], w_i &= u_i \\ \forall j \in [1;|v|], w_{|u|+j} &= v_j \end{aligned}$$

4.6 Mémoire

L'interprète que nous nous apprêtons à définir manipule des valeurs qui sont associées aux variables du programme.

La mémoire est constituée de variables, qui contiennent des valeurs. Ces variables sont organisées, d'une part en un ensemble de variables globales, et d'autre part en une pile de contextes d'appel (qu'on appellera donc aussi cadres de pile, ou *stack frames* en anglais). Cette structure empilée permet de représenter les différents contextes à chaque appel de fonction : par exemple, si une fonction s'appelle récursivement, plusieurs instances de ses variables locales sont présentes dans le programme.

La structure de pile des locales permet de les organiser en niveaux indépendants : à chaque appel de fonction, un nouveau cadre de pile est créé, comprenant ses paramètres et ses variables locales. Au contraire, pour les globales il n'y a pas de système d'empilement, puisque ces variables sont accessibles depuis tout point du programme.

Pour identifier de manière non ambigüe une variable, on note simplement x la variable globale nommée x, et (n, x) la variable locale nommée x dans le n^{e} cadre de pile [3].

Les affectations peuvent avoir la forme $x \leftarrow e$ où x est une variable et e est une expression, mais pas seulement. En effet, à gauche de $\leftarrow$ on trouve en général non pas une variable mais une left-value. Pour représenter quelle partie de la mémoire doit être accédée par cette left-value, on introduit la notion de chemin φ. Un chemin est en quelque sorte une left-value symbolique évaluée : les cas sont similaires, sauf que tous les indices sont évalués. Par exemple, $\varphi = (5, x).p$ représente le champ "p" de la variable x dans le 5^{e} cadre de pile.

Les valeurs, quant à elles, peuvent avoir les formes suivantes (résumé sur la figure 4.7) :

- $\widehat{c}$: une constante. La notation circonflexe permet de distinguer les constructions syntaxique des constructions sémantiques. Par exemple, à la syntaxe 3 correspond la valeur $\widehat{3}$.

  Les valeurs entières sont les entiers signés sur 32 bits, c'est-à-dire entre -2^{31} à $2^{31} - 1$. Mais ce choix est arbitraire : de la même manière, on aurait pu choisir des nombres à 64 bits ou même de précision arbitraire. Les flottants sont les flottants IEEE 754 de 32 bits [oEE08].
- φ : une référence mémoire. Ce chemin correspond à un pointeur sur une left-value. Par exemple, l'expression $\&x$ s'évalue en $\varphi = (5, x)$ si x désigne lexicalement une variable dans le 5^{e} cadre de pile.

3. Les paramètres de fonction sont traités comme des variables locales et se retrouvent dans le cadre correspondant.

Valeurs	$v ::= \widehat{c}$	Constante
	$\mid \widehat{\&}\,\varphi$	Référence mémoire
	$\mid \{l_1 : \widehat{v_1;\ldots;l_n} : v_n\}$	Structure
	$\mid [\widehat{v_1;\ldots;v_n}]$	Tableau
	$\mid \widehat{f}$	Fonction
	$\mid \Omega$	Erreur
Chemins	$\varphi ::= a$	Adresse
	$\mid *\varphi$	Déréférencement
	$\mid \varphi.l$	Accès à un champ
	$\mid \varphi[n]$	Accès à un élément
Adresses	$a ::= (n, x)$	Variable locale
	$\mid (x)$	Variable globale
Erreur	$\Omega ::= \Omega_{array}$	Débordement de tableau
	$\mid \Omega_{ptr}$	Erreur de pointeur
	$\mid \Omega_{div}$	Division par zéro
	$\mid \Omega_{field}$	Erreur de champ

FIGURE 4.7: Valeurs

- $\{l_1 : \widehat{v_1;\ldots;l_n} : v_n\}$: une structure. Comme précédemment, on note $\widehat{\{\cdot\}}$ pour dénoter les valeurs.
- $[\widehat{v_1;\ldots;v_n}]$: un tableau. Pareillement, $\widehat{[\cdot]}$ permet de désigner les valeurs. Par exemple, si x vaut 2 et y vaut 3, l'expression $[x, y]$ s'évaluera en valeur $\widehat{[2,3]}$
- $\widehat{f}$: une fonction. Les valeurs fonctions comportent l'intégralité de la définition de la fonction (liste de paramètres, de variables locales et corps). Remarquons que contrairement à certains langages, l'environnement n'est pas capturé (il n'y a pas de clôture lexicale).
- Ω : une erreur. Par exemple le résultat de $5/0$ est Ω_{div}.

La figure 4.8 résume comment ces valeurs sont organisées. Une pile est une liste de cadre de piles, et un cadre de pile est une liste de couples (nom, valeur). Un état mémoire m est un couple (s, g) où s est une pile et g un cadre de pile (qui représente les variables globales).

Enfin, l'interprétation est définie comme une relation $\cdot \to \cdot$ entre états Ξ ; ces états sont d'une des formes suivantes :

- un couple $\langle e, m\rangle$ où e est une expression et m un état mémoire. m est l'état mémoire sous lequel l'évaluation sera réalisée. Par exemple $\langle 3, ([], [x, 3])\rangle \to \langle \widehat{3}, ([], [x, 3])\rangle$ L'évaluation des expressions est détaillée dans la section 4.11.
- un couple $\langle i, m\rangle$ où i est une instruction et m un état mémoire. La réduction instructions est traitée dans la section 4.12.

Pile	$s ::= []$	Pile vide
	$\mid \{x_1 \mapsto v_1; \ldots; x_n \mapsto v_n\} :: s$	Ajout d'un cadre
État mémoire	$m ::= (s, \{x_1 \mapsto v_1; \ldots; x_n \mapsto v_n\})$	Pile, globales
État d'interpréteur	$\Xi ::= \langle e, m \rangle$	Expression, mémoire
	$\mid \langle i, m \rangle$	Instruction, mémoire
	$\mid \Omega$	Erreur

FIGURE 4.8: Composantes d'un état mémoire

- une erreur Ω. La propagation des erreurs est détaillée dans la section 4.13.

4.7 Opérations sur les valeurs

Un certain nombre d'opérations est possibles sur les valeurs (figure 4.6) :

- les opérations arithmétiques $+$, $-$, $\times$, $/$ et % sur les entiers. L'opérateur % correspond au modulo (reste de la division euclidienne). En cas de division par zéro, l'erreur Ω_{div} est levée.
- les versions "pointées" $+.$, $-.$, $\times.$ et $/.$ sur les flottants
- les opérations d'arithmétique de pointeur $+_p$ et $-_p$ qui à un chemin mémoire et un entier associent un chemin mémoire.
- les opérations d'égalité $=$ et $\neq$. L'égalité entre entiers ou entre flottants est immédiate. Deux valeurs composées (tableaux ou structures) sont égales si elles ont la meme "forme" (même taille pour les tableaux, et même champs pour les structures) et que toutes leurs sous-valeurs sont égales deux à deux. Les références mémoire φ sont égales lorsque les chemins qu'ils décrivent sont syntaxiquement égaux[4].
- les opérations de comparaison $\leq, \geq, <, >$ sont définies avec leur sémantique habituelle sur les entiers et les flottants. Sur les références mémoires, elles sont définies dans le cas où les deux opérarandes sont de la forme $\varphi[\cdot]$ par : $\varphi[n] \boxplus \varphi[m] \stackrel{\text{def}}{=\!=} n \boxplus m$. Dans les autres cas, l'erreur Ω_{ptr} est renvoyée.
- les opérateurs bit à bit sont définis sur les entiers. &, | et ^ représentent respectivement la conjonction, la disjonction et la disjonction exclusive (XOR).
- des versions logiques de la conjonction (&&) et de la disjonction (||) sont également présentes. Leur sémantique est donnée par le tableau suivant :

n	m	n && m	n || m
0	0	0	0
0	$\neq 0$	0	1
$\neq 0$	0	0	1
$\neq 0$	$\neq 0$	1	1

4. Il est donc possible que deux pointeurs pointent sur la même adresse mais soient considérés différents. La raison pour ce choix est que la comparaison doit pouvoir se faire sans accéder à la mémoire.

- des opérateurs de décalage à gauche ($\ll$) et à droite ($\gg$) sont présents. Eux aussi ne s'appliquent qu'aux entiers.
- les opérateurs arithmétiques unaires +, −, +. et −. sont équivalents à l'opération binaire correspondante avec 0 ou 0. comme première opérande.
- $\sim$ inverse tous les bits de son opérande. ! est une version logique, c'est-à-dire que $!0 = 1$ et si $n \neq 0$, $!n = 0$.

4.8 Opérations sur les états mémoire

Définition 4.4 (Recherche de variable)**.** *La recherche de variable permet d'associer à une variable x une adresse a.*

Chaque fonction peut accéder aux variables locales de la fonction en cours, ainsi qu'aux variables globales.

$$\text{Lookup}((s,g),x) = (|s|,x) \; si \; |s| > 0 \; et \; \exists (x,v) \in s_1$$
$$\text{Lookup}((s,g),x) = x \; si \; (x,v) \in g$$

En entrant dans une fonction, on rajoutera un cadre de pile qui contient les paramètres de la fonction ainsi que ses variables locales. En retournant à l'appelant, il faudra supprimer ce cadre de pile.

Définition 4.5 (Manipulations de pile)**.** *On définit l'empilement d'un cadre de pile* $c = ((x_1, v_1), \ldots, (x_n, v_n))$ *sur un état mémoire* $m = (s, g)$ *(figure 4.9(a)) :*

$$\text{Push}((s,g),c) = (c :: s, g)$$

On définit aussi l'extension du dernier cadre de pile (figure 4.9(b)) :

$$\text{Extend}((c :: s, g), x) = (((x@c) :: s), g)$$

De même on définit le dépilement (figure 4.9(c)) :

$$\text{Pop}(c :: s, g) = (s, g)$$

On définit aussi une opération de nettoyage de pile, qui sera utile pour les retours de fonction.

En effet, si une référence au dernier cadre est toujours présente après le retour, elle pourra se résoudre en un objet différent plus tard dans l'exécution du programme.

La fonction Cleanup est donnée par :

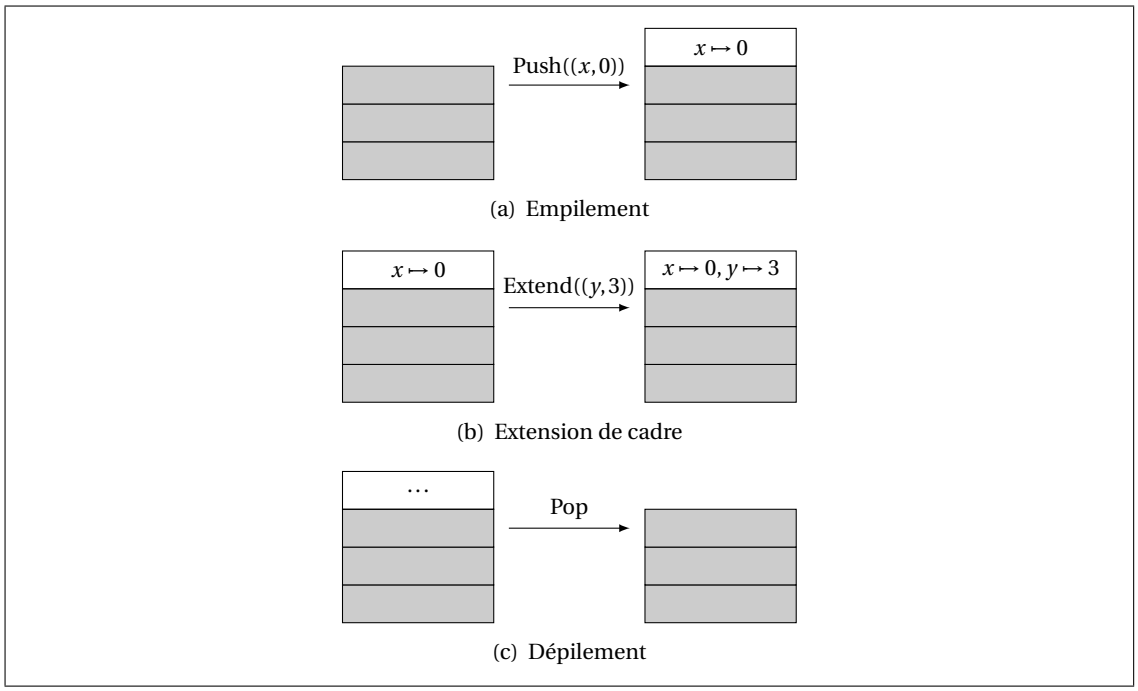


FIGURE 4.9: Opérations de pile

$$
\begin{aligned}
\text{Cleanup}(s,g) &= (s',g') \\
o\grave{u} g' &= \text{CleanupList}(|s|,g) \\
s' &= [\text{CleanupList}(|s|,s_1),\ldots,\text{CleanupList}(|s|,s_n)] \\
\text{CleanupList}(p,u) &= \{(x,v) \in u / v \text{ n'est pas une adresse}\} \\
&\quad \cup \{(x,\varphi) \in u / \text{Live}(p,\varphi)\} \\
\text{Live}(p,(n,x)) &= n < p \\
\text{Live}(p,(x)) &= \text{Vrai} \\
\text{Live}(p,*\varphi) &= \text{Live}(p,\varphi) \\
\text{Live}(p,\varphi.l) &= \text{Live}(p,\varphi) \\
\text{Live}(p,\varphi[n]) &= \text{Live}(p,\varphi)
\end{aligned}
$$

Sans cette règle, examinons le programme suivant :

```
f ()                g (p)           h ()
  (x=0)               (x=0.0)         (p=f())
{                   {               {
  return (&x);        *p = 1;         g(p);
}                   }               }
```

L'exécution de $h()$ donne à p la valeur $(1,x)$. Donc en arrivant dans g, le déréférencement de p va modifier x.

4.9 Accesseurs

On définit ici quelques lentilles.

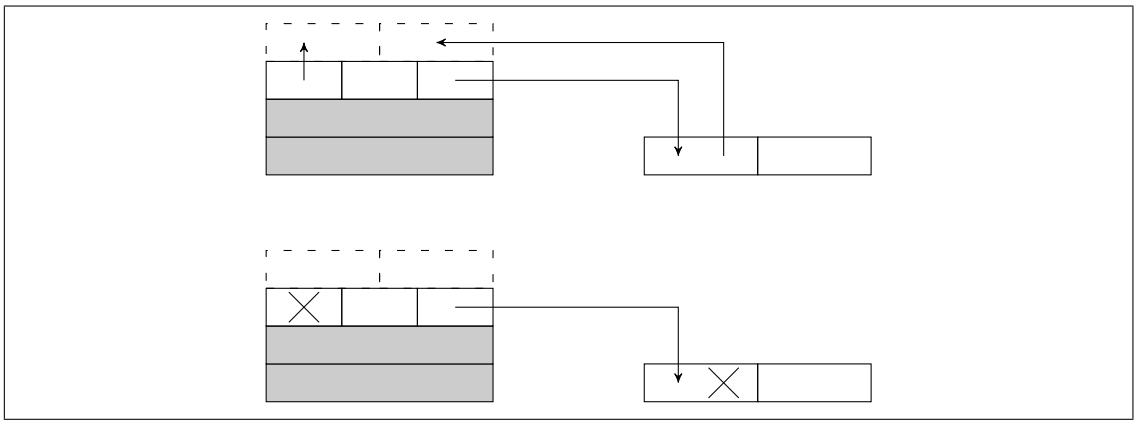


FIGURE 4.10: Nettoyage d'un cadre de pile

Définition 4.6 (Accès à une liste d'associations)**.** *Une liste d'association est une liste de paires (clef, valeur) avec l'invariant supplémentaire que les clefs sont uniques. Il est donc possible de trouver au plus une valeur associée à une clef donnée. L'écriture est également possible, en remplaçant un couple par un couple avec une valeur différente.*

L'accesseur $[\cdot]_L$ *est défini par :*

$$l[x]_L = v \text{ où } \{v\} = \{y/(x,y) \in l\}$$
$$l[x \leftarrow v]_L = (x,v) :: \{(y,v) \in g(x)/y \neq x\})$$

Définition 4.7 (Accès par adresse)**.** *Les états mémoire sont constitués des listes d'association (nom, valeur).*

L'accesseur par adresse $[\cdot]_A$ *permet de généraliser l'accès à ces valeurs en utilisant comme clef non pas un nom mais une adresse.*

Selon cette adresse, on accède soit à la liste des globales, soit à une des listes de la pile des locales.

Pour $m = (s,g)$*,*

$$\begin{aligned}
m[x]_A &= g[x]_L && \textit{Lecture d'une globale}\\
m[(n,x)]_A &= s_{|l|-n+1}[x]_L && \textit{Lecture d'une locale}\\
m[x \leftarrow v]_A &= (s, g[x \leftarrow v]_L) && \textit{Écriture d'une globale}\\
m[(n,x) \leftarrow v]_A &= (s', g) && \textit{Écriture d'une locale}\\
&\text{où } s'_{|l|-n+1} = s_{|l|-n+1}[x \leftarrow v]_L\\
&\forall i \neq |l|-n+1, s'_i = s_i
\end{aligned}$$

Les numéros de cadre qui permettent d'identifier les locales (le n dans (n,x)) croissent avec la pile. D'autre part, l'empilement se fait en tête de liste (près de l'indice 1). Donc pour accéder aux plus vieilles locales (numérotées 1), il faut accéder au dernier élément de la liste. Ceci explique pourquoi un indice $|l|-n+1$ apparaît dans la définition précédente.

Définition 4.8 (Accès par champ)**.** *Les valeurs qui sont des structures possèdent des sous-valeurs, associées à des noms de champ.*

L'accesseur $[\cdot]_L$ *permet de lire et de modifier un champ de ces valeurs.*

L'erreur Ω_{field} *est levée si on accède à un champ non existant.*

$$
\begin{aligned}
\{l_1 : v_1; \ldots; l_n : v_n\}[l]_L &= v_i \ si \ \exists i \in [1;n], l = l_i \\
\{l_1 : v_1; \ldots; l_n : v_n\}[l \leftarrow v]_L &= \{l_1 : v_1 \\
&\quad ; \ldots \\
&\quad ; l_{p-1} : v_{p-1} \\
&\quad ; l_p : v \\
&\quad ; l_{p+1} : v_{p+1} \\
&\quad ; \ldots \\
&\quad ; l_n : v_n\} \ si \ \exists p \in [1;n], l = l_p \\
\{l_1 : v_1; \ldots; l_n : v_n\}[l]_L &= \Omega_{field} \ sinon \\
\{l_1 : v_1; \ldots; l_n : v_n\}[l \leftarrow v]_L &= \Omega_{field} \ sinon
\end{aligned}
$$

Définition 4.9 (Accès par indice)**.** *On définit de même un accesseur $[\cdot]_I$ pour les accès par indice à des valeurs tableaux. Néanmoins le paramètre indice est toujours un entier et pas une expression arbitraire.*

$$
\begin{aligned}
[v_1; \ldots; v_n][i]_I &= v_{i+1} \ si \ i \in [0; n-1] \\
[v_1; \ldots; v_n][i]_I &= \Omega_{array} \ sinon \\
[v_1; \ldots; v_n][i \leftarrow v]_I &= [v'_1; \ldots; v'_n] \ si \ i \in [0; n-1] \\
&\quad o\grave{u} \begin{cases} v'_i = v \\ \forall j \neq i, v'_j = v_j \end{cases} \\
[v_1; \ldots; v_n][i \leftarrow v]_I &= \Omega_{array} \ sinon
\end{aligned}
$$

Définition 4.10 (Accès par chemin)**.** *L'accès par chemin Φ permet de lire et de modifier la mémoire en profondeur.*

On peut accéder directement à une variable :

$$\Phi(a) = A(a)$$

Les accès à des sous-valeurs se font en composant les accesseurs (définition 4.3) :

$$
\begin{aligned}
\Phi(\varphi.l) &= \Phi(\varphi) \ggg L(l) \\
\Phi(\varphi[i]) &= \Phi(\varphi) \ggg I(i)
\end{aligned}
$$

Le déréférencement est défini comme suit :

$$
\begin{aligned}
m[*\varphi]_\Phi &= m[\varphi']_\Phi \ o\grave{u} \ \varphi' = m[\varphi]_\Phi \\
m[*\varphi \leftarrow v]_\Phi &= m[\varphi' \leftarrow v]_\Phi \ o\grave{u} \ \varphi' = m[\varphi]_\Phi
\end{aligned}
$$

Enfin, l'accès à la mémoire par le pointeur nul provoque une erreur :

$$m[\text{Null}]_\Phi = \Omega_{ptr}$$
$$m[\text{Null} \leftarrow v]_\Phi = \Omega_{ptr}$$

Cette dernière définition mérite une explication. Dans le cas de la lecture, il suffit d'appliquer les bons accesseurs : $[\cdot]_L$ pour $\varphi.l$, etc.

En revanche, la modification est plus complexe. Les deux premiers cas ($\varphi = a$ et $\varphi = *\varphi'$) modifient directement une valeur complète (en modifiant une association), mais les deux suivants ($\varphi = \varphi'.l$ et $\varphi = \varphi'[i]$) ne font qu'altérer une sous-valeur existante. Il est donc nécessaire de procéder en 3 étapes :

- obtenir la valeur à modifier (soit $m[\varphi]_\varphi$)
- construitre une valeur altérée (en appliquant par exemple $[l \leftarrow v]_L$)
- affecter cette valeur au même chemin (le $m[\varphi \leftarrow \ldots]_\varphi$ externe)

Dans la suite, on notera uniquement $[\cdot]$ tous ces accesseurs lorsque ce n'est pas ambigü.

4.10 Contextes d'évaluation

L'évaluation des expressions repose sur la notion de contextes d'évaluation. L'idée est que si on peut évaluer une expression, alors on peut évaluer une expression qui contient celle-ci.

Par exemple, supposons que $\langle f(3), m\rangle \rightarrow \langle 2, m\rangle$. Alors on peut ajouter la constante 1 à gauche de chaque expression sans changer le résultat : $\langle 1 + f(3), m\rangle \rightarrow \langle 1 + 2, m\rangle$. On a utilisé le même contexte $C = 1 + \bullet$.

Pour pouvoir raisonner en termes de contextes, 3 points sont nécessaires :

- comment découper une expression selon un contexte
- comment appliquer une règle d'évaluation sous un contexte
- comment regrouper une expression et un contexte

Le premier point consiste à définir les contextes eux-mêmes (figure 4.11).

Le deuxième est résolu les règles d'inférence suivantes :

$$\frac{\langle e, m\rangle \rightarrow \langle e', m'\rangle}{\langle C(\!|e|\!), m\rangle \rightarrow \langle C(\!|e'|\!), m'\rangle}\ (\textsc{Ctx}) \qquad \frac{\langle l\,v, m\rangle \rightarrow \langle l\,v', m'\rangle}{\langle C_L(\!|l\,v|\!)_L, m\rangle \rightarrow \langle C_L(\!|l\,v'|\!)_L, m'\rangle}\ (\textsc{Ctx-Lv})$$

$$\frac{\langle i, m\rangle \rightarrow \langle i', m'\rangle}{\langle C_I(\!|i|\!)_I, m\rangle \rightarrow \langle C_I(\!|i'|\!)_I, m'\rangle}\ (\textsc{Ctx-Instr})$$

Enfin, le troisième revient à définit l'opérateur de substitution $\cdot(\!|\cdot|\!)$ présent dans la règle précédente. Afin de pouvoir appliquer des substitution au niveau des left-values et des instructions, on définit aussi respectivement $\cdot(\!|\cdot|\!)_L$ et $\cdot(\!|\cdot|\!)_I$.

Dans la définition de l'ensemble des contextes, chaque cas hormis le cas de base fait apparaître exactement un "C". Chaque contexte est donc constitué d'exactement un "trou" $\bullet$ (une dérivation de C est toujours linéaire). L'opération de substitution consiste à remplacer ce trou, comme décrit dans la figure 4.12.

Par exemple, décomposons l'évaluation de $e_1 \boxplus e_2$ en $v = v_1 \mathbin{\widehat{\boxplus}} v_2$ depuis un état mémoire m :

Contextes	$C ::= C_L$
	$\mid\ C \boxplus e$
	$\mid\ v \boxplus C$
	$\mid\ \boxminus C$
	$\mid\ C \leftarrow e$
	$\mid\ \varphi \leftarrow C$
	$\mid\ \{l_1 : v_1; \ldots; l_i : C; \ldots; l_n : e_n\}$
	$\mid\ [v_1; \ldots; C; \ldots; e_n]$
	$\mid\ C(e_1, \ldots, e_n)$
	$\mid\ f(v_1, \ldots, C, \ldots, e_n)$
Contextes (left-values)	$C_L ::= \bullet$
	$\mid\ * C_L$
	$\mid\ C_L.l_S$
	$\mid\ C_L[e]$
	$\mid\ \varphi[C]$
Contextes (instructions)	$C_I ::= C_I; i$
	$\mid\ \text{IF}(C)\{i_1\}\text{ELSE}\{i_2\}$
	$\mid\ \text{RETURN}(C)$
	$\mid\ \text{DECL}\ x = C\ \text{IN}\{i\}$
	$\mid\ C$

FIGURE 4.11: Contextes d'exécution

1. on commence par évaluer, d'une manière ou d'une autre, l'expression e_1 en une valeur v_1. Le nouvel état mémoire est noté m'. Soit donc $\langle e_1, m\rangle \rightarrow \langle v_1, m'\rangle$.
2. En appliquant la règle CTX avec $C = \bullet \boxplus e_2$ (qui est une des formes possibles pour un contexte d'évaluation), on déduit de 1. que $\langle e_1 \boxplus e_2, m\rangle \rightarrow^* \langle v_1 \boxplus e_2, m'\rangle$
3. D'autre part, on évalue e_2 depuis m'. En supposant encore que l'évaluation converge, notons v_2 la valeur calculée et m'' l'état mémoire résultant : $\langle e_2, m'\rangle \rightarrow \langle v_2, m''\rangle$.
4. Appliquons la règle CTX à 3. avec $C = v_1 \boxplus \bullet$. On obtient $\langle v_1 \boxplus e_2, m\rangle \rightarrow^* \langle v_1 \boxplus v_2, m'\rangle$.
5. En combinant les résultats de 2. et 4. on en déduit que $\langle e_1 \boxplus e_2, m\rangle \rightarrow^* \langle v_1 \boxplus v_2, m''\rangle$.
6. D'après la règle EXP-BINOP, $\langle v_1 \boxplus v_2, m''\rangle \rightarrow^* \langle v_1 \widehat{\boxplus} v_2, m''\rangle$
7. D'après 5. et 6., on a par combinaison $\langle e_1 \boxplus e_2, m\rangle \rightarrow^* \langle v, m''\rangle$ en posant $v = v_1 \widehat{\boxplus} v_2$.

4.11 Expressions

Définition 4.11 (Évaluation d'une expression)**.** *L'évaluation d'une expression e se fait sous un état mémoire particulier m et est susceptible de modifier celui-ci en le transformant en un nouveau m'. Le résultat est toujours une valeur v, c'est-à-dire que nous présentons pour les expressions une sémantique à grands pas. Cette évaluation est notée :*

$$
\begin{aligned}
\bullet(\!|e_0|\!) &= e_0 \\
(C \boxplus e)(\!|e_0|\!) &= C(\!|e_0|\!) \boxplus e \\
(v \boxplus C)(\!|e_0|\!) &= v \boxplus C(\!|e_0|\!) \\
(\boxminus C)(\!|e_0|\!) &= \boxminus C(\!|e_0|\!) \\
(*C)(\!|e_0|\!) &= * C(\!|e_0|\!) \\
(C.l_S)(\!|e_0|\!) &= C(\!|e_0|\!).l_S \\
(\varphi[C])(\!|e_0|\!) &= \varphi[C(\!|e_0|\!)] \\
(C[e])(\!|e_0|\!) &= C(\!|e_0|\!)[e] \\
(C \leftarrow e)(\!|e_0|\!) &= C(\!|e_0|\!) \leftarrow e \\
(\varphi \leftarrow C)(\!|e_0|\!) &= \varphi \leftarrow C(\!|e_0|\!) \\
\{l_1 : v_1; \ldots; l_i : C; \ldots; l_n : e_n\}(\!|e_0|\!) &= \{l_1 : v_1; \ldots; l_i : C(\!|e_0|\!); \ldots; l_n : e_n\} \\
[v_1; \ldots; C; \ldots; e_n](\!|e_0|\!) &= [v_1; \ldots; C(\!|e_0|\!); \ldots; e_n] \\
C(e_1, \ldots, e_n)(\!|e_0|\!) &= C(\!|e_0|\!)(e_1, \ldots, e_n) \\
f(v_1, \ldots, C, \ldots, e_n)(\!|e_0|\!) &= f(v_1, \ldots, C(\!|e_0|\!), \ldots, e_n)
\end{aligned}
$$

$$
\begin{aligned}
(C; i)(\!|e_0|\!) &= C(\!|e_0|\!); i \\
(\textsc{If}(C)\{i_1\}\textsc{Else}\{i_2\})(\!|e_0|\!) &= \textsc{If}(C(\!|e_0|\!))\{i_1\}\textsc{Else}\{i_2\} \\
(\textsc{Return}(C))(\!|e_0|\!) &= \textsc{Return}(C(\!|e_0|\!))
\end{aligned}
$$

$$
\begin{aligned}
C_L(\!|e_0|\!) &= C_L(\!|e_0|\!)_L \\
C \boxplus e(\!|e_0|\!) &= C(\!|e_0|\!) \boxplus e \\
v \boxplus C(\!|e_0|\!) &= v \boxplus C(\!|e_0|\!) \\
\boxminus C(\!|e_0|\!) &= \boxminus C(\!|e_0|\!) \\
C \leftarrow e(\!|e_0|\!) &= C(\!|e_0|\!) \leftarrow e \\
\varphi \leftarrow C(\!|e_0|\!) &= \varphi \leftarrow C(\!|e_0|\!) \\
\{l_1 : v_1; \ldots; l_i : C; \ldots; l_n : e_n\}(\!|e_0|\!) &= \{l_1 : v_1; \ldots; l_i : C(\!|e_0|\!); \ldots; l_n : e_n\} \\
[v_1; \ldots; C; \ldots; e_n](\!|e_0|\!) &= [v_1; \ldots; C(\!|e_0|\!); \ldots; e_n] \\
C(e_1, \ldots, e_n)(\!|e_0|\!) &= C(\!|e_0|\!)(e_1, \ldots, e_n) \\
f(v_1, \ldots, C, \ldots, e_n)(\!|e_0|\!) &= f(v_1, \ldots, C(\!|e_0|\!), \ldots, e_n)
\end{aligned}
$$

$$
\begin{aligned}
\bullet(\!|l_0|\!)_L &= \bullet \\
*C_L(\!|l_0|\!)_L &= * C_L(\!|l_0|\!)_L \\
C_L.l_S(\!|l_0|\!)_L &= C_L(\!|l_0|\!)_L.l_S \\
C_L[e](\!|l_0|\!)_L &= C_L(\!|l_0|\!)_L[e] \\
\varphi[C](\!|e_0|\!) &= \varphi[C(\!|e_0|\!)]
\end{aligned}
$$

FIGURE 4.12: Substitution dans les contextes d'évaluation

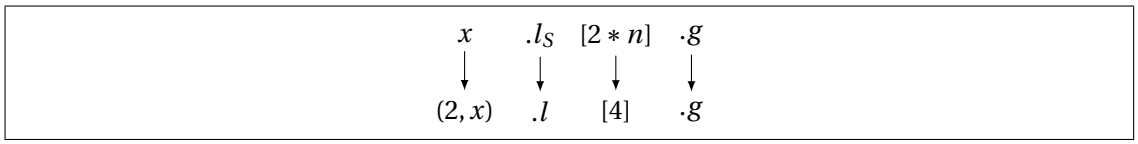


FIGURE 4.13: Évaluation des left-values.

$$\langle e, m\rangle \to \langle v, m'\rangle$$

Définition 4.12 (Évaluation d'une left-value)**.** *L'évaluation d'une left-value l v produit un "chemin" φ dans une variable, qui est en fait équivalent à une left-value dont toutes les sous-expressions (d'indices) ont été évaluées.*

On note :

$$\langle l\, v, m\rangle \to \langle \varphi, m'\rangle$$

Puisque des left-values peuvent apparaitre dans les expressions, et des expressions dans les left-values (en indice de tableau), leurs règles d'évaluation sont mutuellement récursives.

Left-values

Obtenir un chemin à partir d'un nom de variable revient à résoudre le nom de cette variable : est-elle accessible ? Le nom désigne-t-il une variable locale ou une variable globale ?

$$\frac{a = \text{Lookup}(x, m)}{\langle x, m\rangle \to \langle a, m\rangle}\ (\textsc{Phi-Var})$$

Les règles portant sur le déréférencement et l'accès à un champ de structure sont similaires : on commence par évaluer la left-value sur laquelle porte ce modificateur, et on place le même modificateur sur le chemin résultant. Dans le cas des champs de structure, la définition de la structure S n'est pas prise en compte.

$$\frac{}{\langle *\varphi, m\rangle \to \langle \widehat{*}\varphi, m\rangle}\ (\textsc{Phi-Deref}) \qquad \frac{}{\langle l\, v.l_S, m\rangle \to \langle l\, \widehat{v.l}, m\rangle}\ (\textsc{Phi-Struct})$$

Enfin, pour évaluer un chemin dans un tableau, on commence par procéder comme précédemment, c'est-à-dire en évaluant la left-value sur laquelle porte l'opération d'indexation. Puis on évalue l'expression d'indice en une valeur qui permet de construire le chemin résultant.

$$\frac{}{\langle \varphi[n], m\rangle \to \langle \widehat{\varphi[n]}, m\rangle}\ (\textsc{Phi-Array})$$

Notons qu'en procédant ainsi, on évalue les left-values de gauche à droite : dans l'expression $x[e_1][e_2][e_3]$, e_1 est évalué en premier, puis e_2, puis e_3.

Un exemple d'évaluation est donné dans la figure 4.13.

Expressions

Évaluer une constante est le cas le plus simple, puisqu'en quelque sorte celle-ci est déjà évaluée. À chaque constante syntaxique c, on peut associer une valeur sémantique $\widehat{c}$. Par exemple, au chiffre (symbole) 3, on associe le nombre (entier) $\widehat{3}$.

$$\frac{}{\langle c, m\rangle \to \langle \widehat{c}, m\rangle} \text{ (EXP-CST)}$$

De même, une fonction est déjà évaluée :

$$\frac{}{\langle f, m\rangle \to \langle \widehat{f}, m\rangle} \text{ (EXP-FUN)}$$

Pour lire le contenu d'un emplacement mémoire (left-value), il faut tout d'abord l'évaluer en un chemin.

$$\frac{}{\langle \varphi, m\rangle \to \langle m[\varphi]_\Phi, m\rangle} \text{ (EXP-LV)}$$

Pour évaluer une expression constituée d'un opérateur, on évalue une expression, puis l'autre (l'ordre d'évaluation, est encore imposé : de gauche à droite). À chaque opérateur $\boxplus$, correspond un opérateur sémantique $\widehat{\boxplus}$ qui agit sur les valeurs. Par exemple, l'opérateur $\widehat{+}$ est l'addition classique entre entiers. Comme précisé dans la section 4.7, la division par zéro via /, % ou /. provoque l'erreur Ω_{div}.

$$\frac{}{\langle \boxminus\ v, m\rangle \to \langle \widehat{\boxminus}\ v, m\rangle} \text{ (EXP-UNOP)} \qquad \frac{}{\langle v_1 \boxplus v_2, m\rangle \to \langle v_1 \widehat{\boxplus} v_2, m\rangle} \text{ (EXP-BINOP)}$$

Il est nécessaire de dire un mot sur les opérations $\widehat{+_p}$ et $\widehat{-_p}$ définissant l'arithmétique des pointeurs. Celles-ci sont uniquement définies pour les références mémoire à un tableau, c'est-à-dire celles qui ont la forme $\varphi[n]$. On a alors :

$$\begin{aligned} \varphi[n] +_p m &= \varphi[n+m] \\ \varphi[n] -_p m &= \varphi[n-m] \end{aligned}$$

Cela implique qu'on ne peut pas faire faire d'arithmétique de pointeurs au sein d'une même structure. Autrement c'est une erreur de manipulation de pointeurs :

$$\begin{aligned} \varphi +_p m &= \Omega_{ptr} \text{ si } \nexists(\varphi', n), \varphi = \varphi'[n] \\ \varphi -_p m &= \Omega_{ptr} \text{ si } \nexists(\varphi', n), \varphi = \varphi'[n] \\ \text{NULL} +_p m &= \Omega_{ptr} \\ \text{NULL} -_p m &= \Omega_{ptr} \end{aligned}$$

Pour prendre l'adresse d'une variable, il suffit de résoudre celle-ci dans l'état mémoire courant.

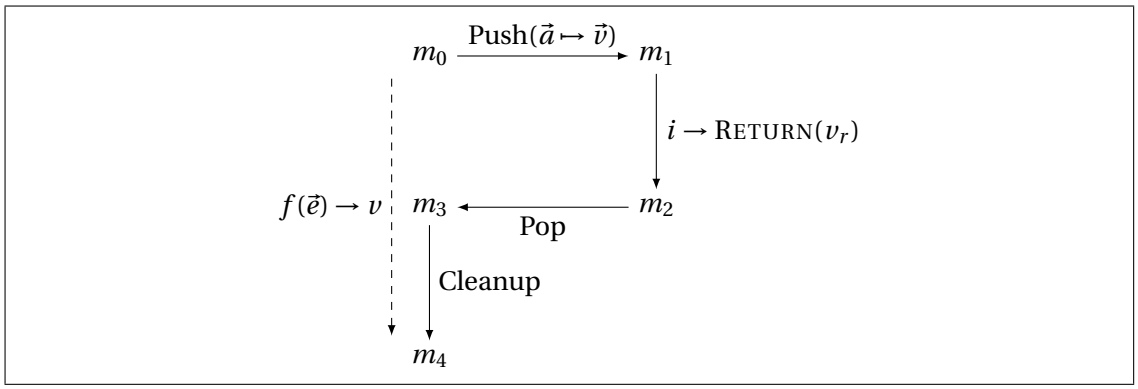


FIGURE 4.14: Appel d'une fonction. La taille de la pile croît de gauche à droite, et les réductions se font de haut en bas.

$$\frac{}{\langle \& \varphi, m \rangle \to \langle \widehat{\&} \varphi, m \rangle} \text{ (EXP-ADDROF)}$$

L'affectation se déroule 3 étapes : d'abord, l'expression est évaluée en une valeur v. Ensuite, la left-value est évaluée en un chemin φ. Enfin, un nouvel état mémoire est construit, où la valeur accessible par φ est remplacée par v. Comme dans le langage C, l'expression d'affectation produit une valeur, qui est celle qui a été affectée.

$$\frac{}{\langle \varphi \leftarrow v, m \rangle \to \langle v, m[\varphi \leftarrow v]_\Phi \rangle} \text{ (EXP-SET)}$$

Expressions composées

Les littéraux de structures sont évalués en leurs constructions syntaxiques respectives. Puisque les contextes d'évaluation sont de la forme $[v_1; \ldots; C; \ldots; e_n]$, l'évaluation se fait toujours de gauche à droite.

$$\frac{}{\langle \{l_1 : v_1; \ldots; l_n : v_n\}, m \rangle \to \langle \{l_1 : \widehat{v_1; \ldots; l_n} : v_n\}, m \rangle} \text{ (EXP-STRUCT)}$$

$$\frac{}{\langle [v_1, \ldots, v_n], m \rangle \to \langle [\widehat{v_1, \ldots, v_n}], m \rangle} \text{ (EXP-ARRAY)}$$

Les contextes utilisés dans l'appel de fonction sont similaires. Tout d'abord, les arguments sont évalués et placés dans un nouveau cadre de pile. Ensuite, le corps de la fonction est évalué jusqu'à se réduire en une instruction RETURN(v). Enfin, le cadre précédemment utilisé est dépilé.

La dernière étape consiste à nettoyer la mémoire de références à l'ancien cadre de pile (on utilise la fonction Cleanup définie dans la section 4.8).

$$\frac{\begin{array}{c} f = \text{fun}(a_1, \ldots, a_n)\{i\} \qquad m_1 = \text{Push}(m_0, ((a_1 \mapsto v_1), \ldots, (a_n \mapsto v_n))) \\ \langle i, m_1 \rangle \to \langle \text{RETURN}(v), m_2 \rangle \qquad m_3 = \text{Pop}(m_2) \qquad m_4 = \text{Cleanup}(m_3) \end{array}}{\langle f(v_1, \ldots, v_n), m_0 \rangle \to \langle v, m_4 \rangle} \text{ (EXP-CALL)}$$

Cette évaluation est décrite dans la figure 4.14.

4.12 Instructions

Contrairement à l'évaluation des expressions, on choisit une sémantique de réécriture à petits pas. La sémantique fonctionne de la manière suivante : partant d'un état mémoire m, on veut exécuter une instruction i. Les règles d'évaluation suivantes permettent de réduire le problème en se ramenant à l'exécution d'une instruction i' "plus simple" en partant d'un état mémoire m'. Un tel pas est noté :

$$\langle i, m\rangle \to \langle i', m'\rangle$$

Par exemple, exécuter $x \leftarrow 3; y \leftarrow x$ revient à évaluer $y \leftarrow x$ depuis un état mémoire dans lequel on a déjà réalisé la première affectation. La seconde affectation se réalise de même et permet de réécrire l'instruction restante en PASS :

$$\begin{aligned}\langle (x \leftarrow 3; y \leftarrow x), m\rangle &\to \langle y \leftarrow x, m[x \mapsto \widehat{3}]\rangle \\ &\to \langle \text{PASS}, m[x \mapsto \widehat{3}][y \mapsto \widehat{3}]\rangle\end{aligned}$$

Il n'est pas possible de réduire plus loin l'instruction PASS. Dans un tel cas, l'évaluation est terminée.

Les seuls cas terminaux sont PASS et RETURN(e).

Les cas de la séquence et de l'affectation ont été utilisés dans l'exemple ci-dessus.

$$\frac{\langle i, m\rangle \to \langle \text{PASS}, m'\rangle}{\langle (i; i'), m\rangle \to \langle i', m'\rangle}\ (\text{SEQ}) \qquad \frac{}{\langle (\text{PASS}; i), m\rangle \to \langle i, m\rangle}\ (\text{PASS}) \qquad \frac{}{\langle v, m\rangle \to \langle \text{PASS}, m\rangle}\ (\text{EXP})$$

L'évaluation de DECL $x = v$ IN$\{i\}$ sous m se fait en trois parties :

- on crée un environnement mémoire m' en ajoutant à m l'association $x \mapsto v$.
- sous m', on évalue i jusqu'à atteindre une instruction PASS ou RETURN(v_r) sous un état m''. C'est cette instruction qui sera retournée.
- on enlève x de m'' et on enlève de cet état mémoire les pointeurs invalides.

$$\frac{m' = \text{Extend}(m, x, v) \qquad \langle i, m'\rangle \to \langle \text{PASS}, m''\rangle \qquad m''' = \text{Cleanup}(m'' - x)}{\langle \text{DECL } x = v \text{ IN}\{i\}, m\rangle \to \langle \text{PASS}, m'''\rangle}\ (\text{DECL})$$

$$\frac{\begin{array}{c} m' = \text{Extend}(m, x, v) \\ \langle i, m'\rangle \to \langle \text{RETURN}(v_r), m''\rangle \qquad m''' = \text{Cleanup}(m'' - x)\end{array}}{\langle \text{DECL } x = v \text{ IN}\{i\}, m\rangle \to \langle \text{RETURN}(v_r), m'''\rangle}\ (\text{DECL-RETURN})$$

Pour traiter l'alternative, on a besoin de 2 règles. Elles commencent de la même manière, en évaluant la condition. Si le résultat est 0 (et seulement dans ce cas), c'est la règle IF-FALSE qui est appliquée et l'instruction revient à évaluer la branche "else". Dans les autres cas, c'est la règle IF-TRUE qui s'applique et la branche "then" qui est prise.

$$\frac{}{\langle \text{IF}(0)\{i_t\}\text{ELSE}\{i_f\}, m\rangle \to \langle i_f, m\rangle}\ (\text{IF-FALSE}) \qquad \frac{v \neq 0}{\langle \text{IF}(v)\{i_t\}\text{ELSE}\{i_f\}, m\rangle \to \langle i_t, m\rangle}\ (\text{IF-TRUE})$$

Pour traiter la boucle, on peut être tenté de procéder de la même manière :

$$\frac{v \neq 0}{\langle \text{WHILE}(v)\{i\}, m\rangle \rightarrow \langle \text{PASS}, m\rangle} \text{ (WHILE-FALSE-BAD)}$$

$$\frac{}{\langle \text{WHILE}(0)\{i\}, m\rangle \rightarrow \langle i; \text{WHILE}(e)\{i\}, m'\rangle} \text{ (WHILE-TRUE-BAD)}$$

Mais la seconde règle est impossible : puisque e a déjà été évaluée, il est impossible de la réintroduire non évaluée en partie droite.

À la place, on exprime la sémantique de la boucle comme une simple règle de réécriture :

$$\frac{}{\langle \text{WHILE}(e)\{i\}, m\rangle \rightarrow \langle \text{IF}(e)\{i; \text{WHILE}(e)\{i\}\}, m\rangle} \text{ (WHILE)}$$

Cette règle revient à dire qu'on peut dérouler une boucle. Pour la comprendre, on peut remarquer qu'une boucle "while" est en réalité équivalente une infinité de "if" imbriqués.

```
while(e) {
    i
}
```

$\widehat{=}$

```
if(e) {
    i;
    if(e) {
        i;
        if(e) {
            i;
            if(e) {
                i;
                ...
            }
        }
    }
}
```

Donc en remplaçant le second "if" par le "while", on obtient :

```
while(e) {
    i
}
```

$\widehat{=}$

```
if(e) {
    i;
    while(e) {
        i
    }
}
```

Enfin, si un "return" apparaît dans une séquence, on peut supprimer la suite :

$$\frac{}{\langle \text{RETURN}(v); i, m\rangle \rightarrow \langle \text{RETURN}(e), m\rangle} \text{ (RETURN)}$$

4.13 Erreurs

Les erreurs se propagent des données vers l'interprète ; c'est-à-dire que si une expression ou instruction est réduite en une valeur d'erreur Ω, alors une transition est faite vers cet état d'erreur.

Cela est aussi vrai d'une sous-expression ou sous-instruction : si l'évaluation de e_1 provoque une erreur, l'évaluation de $e_1 + e_2$ également. La notion de sous-expression ou sous instruction est définie en fonction des contextes C (figure 4.15).

$\boxed{\Xi \to \Omega}$

$$\frac{}{\langle \Omega, m\rangle \to \Omega}\ (\textsc{Exp-Err}) \qquad \frac{\langle e, m\rangle \to \Omega}{\langle C(\!|e|\!), m\rangle \to \Omega}\ (\textsc{Eval-Err})$$

FIGURE 4.15: Évaluation – cas d'erreur

$\boxed{m \Vdash p \to m'}$

$$\frac{\langle e, m\rangle \to \langle v, m'\rangle}{m \Vdash e \to m'}\ (\text{T-}\textsc{Exp}) \qquad \frac{\langle e, m\rangle \to \langle v, m'\rangle}{(s, g) \Vdash x = e \to (s, (x \mapsto v) :: g)}\ (\text{T-}\textsc{Var})$$

FIGURE 4.16: Évaluation des phrases d'un programme

4.14 Phrases

Un programme est constitué d'une suite de phrases : déclarations de fonctions, de variables et de types, et évaluation d'expressions.

Donc l'évaluation d'une phrase p fait passer d'un état mémoire m à un autre m', ce que l'on note $m \Vdash p \to m'$.

L'évaluation d'une expression est uniquement faite pour ses effets de bord. Par exemple, après avoir défini les fonctions du programme, on pourra appeller main(). La déclaration d'une variable globale (avec un initialiseur), quant à elle consiste à évaluer cet initialiseur et à étendre l'état mémoire avec ce couple (variable, valeur) (figure 4.16).

4.15 Exécution

L'exécution d'un programme est sans surprise l'exécution de ses phrases, les unes à la suite des autres.

On commence par étendre l'extension $\to^*$ au listes de la relation $\to$:

$$\frac{}{m \Vdash [\,] \to^* m}\ (\text{T}^*\text{-}\textsc{Nil}) \qquad \frac{m \Vdash p \to m' \qquad m' \Vdash ps \to^* m''}{m \Vdash p :: ps \to^* m''}\ (\text{T}^*\text{-}\textsc{Cons})$$

L'exécution d'un programme est définie par :

$$\frac{([\,], [\,]) \Vdash P \to^* m}{\Vdash P \to^* m}\ (\textsc{Prog})$$

4.16 Exemple : l'algorithme d'Euclide

Version par divisions successives :

```
function gcd(a, b)
  var t = 0;
  while b != 0
    t = b
    b = a mod b
    a = t
  return a
```

Soit :

$$f(a,b)(t=0)\{\textsc{While}(b\neq 0)\{t\leftarrow b; b\leftarrow a\%b; a\leftarrow t\}; \textsc{Return}(a)\}$$

$$\langle f(1071,462), m\rangle \rightarrow ?$$

$$\langle \textsc{While}(b\neq 0)\{t\leftarrow b; b\leftarrow a\%b; a\leftarrow t\}; \textsc{Return}(a), m[a\mapsto 1071][b\mapsto 462][t\mapsto 0]\rangle \rightarrow ?$$

(on notera cet état $s_0 = \langle i_0, m_0\rangle$)

$$\langle a=0, m_0\rangle \rightarrow \langle 0, m_0\rangle$$

donc

$$\langle \textsc{If}(a=0)\{\textsc{Return}(b)\}, m_0\rangle \rightarrow \langle \textsc{Pass}, m[a\mapsto 1071][b\mapsto 462]\rangle$$

$$
\begin{aligned}
s_0 &\to \langle \textsc{If}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t; \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}\}; \textsc{Return}(a), m_0\rangle && (4.1)\\
&\to \langle t \leftarrow b; b \leftarrow a\%b; a \leftarrow t; \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}; \textsc{Return}(a), m_0\rangle && (4.2)\\
&\to \langle b \leftarrow a\%b; a \leftarrow t; \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}; \textsc{Return}(a), m_0\rangle && (4.3)\\
&\to \langle a \leftarrow t; \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}; \textsc{Return}(a), m_0''\rangle && (4.4)\\
&\to \langle \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}; \textsc{Return}(a), m_1\rangle && (4.5)\\
&\to \langle \textsc{If}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t; \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}\}; \textsc{Return}(a), m_1\rangle && (4.6)\\
&\to \langle t \leftarrow b; b \leftarrow a\%b; a \leftarrow t; \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}; \textsc{Return}(a), m_1\rangle && (4.7)\\
&\to \langle \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}; \textsc{Return}(a), m_2\rangle && (4.8)\\
&\to \langle \textsc{If}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t; \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}\}; \textsc{Return}(a), m_2\rangle && (4.9)\\
&\to \langle t \leftarrow b; b \leftarrow a\%b; a \leftarrow t; \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}; \textsc{Return}(a), m_2\rangle && (4.10)\\
&\to \langle \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}; \textsc{Return}(a), m_3\rangle && (4.11)\\
&\to \langle \textsc{If}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t; \textsc{While}(b \neq 0)\{t \leftarrow b; b \leftarrow a\%b; a \leftarrow t\}\}; \textsc{Return}(a), m_3\rangle && (4.12)\\
&\to \langle \textsc{Pass}; \textsc{Return}(a), m_3\rangle && (4.13)\\
&\to \langle \textsc{Return}(a), m_3\rangle && (4.14)
\end{aligned}
$$

$$
\begin{aligned}
m_0' &= m_0[t \mapsto 462] = m[a \mapsto 1071][b \mapsto 462][t \mapsto 462]\\
m_0'' &= m_0'[b \mapsto 147] = m[a \mapsto 1071][b \mapsto 147][t \mapsto 462]\\
m_1 &= m_0''[a \mapsto 462] = m[a \mapsto 462][b \mapsto 147][t \mapsto 462]\\
m_2 &= m_1[t \mapsto 147][b \mapsto 21][a \mapsto 147] = m[a \mapsto 147][b \mapsto 21][t \mapsto 147]\\
m_3 &= m_2[t \mapsto 21][b \mapsto 0][a \mapsto 21] = m[a \mapsto 21][b \mapsto 0][t \mapsto 21]
\end{aligned}
$$

CHAPITRE 5

TYPAGE

Dans ce chapitre, nous enrichissons le langage défini dans le chapitre 4 d'un système de types. Celui-ci permet de séparer les programmes bien formés, comme celui de la figure 5.1(a) des programmes mal formés comme celui de la figure 5.1(b).

Le but d'un tel système de types est de rejeter les programmes qui sont "évidemment faux", c'est-à-dire dont on peut prouver qu'il provoqueraient des erreurs à l'exécution dues à une incompatibilité entre valeurs. En ajoutant cette étape, on restreint la classe d'erreurs qui pourraient bloquer la sémantique.

5.1 Principe

Le principe est d'associer à chaque construction syntaxique une étiquette représentant le genre de valeurs qu'elle produira. Dans le programme de la figure 5.1(a), la variable x est initialisée avec la valeur 0, c'est donc un entier. Cela signifie que dans tout le programme, toutes les instances de cette variable [1] porteront ce type. La première instruction est l'affectation de la constante 1 (entière) à x dont on sait qu'elle porte des valeurs entières, ce qui est donc correct. Le fait de rencontrer RETURN(x) permet de conclure que le type de la fonction est () → INT.

Dans la seconde fonction, au contraire, l'opérateur $*$ est appliqué à x (le début de l'analyse est identique et permet de conclure que x porte des valeurs entières). Or cet opérateur prend un argument d'un type pointeur de la forme $t*$ et renvoie alors une valeur de type t.

1. Deux variables peuvent avoir le même nom dans deux fonctions différentes, par exemple. Dans ce cas il n'y a aucune contrainte particulière entre ces deux variables. L'analyse de typage se fait toujours dans un contexte précis.

```
f()
(x=0)
{
  x = 1
  return x
}
```

(a) Programme bien formé

```
f()
(x=0)
{
  x = 1
  return (*x)
}
```

(b) Programme mal formé

FIGURE 5.1: Programmes bien et mal formés

Type	$t ::=$ INT	Entier
	| FLOAT	Flottant
	| UNIT	Unité
	| $t*$	Pointeur
	| $t[]$	Tableau
	| S	Structure
	| $(t_1, \ldots, t_n) \rightarrow t$	Fonction
Structure	$S ::= \{l_1 : t_1; \ldots; l_n : t_n\}$	Structure simple
Environnement de typage	$\Gamma ::= []$	Environnement vide
	| $(a, t) :: \Gamma'$	Extension

FIGURE 5.2: Types et environnements de typage

Ceci est valable pour tout t (INT, FLOAT où même $t'*$: le déréférencement d'un pointeur sur pointeur donne un pointeur), mais le type de x, INT, n'est pas de cette forme. Ce programme est donc mal typé.

5.2 Environnements et notations

Les types associés aux expressions sont décrits dans la figure 5.2. Tous sont des types concrets : il n'y a pas de polymorphisme.

Pour maintenir les contextes de typage, un environnement Γ associe un type à un ensemble de variables.

Plus précisément, un environnement Γ est une liste de couples (variable, type).

Par exemple, $(p, \text{INT}*) \in \Gamma$ permet de typer (sous Γ) l'expression p en INT$*$, $*p$ en INT et $p +_p 4$ en INT$*$.

Le type des fonctions semble faire apparaître un n-uplet $(t_1, \ldots, t_n)$ mais ce n'est qu'une notation : il n'y a pas de n-uplets de première classe, ils sont toujours présents dans un type fonctionnel.

Typage d'une expression : on note de la manière suivante le fait qu'une expression e (telle que définie dans la figure 4.4) ait pour type t dans le contexte Γ.

$$\Gamma \vdash e : t$$

Typage d'une instruction : les instructions n'ont en revanche pas de type. Mais il est tout de même nécessaire de vérifier que toutes les sous-expressions apparaissant dans une instruction sont cohérentes ensemble.

On note de la manière suivante le fait que sous l'environnement Γ l'instruction i est bien typée :

$$\Gamma \vdash i$$

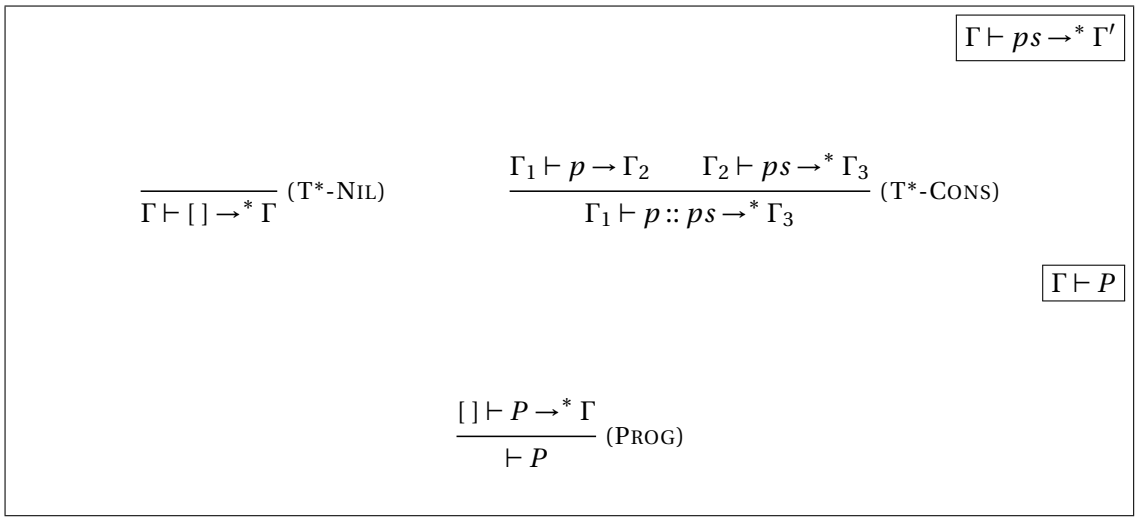


FIGURE 5.3: Typage d'une suite de phrases et d'un programme

Typage d'une phrase : De par leur nature séquentielle, les phrases qui composent un programme altèrent l'environnement de typage. Par exemple, la déclaration d'une variable globale ajoute une valeur dans l'environnement.

On note

$$\Gamma \vdash p \rightarrow \Gamma'$$

si le typage de la phrase p transforme l'environnement Γ en Γ'.

On étend cette notation aux suites de phrases, ce qui définit le typage d'un programme, ce que l'on note $\vdash P$ (figure 5.3).

5.3 Expressions

Littéraux

Le typage des littéraux numériques ne dépend pas de l'environnement de typage : ce sont toujours des entiers ou des flottants.

$$\frac{}{\Gamma \vdash i : \text{INT}}\ (\text{CST-INT}) \qquad \frac{}{\Gamma \vdash d : \text{FLOAT}}\ (\text{CST-FLOAT})$$

Le pointeur nul, quant à lui, est compatible avec tous les types pointeur.

$$\frac{}{\Gamma \vdash \text{NULL} : t*}\ (\text{CST-NULL})$$

Enfin, le littéral unité a le type UNIT.

$$\frac{}{\Gamma \vdash () : \text{UNIT}}\ (\text{CST-UNIT})$$

Left-values

Rappelons que l'environnement de typage Γ contient le type des variables accessibles du programme. Le cas où la left-value à typer est une variable est donc direct : il suffit de retrouver son type dans l'environnement.

$$\frac{x : t \in \Gamma}{\Gamma \vdash x : t} \text{ (LV-VAR)}$$

Dans le cas d'un déréférencement, on commence par typer la left-value déréférencée. Si elle a un type pointeur, la valeur déréférencée est du type pointé.

$$\frac{\Gamma \vdash lv : t*}{\Gamma \vdash *lv : t} \text{ (LV-DEREF)}$$

Pour une left-value indexée (l'accès à tableau), on s'assure que l'indice soit entier, et que la left-value a un type tableau : le type de l'élement est encore une fois le type de base du type tableau (t pour $t[]$).

$$\frac{\Gamma \vdash e : \text{INT} \qquad \Gamma \vdash lv : t[\,]}{\Gamma \vdash lv[e] : t} \text{ (LV-INDEX)}$$

Le typage de l'accès à un champ est facilité par le fait que dans le programme, le type complet de la structure est accessible sur le champ.

Dans la définition de cette règle on utilise la notation :

$$(l, t) \in \{l_1 : t_1; \ldots; l_n : t_n\} \stackrel{\text{def}}{=\!=} \exists i \in [1; n], l = l_i \wedge t = t_i$$

$$\frac{(l, t) \in S \qquad \Gamma \vdash lv : S}{\Gamma \vdash lv.l_S : t} \text{ (LV-FIELD)}$$

Opérateurs

Un certain nombre d'opérations est possible sur le type INT.

$$\frac{\boxplus \in \{+, -, \times, /, \&, |, \wedge, \&\&, ||, \ll, \gg, \leq, \geq, <, >\} \qquad \Gamma \vdash e_1 : \text{INT} \qquad \Gamma \vdash e_2 : \text{INT}}{\Gamma \vdash e_1 \boxplus e_2 : \text{INT}} \text{ (OP-INT)}$$

De même sur FLOAT.

$$\frac{\boxplus \in \{+_., -_., \times_., /_., \leq_., \geq_., <_., >_.\} \qquad \Gamma \vdash e_1 : \text{FLOAT} \qquad \Gamma \vdash e_2 : \text{FLOAT}}{\Gamma \vdash e_1 \boxplus e_2 : \text{FLOAT}} \text{ (OP-FLOAT)}$$

Les opérateurs de comparaison peuvent s'appliquer à deux opérandes qui sont d'un type qui supporte l'égalité. Ceci est représenté par un jugement $\text{EQ}(t)$ qui est vrai pour les types

$\boxed{\text{EQ}(t)}$

$$\frac{t \in \{\text{INT}, \text{FLOAT}\}}{\text{EQ}(t)}\ (\text{EQ-NUM}) \qquad \frac{}{\text{EQ}(t*)}\ (\text{EQ-PTR}) \qquad \frac{\text{EQ}(t)}{\text{EQ}(t[\,])}\ (\text{EQ-ARRAY})$$

$$\frac{\forall i \in [1;n].\text{EQ}(t_i)}{\text{EQ}(\{l_1 : t_1; \ldots l_n : t_n\})}\ (\text{EQ-STRUCT})$$

FIGURE 5.4: Jugements d'égalité sur les types

INT, FLOAT et pointeurs, ainsi que les types composés si les types de leurs composantes (figure 5.4). Les opérateurs = et $\neq$ renvoient alors un INT :

$$\frac{\boxplus \in \{=, \neq\} \qquad \Gamma \vdash e_1 : t \qquad \Gamma \vdash e_2 : t \qquad \text{EQ}(t)}{\Gamma \vdash e_1 \boxplus e_2 : \text{INT}}\ (\text{OP-EQ})$$

Les opérateurs unaires "+" et "−" appliquent aux INT, et leurs équivalents "+." et "−." aux FLOAT.

$$\frac{\Gamma \vdash e : \text{INT}}{\Gamma \vdash +e : \text{INT}}\ (\text{UNOP-PLUS-INT}) \qquad \frac{\Gamma \vdash e : \text{FLOAT}}{\Gamma \vdash +.e : \text{FLOAT}}\ (\text{UNOP-PLUS-FLOAT})$$

$$\frac{\Gamma \vdash e : \text{INT}}{\Gamma \vdash -e : \text{INT}}\ (\text{UNOP-MINUS-INT}) \qquad \frac{\Gamma \vdash e : \text{FLOAT}}{\Gamma \vdash -.e : \text{FLOAT}}\ (\text{UNOP-MINUS-FLOAT})$$

Les opérateurs de négation unaires, en revanche, ne s'appliquent qu'aux entiers.

$$\frac{\boxminus \in \{\sim, !\} \qquad \Gamma \vdash e : \text{INT}}{\Gamma \vdash \boxminus e : \text{INT}}\ (\text{UNOP-NOT})$$

L'arithmétique de pointeurs préserve le type des pointeurs.

$$\frac{\boxplus \in \{+_p, -_p\} \qquad \Gamma \vdash e_1 : t* \qquad \Gamma \vdash e_2 : \text{INT}}{\Gamma \vdash e_1 \boxplus e_2 : t*}\ (\text{PTR-ARITH})$$

Autres expressions

Prendre l'adresse d'une left-value rend un type pointeur sur le type de celle-ci.

$$\frac{\Gamma \vdash lv : t}{\Gamma \vdash \&lv : t*}\ (\text{ADDR})$$

Pour typer une affectation, on vérifie que la left-value (à gauche) et l'expression (à droite) ont le même type. C'est alors le type résultat de l'expression d'affectation.

$$\frac{\Gamma \vdash lv : t \qquad \Gamma \vdash e : t}{\Gamma \vdash lv \leftarrow e : t} \text{ (SET)}$$

Un littéral tableau a pour type $t[]$ où t est le type de chacun de ses éléments.

$$\frac{\forall i \in [1;n], \Gamma \vdash e_i : t}{\Gamma \vdash [e_1;\ldots;e_n] : t[\,]} \text{ (ARRAY)}$$

Un littéral de structure est bien typé si ses champs sont bien typés.

$$\frac{\forall i \in [1;n], \Gamma \vdash e_i : t_i}{\Gamma \vdash \{l_1 : e_1;\ldots;l_n : e_n\} : \{l_1 : t_1;\ldots;l_n : t_n\}} \text{ (STRUCT)}$$

Pour typer un appel de fonction, on s'assure que la fonction a bien un type fonctionnel. On type alors chacun des arguments avec le type attendu. Le résultat est du type de retour de la fonction.

$$\frac{\Gamma \vdash e : (t_1,\ldots,t_n) \to t \qquad \forall i \in [1;n], \Gamma \vdash e_i : t_i}{\Gamma \vdash e(e_1,\ldots,e_n) : t} \text{ (CALL)}$$

5.4 Instructions

La séquence est simple à traiter : l'instruction vide est toujours bien typée, et la suite de deux instructions est bien typée si celles-ci le sont également.

$$\frac{}{\Gamma \vdash \text{PASS}} \text{ (PASS)} \qquad \frac{\Gamma \vdash i_1 \qquad \Gamma \vdash i_2}{\Gamma \vdash i_1; i_2} \text{ (SEQ)}$$

Une instruction constituée d'une expression est bien typée si celle-ci peut être typée dans ce même contexte.

$$\frac{\Gamma \vdash e : t}{\Gamma \vdash e} \text{ (EXP)}$$

Une déclaration de variable est bien typée si son bloc interne est bien typé quand on ajoute à l'environnement la variable avec le type de son initialiseur.

$$\frac{\Gamma \vdash e : t \qquad \Gamma, x : t \vdash i}{\Gamma \vdash \text{DECL } x = e \text{ IN}\{i\}} \text{ (DECL)}$$

Les constructions de contrôle sont bien typées si leurs sous-instructions sont bien typées, et si la condition est d'un type entier.

$$\frac{\Gamma \vdash e : \text{INT} \qquad \Gamma \vdash i_1 \qquad \Gamma \vdash i_2}{\Gamma \vdash \text{IF}(e)\{i_1\}\text{ELSE}\{i_2\}} \text{ (IF)} \qquad \frac{\Gamma \vdash e : \text{INT} \qquad \Gamma \vdash i}{\Gamma \vdash \text{WHILE}(e)\{i\}} \text{ (WHILE)}$$

$\boxed{\Gamma \vdash p \to \Gamma'}$

$$\frac{\Gamma \vdash e : t}{\Gamma \vdash e \to \Gamma} \text{ (T-EXP)} \qquad \frac{\Gamma \vdash e : t \qquad \Gamma' = (x, t), \Gamma}{\Gamma \vdash x = e \to \Gamma'} \text{ (T-VAR)}$$

FIGURE 5.5: Typage des phrases

5.5 Fonctions

Le typage des fonctions fait intervenir une variable virtuelle $\underline{R}$. Cela revient à typer l'instruction RETURN(e) comme $\underline{R} \leftarrow e$.

$$\frac{\Gamma \vdash \underline{R} \leftarrow e}{\Gamma \vdash \text{RETURN}(e)} \text{ (RETURN)}$$

Pour typer une définition de fonction, on commence par créer un nouvel environnement de typage Γ' obtenu par la suite d'opérations suivantes :

- on enlève (s'il existe) le couple $\underline{R} : t_f$ correspondant à la valeur de retour de la fonction appelante
- on ajoute les types des arguments $a_i : t_i$
- on ajoute le type de la valeur de retour de la fonction appelée, $\underline{R} : t$

Si le corps de la fonction est bien typé sous Γ', alors la fonction est typable en $(t_1, \ldots, t_n) \to t$ sous Γ.

$$\frac{\Gamma' = (\Gamma - \underline{R}), \vec{a} : \vec{t}, \underline{R} : t_r \qquad \Gamma' \vdash i}{\Gamma \vdash \text{fun}(\vec{a})\{i\} : \vec{t} \to t_r} \text{ (FUN)}$$

Cette règle utilise les notations suivantes :

$$\vec{a} \stackrel{\text{def}}{=\!=} (a_1, \ldots, a_n) \text{ où } n = |a|$$
$$\vec{a} : \vec{t} \stackrel{\text{def}}{=\!=} a_1 : t_1, \ldots a_n : t_n \text{ où } n = |a|$$

5.6 Phrases

Le typage des phrases est détaillé dans la figure 5.5. Le typage d'une expression est le cas le plus simple. En effet, il y a juste à vérifier que celle-ci est bien typable (avec ce type) dans l'environnement de départ : l'environnement n'est pas modifié. En revanche, la déclaration d'une variable globale commence de la même manière, mais on enrichit l'environnement de cette nouvelle valeur.

Type sémantique

$\tau ::=$	INT	Entier
|	FLOAT	Flottant
|	UNIT	Unité
|	$\tau *$	Pointeur
|	$\tau[\,]$	Tableau
|	$\{l_1 : \tau_1; \ldots; l_n : \tau_n\}$	Structure
|	FUN_n	Fonction

FIGURE 5.6: Types sémantiques

5.7 Sûreté du typage

Comme nous l'évoquions au début de ce chapitre, le but du typage est de rejeter certains programmes afin de ne garder que ceux qui ne provoquent pas un certain type d'erreurs à l'exécution.

Dans cette section, nous donnons des propriétés que respectent tous les programmes bien typés. Il est traditionnel de rappeler l'adage de Robin Milner :

> Well-typed programs don't go wrong.

To go wrong reste bien sûr à définir ! Cette sûreté du typage repose sur les deux théorèmes :

- progrès : si un terme est bien typé, il y a toujours une règle d'évaluation qui s'applique.
- préservation (ou *subject reduction*) : l'évaluation transforme un terme bien typé en un terme du même type.

5.8 Typage des valeurs

Puisque nous allons manipuler les propriétés statiques et dynamiques des programmes, nous allons avoir à traiter des environnements de typage Γ et des états mémoires m. La première chose à faire est donc d'établir une correspondance entre ces deux mondes.

Étant donné un état mémoire m, on associe un type de valeur (ou type sémantique) τ aux valeurs v. Cela est fait sous la forme d'un jugement $m \vDash v : \tau$.

Ces types sémantiques ne sont pas exactement les mêmes que les types statiques. Pour les calculer, on n'a pas accès au code du programme, seulement à ses données. Il est par exemple possible de reconnaître le type des constantes, mais pas celui des fonctions. Celles-ci sont en fait le seul cas qu'il est impossible de déterminer statiquement. On le remplace donc par un cas plus simple où seul l'arité est conservée (figure 5.6).

Les règles sont détaillées dans la figure 5.7 : les types des constantes sont simples à retrouver car il y a assez d'information en mémoire. Pour les références, ce qui peut être déréférencé en une valeur de type τ est un $\tau *$. Le typage des valeurs composées se fait en profondeur. Enfin, la seule information restant à l'exécution sur les fonctions est son arité.

La prochaine étape est de définir une relation de compatibilité entre les types sémantiques τ et statiques t. Nous noterons ceci sous la forme d'un jugement $\text{Comp}(\tau, t)$. Les règles sont décrites dans la figure 5.8, la règle importante étant COMP-FUN.

Grâce à ce jugement, on peut donner la définition suivante.

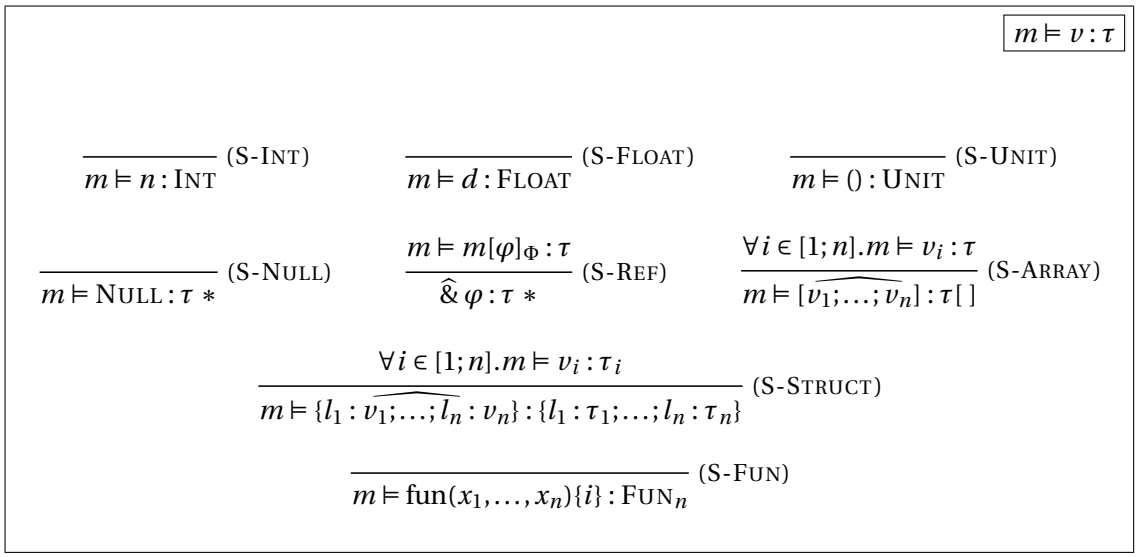


FIGURE 5.7: Règles de typage sémantique

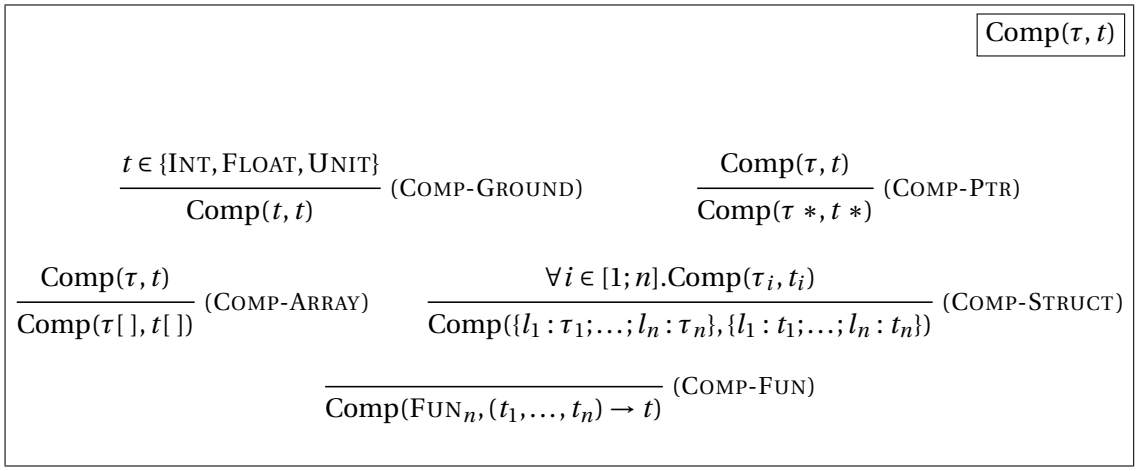


FIGURE 5.8: Compatibilité entre types sémantiques et statiques

Définition 5.1 (État mémoire bien typé)**.** *On dit qu'un état mémoire m est bien typé sous un environnement Γ, ce que l'on note $\Gamma \vDash m$, si les types sémantiques des variables visibles coïncident avec leurs types statiques.*

$$\Gamma \vDash m \stackrel{\text{def}}{=\!=} \forall (x \mapsto v) \in \text{Visible}(m).\exists \tau, t. \begin{cases} \Gamma \vdash x : t \\ m \vDash v : \tau \\ \text{Comp}(\tau, t) \end{cases}$$

Visible(m) *désigne l'ensemble des couples d'associations (variable, valeur) du cadre de pile le plus récents ainsi que des variables globales. Cela correspond aussi aux variables présentes dans l'environnement de typage.*

5.9 Progrès et préservation

On commence par énoncer quelques lemmes utiles dans la démonstration de ces théorèmes.

Les règles précédentes ont la particularité suivante : pour chaque forme syntaxique, il n'y a souvent qu'une règle qui peut s'appliquer. Cela permet de déduire quelle règle il faut appliquer pour vérifier (ou inférer) le type d'une expression.

Lemme 5.1 (Inversion)**.** *À partir d'un jugement de typage, on peut en déduire des informations sur les types de ses sous-expressions.*

- *Constantes*
  - *si* $\Gamma \vdash n : t$, *alors* $t = \text{Int}$
  - *si* $\Gamma \vdash d : t$, *alors* $t = \text{Float}$
  - *si* $\Gamma \vdash \text{Null} : t$, *alors* $\exists t', t = t'*$ *textsc passe pas*
  - *si* $\Gamma \vdash () : t$, *alors* $t = \text{Unit}$
- *Références mémoire :*
  - *si* $\Gamma \vdash x : t$, *alors* $x : t \in \Gamma$
  - *si* $\Gamma \vdash *lv : t$, *alors* $\Gamma \vdash lv : t*$
  - *si* $\Gamma \vdash lv[e] : t$, *alors* $\Gamma \vdash lv : t[]$ *et* $\Gamma \vdash e : \text{Int}$
  - *si* $\Gamma \vdash lv.l_S : t$, *alors* $\Gamma \vdash lv : S$
- *Opérations :*
  - *si* $\Gamma \vdash \boxminus\, e : t$, *alors on est dans un des cas suivants :*
    - $\boxminus \in \{+, -, \sim, !\}$, $t = \text{Int}$, $\Gamma \vdash e : \text{Int}$
    - $\boxminus \in \{+., -.\}$, $t = \text{Float}$, $\Gamma \vdash e : \text{Float}$
  - *si* $\Gamma \vdash e_1 \boxplus e_2 : t$, *un des cas suivants se présente :*
    - $\boxplus \in \{+, -, \times, /, \&, |, \hat{}, \&\&, ||, \ll, \gg, \leq, \geq, <, >\}$, $\Gamma \vdash e_1 : \text{Int}$, $\Gamma \vdash e_2 : \text{Int}$, $t = \text{Int}$
    - $\boxplus \in \{+., -., \times., /., \leq., \geq., <., >.\}$, $\Gamma \vdash e_1 : \text{Float}$, $\Gamma \vdash e_2 : \text{Float}$, $t = \text{Float}$
    - $\boxplus \in \{=, \neq\}$, $\Gamma \vdash e_1 : t'$, $\Gamma \vdash e_2 : t'$, $\text{Eq}(t')$, $t = \text{Int}$
    - $\boxplus \in \{\leq, \geq, <, >\}$, $t = \text{Int}$, $\Gamma \vdash e_1 : t'$, $\Gamma \vdash e_2 : t'$, $t' \in \{\text{Int}, \text{Float}\}$
    - $\boxplus \in \{+_p, -_p\}$, $\exists t', t = t'*$, $\Gamma \vdash e_1 : t'*$, $\Gamma \vdash e_2 : \text{Int}$
- *Appel de fonction : si* $\Gamma \vdash e(e_1, \ldots, e_n) : t$, *il existe* $(t_1, \ldots, t_n)$ *tels que*

$$\begin{cases} \Gamma \vdash e : (t_1, \ldots, t_n) \to t \\ \forall i \in [1; n], \Gamma \vdash e_i : t_i \end{cases}$$

- *Fonction : si* $\Gamma \vdash \text{fun}(a_1, \ldots, a_n)\{i\} : t$, *alors il existe* $(t_1, \ldots, t_n)$ *et* t' *tels que* $t' = (t_1, \ldots, t_n) \to t$.

Démonstration. Pour chaque jugement, on considère les règles qui peuvent amener à cette conclusion.

- Références mémoire :
  - $\Gamma \vdash x : t$
    La seule règle de cette forme est Lv-Var. Puisque sa prémisse est vraie, on en conclut que $x : t \in \Gamma$.
  - $\Gamma \vdash *\varphi : t$
    Idem avec Lv-Deref.

- $\Gamma \vdash \varphi[] : t$
  Idem avec LV-INDEX.
- $\Gamma \vdash \varphi.l : t$
  Idem avec LV-FIELD.

- Appel de fonction : pour en arriver à $\Gamma \vdash e(e_1, \ldots, e_n) : t$, seule la règle CALL s'applique, ce qui permet de conclure.
- Fonction : la seule règle possible pour conclure une dérivation de $\Gamma \vdash \text{fun}(a_1, \ldots, a_n)\{i\} : t$ est FUN.

□

Il est aussi possible de réaliser l'opération inverse : à partir du type d'une valeur, on peut déterminer sa forme syntaxique. C'est bien sûr uniquement possible pour les valeurs, pas pour n'importe quelle expression (par exemple l'expression x (variable) peut avoir n'importe quel type t dans le contexte $\Gamma = x : t$).

Lemme 5.2 (Formes canoniques)**.** *Il est possible de déterminer la forme syntaxique d'une valeur étant donné son type, comme décrit dans le tableau suivant. Par exemple, d'après la première ligne, si* $\Gamma \vdash v : \text{Int}$*, alors* $\exists n, v = d$*.*

Type de v	*Forme de v*
Int	n
Float	d
Unit	()
$t*$	φ *ou* Null
$t[\,]$	$[v_1; \ldots; v_n]$
$\{l_1 : t_1; \ldots; l_n : t_n\}$	$\{l_1 : v_1; \ldots; l_n : v_n\}$
$(t_1, \ldots, t_n) \to t$	$\text{fun}(a_1, \ldots, a_n)\{i\}$

Ces lemmes étant établis, on énonce maintenant le théorème de progrès.

Théorème 5.1 (Progrès)**.** *Supposons que* $\Gamma \vdash e : t$*. Soit* m *un état mémoire tel que* $\Gamma \vDash m$*. Alors l'un des cas suivant est vrai :*

- $\exists v \neq \Omega, e = v$
- $\exists (e', m'), \Gamma \vDash m' \wedge \langle e, m\rangle \to \langle e', m'\rangle$
- $\exists \Omega \in \{\Omega_{div}, \Omega_{array}, \Omega_{ptr}\}, \langle e, m\rangle \to \Omega$

C'est-à-dire, soit :

- *e est complètement évaluée*
- *un pas d'évaluation préservant la compatibilité mémoire est possible*
- *une erreur de division, tableau ou pointeur se produit*

Théorème 5.2 (Progrès pour les left-values)**.** *Si* $\langle l\, v, m\rangle \to^* \langle v, m'\rangle$*,* v *est de la forme* φ *(référence mémoire) ou* Ω *(erreur).*

La preuve des théorèmes 5.1 et 5.2 se trouve en annexe D.2.

Théorème 5.3 (Préservation)**.** *Soient* Γ *un environnement de typage, e une expression et m un état mémoire.*

On suppose que les hypothèse suivantes sont vérifiées :

- $\Gamma \vdash e : t$
- $\Gamma \vDash m$
- $\langle e, m \rangle \rightarrow \langle e', m' \rangle$

Alors $\Gamma \vdash e' : t$ *et* $\Gamma \vDash m'$.
Autrement dit, si une expression est typable, alors un pas d'évaluation ne modifie pas son type et préserve le typage de la mémoire.

La preuve de ce théorème se trouve en annexe D.3.

Cela prouve qu'aucun terme ne reste "bloqué" parce qu'aucune règle ne s'applique, et que la sémantique respecte le typage. En quelque sorte, les types sont un contrat entre les expressions et les fonctions : si leur évaluation converge, alors une valeur du type inféré sera produite.

CHAPITRE 6

QUALIFICATEURS DE TYPE

Jusqu'ici SAFESPEAK est un langage impératif généraliste, ne prenant pas en compte les spécificités de l'adressage utilisé dans les systèmes d'exploitation.

Dans ce chapitre, on commence par l'étendre en ajoutant des constructions modélisant les variables présentes dans l'espace utilisateur (cf. chapitre 2). Pour accéder à celles-ci, on ajoute un opérateur de déréférencement sûr qui vérifie à l'exécution que l'invariant suivant est respecté :

> Les pointeurs dont la valeur est contrôlée par l'utilisateur, pointent vers l'espace utilisateur.

La terminologie mérite d'être détaillée :

Un pointeur contrôlé par l'utilisateur, ou *pointeur utilisateur*, est référence mémoire dont la valeur est modifiable par le code utilisateur (opposé au code noyau, que nous analysons ici). Ceci correspond à des données provenant de l'extérieur du système vérifié. Pour faire un parallèle avec la programmation réseau, cela correspond à des variables dont la valeur proviendrait d'un client distant. C'est une propriété statique, qui peut être déterminée à la compilation à partir de considérations syntaxiques.

Un pointeur pointant vers l'espace utilisateur fait référence à une variable allouée en espace utilisateur. Cela veut dire qu'y accéder ne risque pas de mettre en péril l'isolation du noyau en faisant fuiter des informations confidentielles ou en déjouant son intégrité. Cette propriété est dynamique : un pointeur utilisateur peut *a priori* pointer vers l'espace utilisateur, ou non.

Pour reprendre le parallèle avec la programmation web, une chaîne provenant du réseau peut être correctement échappée ou non. Pour s'assurer de la sécurité du traitement d'une telle chaîne, il faut s'assurer que celles-ci sont correctement échappées.

Pour prouver que cette approche est bien fondée, on procède en plusieurs étapes.

Tout d'abord, on définit un nouveau type d'erreur Ω_{taint}, déclenché lorsqu'un pointeur contrôlé par l'utilisateur et pointant vers le noyau est déréférencé (le cas que l'on cherche à éviter) [1].

Ensuite, on montre qu'avec cet ajout, si on étend naïvement le système de types en donnant le même type aux pointeurs contrôlés par l'utilisateur et le noyau, le théorème de progrès (5.1) n'est plus valable.

L'étape suivante est d'étendre, à son tour, le système de types de SAFESPEAK en distinguant les types des pointeurs contrôlés par l'utilisateur des pointeurs contrôlés par le noyau.

1. Il est important de noter que ce cas d'erreur est "virtuel", la distinction entre pointeur contrôlé par l'utilisateur et par le noyau n'étant normalement pas visible à l'exécution.

```
sys_getver = fun(p){
        DECL v = { major : 3; minor : 14; patch : 15 } IN
        copy_to_user(p, &v)
}
```

FIGURE 6.1: Implantation d'un appel système qui remplit une structure par pointeur

Puisqu'on veut interdire le déréférencement des premiers par l'opérateur $*$, on modifie également le typage de celui-ci.

Enfin, une fois ces modifications faites, on prouve que les propriétés de progrès et de préservation sont rétablies.

6.1 Extensions noyau pour SAFESPEAK

On ajoute à SAFESPEAK la notion de valeur provenant de l'espace utilisateur. Pour marquer la séparation entre les deux espaces d'adressage, on ajoute une construction $\varphi ::= \widehat{\diamondsuit}\,\varphi'$. Le chemin interne φ' désigne une variable classique (un pointeur noyau) et l'opérateur $\widehat{\diamondsuit}\,\cdot$ permet de l'interpréter comme un pointeur vers l'espace utilisateur, contrairement à $\widehat{\&}\cdot$ qui crée des pointeurs vers l'espace noyau. En quelque sorte, on ne classifie pas les valeur selon la variable pointée mais selon la construction du pointeur.

En plus du déréférencement par $*$ (qui devra donc renvoyer Ω_{taint} pour les valeurs de la forme $\widehat{\diamondsuit}\,\varphi'$), il faut aussi ajouter des constructions de lecture et d'écriture à travers les pointeurs utilisateurs. Ceci sera fait sous forme de deux fonctions, copy_from_user et copy_to_user. Celles-ci prennent deux pointeurs en paramètre et renvoient un booléen indiquant si la copie pu être faite (si le paramètre contrôlé par l'utilisateur pointe en espace noyau, les fonctions ne font pas la copie et signalent l'erreur).

Illustrons ceci par un exemple. Imaginons un appel système fictif qui renvoie la version du noyau, en remplissant par pointeur une structure contenant les champs entiers `major`, `minor` et `patch` (un équivalent dans Linux est l'appel système `uname()`). Celui-ci peut être alors écrit comme dans la figure 6.1. Une fois la structure noyau v remplie, il faut la copier vers l'espace utilisateur. La fonction copy_to_user va réaliser cette copie (de la même manière qu'avec un `memcpy()`, mais après avoir testé dynamiquement que p pointe en espace utilisateur (dans le cas contraire, la copie n'est pas faite).

On commence donc par ajouter aux instructions des constructions copy_to_user$(\cdot,\cdot)$ et copy_from_user$(\cdot,\cdot)$ de copie sûre. Afin de leur donner une sémantique, il faut étendre l'ensemble des valeurs pointeur φ aux constructions de la forme $\widehat{\diamondsuit}\,\varphi'$. Pour créer des termes s'évaluant en de telles valeurs, il faut une construction syntaxique $\diamondsuit\ e$ telle que si e s'évalue en φ, $\diamondsuit\ e$ s'évalue en $\widehat{\diamondsuit}\,\varphi$. Cela correspond à 2 ajouts : une un nouveau contexte d'évaluation $\diamondsuit\ \bullet$ et une règle d'évaluation. Enfin, on ajoute une nouvelle erreur Ω_{taint} à déclencher lorsqu'on déréférence directement un pointeur utilisateur. Ces étapes sont résumées dans la figure 6.2.

6.2 Extensions sémantiques

En ce qui concerne l'évaluation des expressions $\diamondsuit\ \cdot$, on ajoute la règle suivante :

Expressions	$e ::= \ldots$	
	$\mid \ \diamondsuit\, e$	Expression souillée
Contextes	$C ::= \ldots$	
	$\mid \ \diamondsuit\, C$	
Chemins	$\varphi ::= \ldots$	
	$\mid \ \widehat{\diamondsuit}\, \varphi$	Valeur souillée
Erreurs	$\Omega ::= \ldots$	
	$\mid \ \Omega_{taint}$	Erreur de souillure

FIGURE 6.2: Ajouts liés aux pointeurs utilisateurs (par rapport à l'évaluateur du chapitre 4)

$$\frac{}{\langle \diamondsuit\, \varphi, m\rangle \rightarrow \langle \widehat{\diamondsuit}\, \varphi, m\rangle}\ (\text{EXPR-TAINTED})$$

Ensuite, il est nécessaire d'adapter les règles d'accès à la mémoire pour déclencher une erreur Ω_{taint} en cas de déréférencement d'un pointeur utilisateur. Les accès mémoire en lecture proviennent de la règle EXP-LV et ceux en lecture, de la règle EXP-SET, rappellées ici :

$$\frac{}{\langle \varphi, m\rangle \rightarrow \langle m[\varphi]_\Phi, m\rangle}\ (\text{EXP-LV}) \qquad \frac{}{\langle \varphi \leftarrow v, m\rangle \rightarrow \langle v, m[\varphi \leftarrow v]_\Phi\rangle}\ (\text{EXP-SET})$$

Les accès à la mémoire sont en effet faits par le biais de la lentille Φ. Il suffit donc d'adapter la définition 4.10 de celle-ci en rajoutant le cas :

$$\Phi(\widehat{\diamondsuit}\, \varphi) = \Omega_{taint}$$

Enfin, il est nécessaire de donner une sémantique aux fonctions copy_from_user et copy_to_user. L'idée est que celles-ci testent dynamiquement la *valeur* du paramètre contrôlé par l'utilisateur afin de vérifier que celui-ci pointe vers l'espace utilisateur (c'est-à-dire, qu'il est de la forme $\widehat{\diamondsuit}\, \varphi$).

Deux cas peuvent se produire. Si ce test est vrai, alors la copie est faite et l'opération de copie retourne la valeur entière 0. Dans le cas contraire, aucune copie n'est faite et la valeur -1 est retournée. Ce comportement est calé celui des fonctions `copy_{from,to}_user` du noyau Linux : en cas de succès elles renvoient 0, et en cas d'erreur `-EFAULT`.

$$\frac{\langle \varphi_d \leftarrow \varphi_s, m\rangle \rightarrow \langle \textsc{Pass}, m'\rangle}{\langle \text{copy\_from\_user}(\varphi_d, \widehat{\diamondsuit}\ \varphi_s), m\rangle \rightarrow \langle 0, m'\rangle}\ (\textsc{User-Get-OK})$$

$$\frac{\nexists\, \varphi_s, \varphi = \widehat{\diamondsuit}\ \varphi_s}{\langle \text{copy\_from\_user}(\varphi_d, \varphi), m\rangle \rightarrow \langle -1, m\rangle}\ (\textsc{User-Get-Err})$$

$$\frac{\langle \varphi_d \leftarrow \varphi_s, m\rangle \rightarrow \langle \textsc{Pass}, m'\rangle}{\langle \text{copy\_to\_user}(\widehat{\diamondsuit}\ \varphi_d, \varphi_s), m\rangle \rightarrow \langle 0, m'\rangle}\ (\textsc{User-Put-OK})$$

$$\frac{\nexists\, \varphi_d, \varphi = \widehat{\diamondsuit}\ \varphi_d}{\langle \text{copy\_to\_user}(\varphi, \varphi_s), m\rangle \rightarrow \langle -1, m\rangle}\ (\textsc{User-Put-Err})$$

Ces règles sont à appliquer en priorité de la règle d'appel de fonction classique, puisque ces fonctions n'ont pas de corps en tant que tel. En effet celles ci ne sont pas implantables directement en SAFESPEAK, puisqu'il n'y a pas par exemple d'opérateur permettant d'extraire φ depuis une valeur $\widehat{\diamondsuit}\ \varphi$. L'opération en "boîte noire" de ces deux fonctions permet d'assurer que l'accès à l'espace utilisateur est toujours couplé à un test dynamique.

6.3 Insuffisance des types simples

Étant donné SAFESPEAK augmenté de cette extension sémantique, on peut étendre trivialement le système de types avec la règle suivante :

$$\frac{\Gamma \vdash e : t*}{\Gamma \vdash \diamondsuit\ e : t*}\ (\textsc{Taint-Ignore})$$

Cette règle est compatible avec l'extension, sauf qu'elle introduit des termes qui sont bien typables mais dont l'évaluation provoque une erreur autre que Ω_{div}, Ω_{array} ou Ω_{ptr}, violant ainsi le théorème 5.1.

Par exemple, supposons que x soit une variable globale entière, et posons :

$$\begin{cases} e = *\diamondsuit\ \&x \\ \Gamma = x : \textsc{Int} \\ m = ([[x \mapsto 0]], [\,]) \end{cases}$$

Les hypothèses du théorème de progrès sont bien vérifiées, mais cependant la conclusion n'est pas vraie :

- On a bien $\Gamma \vDash m$.
- e est bien typée sous Γ :

$$\dfrac{\dfrac{\dfrac{\dfrac{x : \textsc{Int} \in \Gamma}{\Gamma \vdash x : \textsc{Int}}\ (\textsc{Lv-Var})}{\Gamma \vdash \&x : \textsc{Int}*}\ (\textsc{Lv-Deref})}{\Gamma \vdash \diamondsuit\ \&x : \textsc{Int}*}\ (\textsc{Taint-Ignore})}{\Gamma \vdash *\diamondsuit\ \&x : \textsc{Int}}\ (\textsc{Lv-Deref})$$

Type	$t ::= \ldots$	
	$\mid \; t\,@$	Pointeur utilisateur
Type sémantique	$\tau ::= \ldots$	
	$\mid \; t\,@$	Pointeur utilisateur

FIGURE 6.3: Ajouts liés aux qualificateurs de types (par rapport aux figures 5.2 et 5.6)

- L'évaluation de e sous m provoque une erreur différente de Ω_{div}, Ω_{array}, ou Ω_{ptr} :

$$\dfrac{\dfrac{\overline{m[*\widehat{\diamondsuit}\ x] = \Omega_{taint}}}{\langle *\diamondsuit\ \&x, m\rangle \to \langle \Omega_{taint}, m\rangle}\ (\text{EXP-LV}) \qquad \dfrac{}{\langle \Omega_{taint}, m\rangle \to \Omega_{taint}}\ (\text{EVAL-ERR})}{\langle e, m\rangle \to \Omega_{taint}}$$

Cela montre que le typage n'apporte plus de garantie de sûreté sur l'exécution : le système de types naïvement étendu par une règle comme TAINT-IGNORE n'est pas en adéquation avec les extensions présentées dans la section 6.1.

6.4 Extensions du système de types

On présente ici un système de types plus expressif permettant de capturer les extensions de sémantique. *In fine*, cela permettra de prouver le théorème 6.1 qui est l'équivalent du théorème 5.1 mais pour le nouveau jugement de typage.

Définir un nouveau système de types revient à étendre le jugement de typage $\cdot \vdash \cdot : \cdot$, en modifiant certaines règles et en en ajoutant d'autres. Naturellement, la plupart des différences porteront sur le traitement des pointeurs.

Pointeurs utilisateurs

Le changement clef est l'ajout de *pointeurs utilisateurs*. En plus des types pointeurs $t\ *$, on ajoute des types pointeurs $t\ @$. La différence entre les deux représente *qui* contrôle leur valeur (section 2.5).

Les différences sont les suivantes :

- Les types "$t\ *$" s'appliquent aux pointeurs contrôlés par le noyau. Par exemple, prendre l'adresse d'un objet de la pile noyau donne un pointeur noyau.
- Les types "$t\ @$", quant à eux, s'appliquent aux pointeurs qui proviennent de l'espace utilisateur. Ces pointeurs proviennent toujours d'interfaces particulières, comme les appels système ou les paramètres passés aux implantations de la fonction `ioctl`.

Cet ajout est précisé dans la figure 6.3.

Puisqu'on s'intéresse à la provenance des pointeurs, détaillons les règles qui créent, manipulent et utilisent des pointeurs.

Sources de pointeurs

La source principale de pointeurs est l'opérateur & qui prend l'adresse d'une variable. Celle-ci est bien entendue contrôlée par le noyau (dans le sens où son déréférencement est toujours sûr). Cette construction crée donc des pointeurs noyau :

$$\frac{\Gamma \vdash lv : t}{\Gamma \vdash \&lv : t\,*}\ (\textsc{AddrOf-Kernel})$$

Manipulations de pointeurs

L'avantage du typage est que celui-ci suit le flot de données : si à un endroit une valeur de type t est affectée à une variable, que le contenu de cette variable est placé puis retiré d'une structure de données, il conserve ce type t. En particulier un pointeur utilisateur reste un pointeur utilisateur.

Une seule règle consomme un pointeur et en retourne un. Elle concerne l'arithmétique des pointeurs. On l'étend aux pointeurs utilisateur :

$$\frac{\boxplus \in \{+_p, -_p\} \qquad \Gamma \vdash e_1 : t\,@ \qquad \Gamma \vdash e_2 : \textsc{Int}}{\Gamma \vdash e_1 \boxplus e_2 : t\,@}\ (\textsc{Ptr-Arith-User})$$

Utilisations de pointeurs

La principale restriction, au cœur de ce chapitre et de cette thèse, est que seuls les pointeurs noyau peuvent être déréférencés de manière sûre. La règle suivante remplace donc Lv-Deref :

$$\frac{\Gamma \vdash e : t\,*}{\Gamma \vdash *e : t}\ (\textsc{Lv-Deref-Kernel})$$

Ainsi, on interdit le déréférencement des expressions de type $t\,@$ à la compilation.

L'opérateur $\diamondsuit\,\cdot$ transforme un pointeur selon la règle suivante :

$$\frac{\Gamma \vdash e : t\,*}{\Gamma \vdash \diamondsuit\, e : t\,@}\ (\textsc{Taint})$$

Les fonctions copy_from_user et copy_to_user sont typées de la manière suivante :

$$\frac{}{\Gamma \vdash \text{copy\_from\_user} : (t\,*, t\,@) \to \textsc{Int}}\ (\textsc{GetU}) \qquad \frac{}{\Gamma \vdash \text{copy\_to\_user} : (t\,@, t\,*) \to \textsc{Int}}\ (\textsc{PutU})$$

Il est à noter qu'il n'y a pas de sous-typage : les pointeurs noyau ne peuvent être utilisés qu'en tant que pointeurs noyau, et les pointeurs utilisateurs qu'en tant que pointeurs utilisateurs.

6.5 Sûreté du typage

Typage sémantique

La définition du typage sémantique doit aussi être étendue au cas $\varphi = \widehat{\Diamond}\ \varphi'$.

$$\frac{m \vDash \varphi' : \tau\ *}{m \vDash \widehat{\Diamond}\ \varphi' : \tau\ @}\ (\text{S-TAINTED}) \qquad \frac{\text{Comp}(\tau, t)}{\text{Comp}(\tau\ @, t\ @)}\ (\text{COMP-PTR})$$

Progrès et préservation

Le déréférencement d'un pointeur dont la valeur est contrôlée par l'utilisateur ne peut se faire qu'à travers une fonction qui vérifie la sûreté de celui-ci.

Théorème 6.1 (Progrès pour les types qualifiés)**.** *Supposons que* $\Gamma \vdash e : t$. *Soit* m *un état mémoire tel que* $\Gamma \vDash m$. *Alors l'un des cas suivants est vrai :*

- $\exists v \neq \Omega, e = v$
- $\exists (e', m'), \Gamma \vDash m' \wedge \langle e, m\rangle \rightarrow \langle e', m'\rangle$
- $\exists \Omega \in \{\Omega_{div}, \Omega_{array}, \Omega_{ptr}\}, \langle e, m\rangle \rightarrow \Omega$

La preuve de ce théorème est en annexe D.4.
Et nous donnons un équivalent du théorème 5.3.

Théorème 6.2 (Préservation pour les types qualifiés)**.** *Si une expression est typable et que son évaluation produit une valeur, alors cette valeur est du même type que l'expression.*
Si $\Gamma \vdash e : t$ *et* $e \rightarrow v$
alors $\Gamma \vdash v : t$.

La preuve de ce théorème est en annexe D.5.
La sûreté du typage étant à nouveau établie, on a montré que les qualificateurs de types sur les pointeurs suffisent pour avoir une adéquation entre les extensions de sémantique de la section 6.1 et les extensions du système de type de la section 6.4.

Conclusion de la partie II

L'idée derrière CQual se rapprochant de la nôtre, et cet outil ayant été appliqué aux types de vulnérabilités qui nous intéressent [JW04], il est utile de détailler pourquoi notre approche est différente de ces travaux.

Tout d'abord, CQual modifie fondamentalement l'ensemble du système de types (via le jugement de typage) et insère des qualificateurs à tous les niveaux de types. Au contraire, nous ne modifions le système de types que là où cela est nécessaire, c'est-à-dire sur les types pointeurs.

D'autre part, le système de types de CQual fait intervenir de manière fondamentale une relation de sous-typage. Le problème de déréférencement des pointeurs utilisateurs peut être traité dans ce cadre en posant Kernel $\preceq$ User pour restreindre certaines opérations aux pointeurs Kernel.

Notre approche, au contraire, n'utilise pas de sous-typage, mais consiste à définir un type abstrait t @ partageant certaines propriétés avec t $*$ (comme la taille et la représentation) mais incompatible avec certaines opérations. C'est à rapprocher du système de masquage par interface des langages ML, ou encore de la définition de types abstraits par le mot-clef `newtype` en Haskell.

En C, il est notamment commun d'utiliser des `int` pour tout et n'importe quoi : pour des entiers bien sûr (au sens de $\mathbb{Z}$), mais aussi comme identificateurs pour lesquels les opérations usuelles comme l'addition n'ont pas de sens. Par exemple, sous Linux, l'opération d'ouverture de fichier renvoie un entier, dit *descripteur de fichier*, qui identifie ce fichier pour ce processus. Il est passé à toutes les fonctions de lecture et d'écriture. Un autre idiome est d'utiliser des entiers comme des ensembles de bits (on parle de *bitmask*). Par exemple, en ouvrant un fichier on précise le mode du fichier (lecture, écriture ou les deux) par les bits 1 et 2, s'il faut créer le fichier ou non s'il n'existe pas par le bit 7, s'il dans ce cas il doit être effacé par le bit 8), etc. On obtient un mode en réalisant un "ou" bit à bit entre des constantes. Ces deux utilisations du type `int` n'ont rien à voir ; il faudrait donc empêcher d'utiliser un descripteur de fichier comme un mode, et vice-versa. De même, aucun opérateur n'a de sens sur les descripteurs de fichier, mais l'opérateur `|` du "ou" bit à bit doit rester possible pour les modes.

Dans le cas des pointeurs, même si le noyau Linux (et la plupart des systèmes d'exploitation) ne comportent que deux espaces d'adressage, il est commun dans les systèmes embarqués de manipuler des pointeurs provenant d'espaces mémoire indépendants : par exemple, de la de la mémoire flash, de la RAM, ou une EEPROM de configuration. Ces différentes mémoires possèdent des adresses, et un pointeur est interprété comme faisant référence à une ou l'autre selon le code dont il est tiré. Dans ces cas, la notion de sous-typage n'est pas adaptée.

Troisième partie

Expérimentation

On décrit ici la démarche expérimentale liée à l'implémentation des analyses décrites dans la partie II.

Le chapitre 7 décrit l'implémentation en elle-même : comment le code source C est compilé vers SAFESPEAK, et comment les types du programme sont vérifiés.

Ensuite, dans le chapitre 8, le cas d'un bogue de pilote graphique dans le noyau Linux est étudié. On montre que les analyses précedentes permettent de distinguer statiquement entre le cas incorrect et le cas corrigé.

Enfin, le chapitre 9 conclut : les limitations de cette approche sont présentées, ainsi qu'un résumé des contributions de cet ouvrage.

CHAPITRE 7

IMPLANTATION

Dans ce chapitre, nous décrivons la mise en œuvre des analyses statiques précédentes. Nous commençons par un tour d'horizon des représentations intermédiaires possibles, avant de décrire celle retenue : Newspeak. La chaîne de compilation est explicitée, partant de C pour aller au langage impératif décrit dans le chapitre 4. Enfin, nous donnons les détails d'un algorithme d'inférence de types à la Hindley-Milner, reposant sur l'unification et le partage de références.

7.1 Newspeak

Newspeak [HL08] est un langage conçu pour être à la fois :

- Précis : sa sémantique est définie formellement dans [HL08]
- Expressif : la plupart des primitives présentes dans les langages de bas niveau sont compilables en Newspeak.
- Simple : peu de primitives sont présentes.
- Minimal : aucun élément syntaxique ne peut être exprimé comme combinaison d'autres.
- Explicite : les constructions sont toutes indépendantes du contexte.
- Orienté analyse : les primitives sont décorées d'informations reflétant leur validité (tests de bornes, etc)
- Indépendant de l'architecture : toutes les caractéristiques comme la taille des types ou l'alignement des structures sont rendues explicites.

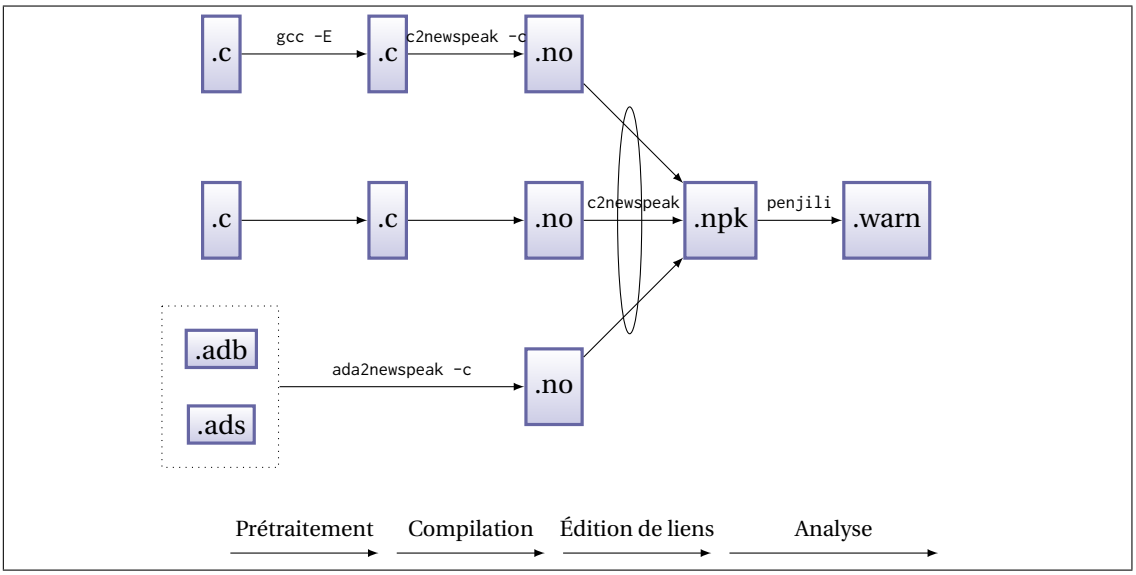


FIGURE 7.1: Compilation depuis Newspeak

7.2 Chaîne de compilation

La compilation vers C est faite en trois étapes (figure 7.1) : prétraitement du code source, compilation de C prétraité vers NEWSPEAK, puis compilation de NEWSPEAK vers ce langage.

La première étape consiste à prétraiter les fichiers C source avec le logiciel `cpp`, comme pour une compilation normale. Cette étape interprète les directives comme `#include`, `#ifdef`, À cet étape, les commentaires sont aussi supprimés.

Une fois cette passe effectuée, le résultats est un ensemble de fichiers C prétraités, c'est-à-dire des unités de compilation.

Puisque la directive `#include` est textuelle, ces fichiers sont très grands et donnent lieu à beaucoup de duplication dans les passes suivantes.

À ce niveau, les fichiers sont passés à l'outil `c2newspeak` qui les traduit vers Newspeak. Dans cette étape, les types et les noms sont résolus, et le programme est annoté de manière à rendre les prochaines étapes indépendante du contexte. Par exemple, chaque déclaration de variable est adjointe d'une description complète du type.

Lors de cette étape, le flôt de contrôle est également simplifié. C en effet propose de nombreuses constructions ambigües ou redondantes.

Au contraire, Newspeak propose un nombre réduit de constructions. Rappelons que le but de ce langage est de faciliter l'analyse statique : des constructions orthogonales permettent donc d'éviter la duplication de règles sémantique, ou de code lors de l'implémentation d'un analyseur.

Par exemple, plutôt que de fournir une boucle *while*, une boucle *do/while* et une boucle for, Newspeak fournit une unique boucle WHILE(1){}. La sortie de boucle est compilée vers un GOTO [EH94], qui est toujours un saut vers l'avant (similaire à un "break" généralisé).

La sémantique de Newspeak et la traduction de C vers Newspeak sont décrites dans [HL08].

Newspeak est conçu pour l'analyse statique par interprétation abstraite. Il a donc une vue de bas niveau sur les programmes. Par exemple, aucune distinction n'est faite entre l'accès à un champ et l'accès un l'élément d'un tableau (tous deux sont traduits par un décalage numérique depuis le début de la zone mémoire). Pour supprimer cette ambiguïté, il faut s'interfacer dans les structures internes de `c2newspeak`, où les informations nécessaires sont encore présentes.

```
int x;
int *p = &x;
x = 0;
```

FIGURE 7.2: Compilation d'un programme C – avant

Ensuite, les différents fichiers sont liés ensemble. Cet étape consiste principalement à s'assurer que les hypothèses faites par les différentes unités de compilation sont cohérentes entre elles. Les objets marqués `static`, invisibles à l'extérieur de leur unité de compilation, sont renommés afin qu'ils aient un nom unique.

Enfin, l'implantation d'un algorithme d'inférence pour les systèmes de types décrits dans les chapitres 5 et 6 est assez simple. Puisqu'ils sont suffisament proches du lambda calcul simplement typé, on peut utiliser une variante de l'algorithme W de Damas et Milner [DM82]. On utilise l'optimisation classique qui consiste à se reposer sur le partage de références pour réaliser l'unification, plutôt que de faire des substitutions explicites. Puisque ces systèmes de types sont monomorphes, on présente une erreur si des variable de type libres sont présentes.

À la fin de cette étape, on obtient soit un programme complètement annoté, soit une erreur de type.

Prenons l'exemple de la figure 7.2 et typons-le "à la main". On commence par oublier toutes les étiquettes de type présentes dans le programme. Celui-ci devient alors :

```
var x, p;
p = &x;
x = 0;
```

La premiere ligne introduit deux variables. On peut noter leurs types respectifs (inconnus pour le moment) t_1 et t_2. La première affectation `p = &x` implique que les deux côtés du signe "=" ont le même type. À gauche, le type est t_2, et à droite $Ptr(t_1)$. On applique le même raisonnement à la seconde affectation : à gauche, le type est t_1 et à droite Int. On en déduit que le type de x est Int et celui de p est Ptr(Int).

```
type var_type =
  | Unknown of int
  | Instanciated of ml_type

and const_type =
  | Int_type
  | Float_type

and ml_type =
  | Var_type of var_type ref
  | Const_type of const_type
  | Pair_type of ml_type * ml_type
  | Fun_type of ml_type * ml_type
```

Pour implanter cet algorithme, on représente les types de données du programmes à typer par une valeur de type `ml_type`. En plus des constantes de types comme int ou float, et des constructeurs de type comme pair et fun, le constructeur Var permet d'exprimer les variables de types (inconnues ou non).

Celles-ci sont numérotées par un int, on suppose avoir à disposition deux fonctions manipulant un compteur global d'inconnues.

```
module Counter : sig
  val reset_unknowns : unit -> unit
  val new_unknown : unit -> int
end
```

De plus, on a un module gérant les environnements de typage. Il pourra être implanté avec des listes d'association ou des tables de hachage, par exemple. Sa signature est :

```
module Env : sig
  type t

  (* Construction *)
  val empty : t
  val extend : ml_ident -> ml_type -> t -> t

  (* Interrogation *)
  val get : ml_ident -> t -> ml_type option
end
```

Reprenons l'exemple précédent. Partant d'un environnement vide (Env.empty), on commence par l'étendre de deux variables. Comme on n'a aucune information, il fait allouer des nouveaux noms d'inconnues (qui correspondent à t_1 et t_2) :

```
let t1 = Var_type (Unknown (new_unknown ())) in
let t2 = Var_type (Unknown (new_unknown ())) in
let env =
  Env.extend "p" t2
    (Env.extend "x" t1
      Env.empty
    ) in
```

La première instruction indique que les deux côtés de l'affectation doivent avoir le même type.

```
let lhs1 = Lv_var "p"
and rhs1 = AddrOf (Exp_var "x") in
let t_lhs1 = typeof lhs1 env
and t_rhs1 = typeof rhs1 env in
unify t_lhs1 t_rhs1;
```

Ici il se passe plusieurs choses intéréssantes. D'une part nous faisont appel à une fonction externe typeof qui retourne le type d'une expression sous un environnement (dans une implantation complète il s'agirait d'un appel récursif). Dans ce cas, `typeof lhs1 env` est identique à `Env.get lhs1 env` et `typeof rhs1 env` à `Ptr_type t1`. L'autre aspect intéressant est la dernière ligne : la fonction `unify` va modifier en place les représentations des types afin de les rendre égales. L'implantation de `unify` sera décrite plus tard. Dans ce cas précis, elle va faire pointer la référence dans t2 vers t1 (figure 7.3).

Enfin, la seconde affectation se déroule à peu près de la même manière.

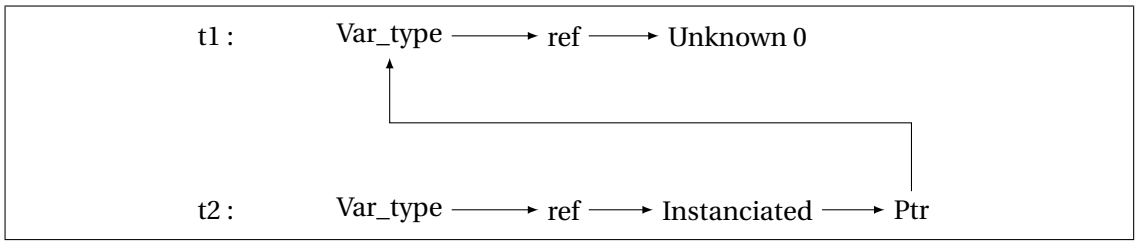


FIGURE 7.3: Unification : partage

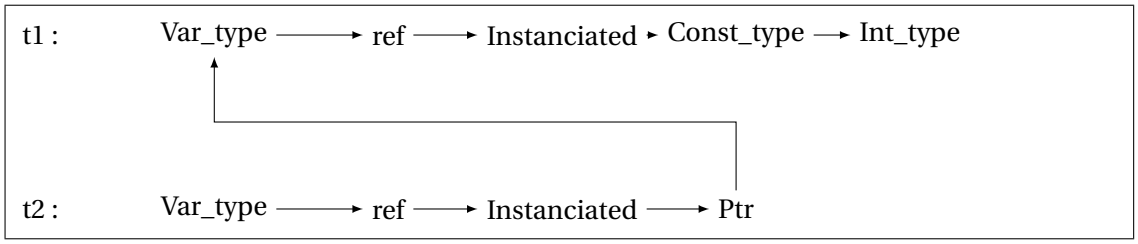


FIGURE 7.4: Unification par mutation de références

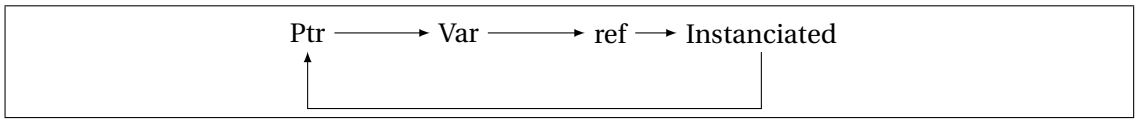


FIGURE 7.5: Cycle dans le graphe de types

```
let lhs2 = Lv_deref (Lv_var "p")
and rhs2 = Exp_int 0 in
let t_lhs2 = typeof lhs2 env
and t_rhs2 = typeof rhs2 env in
unify t_lhs2 t_rhs2;
```

Ici `typeof lhs2 env` est identique à `Ptr_type (Env.get "p" env)` et `typeof lhs2 env` à `Const_type Int_type`. Et dans cas, l'unification doit se faire entre t1 et `Const_type Int_type` : cela mute la référence derrière t1 (figure 7.4).

L'essence de l'algorithme d'inférence se situe donc dans 2 fonctions. D'une part, `unify` qui réalise l'unification des types grâce à au partage des références. D'autre part, la `typeof` qui encode les règles de typage elles-mêmes et les applique à l'aide de `unify`.

Algorithme d'unification

Voici une implantation de la fonction `unify`.

Celle-ci prend en entrée deux types t_1 et t_2. À l'issue de l'exécution de `unify`, ces deux types doivent pouvoir être considérés comme égaux. Si ce n'est pas possible, une erreur sera levée.

La première étape est de réduire ces deux types, c'est-à-dire à transformer les constructions `Var (ref (Instanciated t))` en `t`.

Ensuite, cela dépend des formes qu'ont les types réduits :

- si les deux types sont inconnus (de la forme `Var (ref (Instanciated t))`), on fait pointer l'une des deux références vers le premier type. Notons que cela crée un type de la forme `Var (ref (Instanciated (Var (ref (Unknown n)))))` qui sera réduit lors d'une prochaine étape d'unification.

```
let process_npk npk =
  let tpk = Npk2tpk.convert_unit npk in
  let order = Topological.topological_sort (Topological.make_graph npk) in

  let function_is_defined f =
    Hashtbl.mem tpk.Tyspeak.fundecs f
  in

  let (internal_funcs, external_funcs) =
    List.partition function_is_defined order
  in

  let exttbl = Printer.parse_external_type_annotations tpk in

  let env =
    env_add_external_fundecs exttbl external_funcs Env.empty
  in
  let s = Infer.infer internal_funcs env tpk in
  begin
    if !Options.do_checks then
      Check.check env s
  end;
  Printer.dump s
```

FIGURE 7.6: Fonction principale de `ptrtype`

- si un type est inconnu et pas l'autre, il faut de la même manière affecter la référence. Mais en faisant ça inconditionnellement, cela peut poser problème : par exemple en tentant d'unifier a avec `Ptr(a)` on pourrait créer un cycle dans le graphe (figure 7.5). Pour éviter cette situation, il suffit de s'assurer que le type inconnu n'est pas présent dans le type à affecter.
- si les deux types sont des types de base (comme INT ou FLOAT) égaux, on ne fait rien.
- si les deux types sont des constructeurs de type, il faut que les constructeurs soient égaux. On unifie en outre leurs arguments deux à deux.
- dans les autres cas, l'algorithme échoue.

7.3 Architecture de `ptrtype`

L'outil `ptrtype` lit un programme Newspeak (ou un fichier C), et réalise l'inférence de qualificateurs. En sortie, il affiche soit le programme complètement annoté, soit une erreur. Le cœur de l'outil est dans la

Si l'argument à `ptrtype` est un fichier C, il est tout d'abord compilé vers Newspeak grâce à l'utilitaire `c2newspeak`. Ensuite, les autres passes travaillent sur une représentation intermédiaire proche de Newspeak, mais où des étiquettes de type supplémentaires sont ajoutées. Ce type de représentation intermédiaire (polymorphe en le type des étiquettes) est `'a Tyspeak.t`.

Le reste de l'outil est résumé dans la fonction `process_npk` (figure 7.6) :

- Grâce à la fonction `convert_unit : Newspeak.t -> unit Tyspeak.t`, on ajoute des étiquettes “vides” (toutes égales à `() : unit`).
- L'ensemble des fonctions du programme est trié topologiquement selon la relation $\preceq$ définie par $f \preceq g \stackrel{\text{def}}{=\!=}$ “g apparaît dans la définition de f”. Cela est fait en construisant une représentation de $\preceq$ sous forme de graphe, puis en faisant un parcours en largeur de celui-ci.
- Les annotations extérieures sont alors lues (variable `exttbl`), ce qui permet de créer un environnement initial.
- Les types de chaque fonction sont inférés, par le biais de la fonction suivante :

```
val infer : Newspeak.fid list
         -> Types.simple Env.t
         -> 'a Tyspeak.t
         -> Types.simple Tyspeak.t
```

- Le programme obtenu, de type `Types.simple Tyspeak.t`, est affiché sur le terminal.

7.4 Inférence de types

L'inférence de types consiste à remplacer les étiquettes de type `unit` par des étiquettes de type `simple` (autrement dit de vraies représentations de types).

Cette étape se fait de manière impérative : on peut créer de nouveaux types avec `new_unknown` et unifier deux types avec `unify`. Leurs types sont :

```
val new_unknown : unit -> Types.simple
val unify : Types.simple -> Types.simple -> unit
```

La fonction `infer` s'appuie sur un ensemble de fonctions récursivement définies portant sur chaque type de fragment : `infer_fdec` pour les déclarations de fonction, `infer_exp` pour les expressions, `infer_stmtkind` pour les instructions, etc.

Les règles de typage sont implantées par `new_unknown` et `unify`. Par exemple, pour typer une déclaration (figure 7.7), on crée un nouveau type `t0`. On étend l'environnement courant avec cette nouvelle association et sous ce nouvel environnement, on type le bloc de portée de la déclaration.

De même, pour typer un appel de fonction, on infère le type de ses arguments et left-values de retour. On obtient également le type de la fonction (à partir du type de la fonction présent dans l'environnement, ou du type du pointeur de fonction qui est déréférencé), et on unifie ces deux informations.

On peut donner quelques exemples, comme :

- addition de deux flottants (dans `infer_binop`) :

```
let infer_binop op (_, a) (_, b) =
  match op with
    (* [...] *)
    | N.PlusF _ ->
        unify a Float;
        unify b Float;
        Float
```

```
let rec infer_stmtkind env sk =
  match sk with
  (* [...] *)
  | T.Decl (n, nty, _ty, blk) ->
      let var = T.Local n in
      let t0 = new_unknown () in
      let new_env = Env.add (VLocal n) (Some nty) t0 env in
      let blk' = infer_blk new_env blk in
      let ty = lval_type new_env var in
      T.Decl (n, nty, ty, blk')
  (* [...] *)
  | T.Call (args, fexp, rets) ->
      let infer_arg (e, nt) =
        let et = infer_exp env e in
        (et, nt)
      in

      let infer_ret (lv, nt) =
        (infer_lv env lv, nt)
      in

      let args' = List.map infer_arg args in
      let rets' = List.map infer_ret rets in

      let t_args = List.map (fun ((_, t), _) -> t) args' in
      let t_rets = List.map (fun (lv, _) -> lval_type env lv) rets' in

      let (fexp', tf) = infer_funexp env fexp in
      let call_type = Fun (t_args, t_rets) in
      unify tf call_type;
      T.Call (args', fexp', rets')
```

FIGURE 7.7: Inférence des déclarations de variable et appels de fonction

- adresse d'une left-value (dans `lval_type`) :

```
| T.AddrOf lv ->
    let lv' = infer_lv env lv in
    let ty = lval_type env lv in
    (T.AddrOf lv', Ptr (Kernel, ty))
```

- déréférencement d'une left-value (dans `lval_type`) :

```
| T.Deref(e, _sz) ->
    let (_, te) = infer_exp env e in
    let t = new_unknown () in
    unify (Ptr (Kernel, t)) te;
    t
```

```
type unknown = { id : int }

type 'a var_type =
  | Unknown of unknown
  | Instanciated of 'a

and qual =
  | Kernel
  | User
  | QVar of qual var_type ref

and simple =
  | Int
  | Float
  | Fun of simple list * simple list
  | Ptr of qual * simple
  | Array of simple
  | Struct of (int * simple) list ref
  | Var of simple var_type ref
```

FIGURE 7.8: Représentation des types

7.5 Unification

Les types de données utilisés sont donnés dans la figure 7.8. Les types SAFESPEAK sont représentés soit par un constructeur de type "résolu" immutable comme `Int`, soit par une référence dans le cas d'une variable inconnue (placée alors derrière le constructeur `Var`).

Ces références contiennent une valeur du type `simple var_type` [1], c'est-à-dire :

- soit un numéro d'inconnue (constructeur `Unknown`) [2].
- soit un type résolu (constructeur `Instanciated`) si cette inconnue a été unifiée avec un type concret.

Ce système peut créer des représentations de types arbitrairement longues, comme par exemple :

```
Var (ref (Instanciated (Var (ref (Instanciated Int)))))
```

Cela est dû au fait que `fun x -> Var (ref (Instanciated x))` est typée `simple -> simple` et peut donc être appliquée à loisir. Pour éviter cet effet d'allongement, on définit une fonction `shorten` qui supprime ces chaînes (figure 7.9).

La fonction d'unification, quant à elle, commence par raccourcir ses deux arguments puis faire une analyse de cas par filtrage.

Les cas principaux pour unifier deux types réduits `sta` et `stb` sont :

- `sta` et `stb` sont deux inconnues. Alors on modifie l'un pour pointer sur l'autre. Les références étant uniques et partagées, cela revient à substituer l'un par l'autre dans toutes les représentations de types.

1. Ce type est polymorphe pour pouvoir être repris dans l'unification des qualificateurs (figure 7.12).

2. On place le numéro dans un enregistrement pour abstraire l'entier sous-jacent, en empêchant par exemple de faire de l'arithmétique sur celui-ci.

```
let rec shorten = function
  | Var ({contents = Instanciated (Var _ as t)} as vt) ->
      let t2 = shorten t in
      vt := Instanciated t;
      t2
  | Var {contents = Instanciated t} -> t
  | t -> t
```

FIGURE 7.9: Fonction de racourcissement des représentations de types

- Un type est concret, l'autre une inconnue. Dans ce cas on modifie le second comme étant un `Instanciated` du premier. Il faut vérifier que le type inconnu remplacé n'apparaît pas dans le type concret, sinon on créé un cycle. Ce cas a lieu par exemple quand on cherche à unifier t avec $t*$. Il faut alors signaler une erreur, ce que fait la fonction `occurs_check_failed`.
- Les deux types sont des types concrets. Alors ils sont de la forme respective $C(t_1,\ldots,t_n)$ et $D(u_1,\ldots,u_m)$ où C et D sont des constructeurs de type avec respectivement n et m arguments[3]. Si $C = D$ et $n = m$, alors on unifie récursivement chaque t_i avec u_i. Sinon on lève une erreur (fonction `type_clash`).

Cette analyse de cas est implantée dans la fonction `unify_now` (figure 7.10).
cas structure (figure 7.11).

Pour les pointeurs, il est nécessaire de définir des fonctions similaires sur les qualificateurs (figure 7.12).

La fonction appellée directement par le reste du code, appellée `unify`, peut retarder l'unification (figure 7.13). Dans ce cas, la paire de types à unifier est mise dans une liste d'attente qui sera unifiée après le parcours du programme. Le but est d'instrumenter l'inférence de types afin de pouvoir en faire une exécution "pas à pas".

7.6 Exemple

Lançons l'analyse sur un petit exemple :

```
int f(int *x) { return (*x + 1); }
```

L'exécution de notre analyseur affiche un programme complètement annoté :

```
 % ptrtype example.c
f : (Ptr (Int)) -> (Int)
Int (example.c:1#4)^f(Ptr (Int) x) {
  (.c:3#4)^!return =(int32)
    (coerce[-2147483648,2147483647]
      ( ( ([(x_Ptr (Int) : Ptr (Int))]32_Int
            : Int
          )
          + (1 : Int)
        ) : Int
```

3. Le cas $n = 0$ ou $m = 0$ correspond aux types concrets comme INT ou FLOAT.

```
let rec unify_now ta tb =
  let sta = shorten ta in
  let stb = shorten tb in
  match (sta, stb) with
  | ((Var ({contents = Unknown na} as ra)),
     (Var  {contents = Unknown nb})) ->
       begin
         if na <> nb then
           ra := Instanciated stb
       end
  | ((Var ({contents = Unknown {id = n}} as r)), t)
    ->
      begin
        if occurs n t then
          occurs_check_failed sta stb
        else
          r := Instanciated t
      end
  | (_, (Var ({contents = Unknown _}))) -> unify_now stb sta

  | Int, Int
  | Float, Float -> ()

  | Ptr (qa, ta), Ptr (qb, tb) ->
      unify_quals qa qb;
      unify_now ta tb

  | Array a, Array b -> unify_now a b

  | Fun (args_a, rets_a), Fun (args_b, rets_b) ->
      List.iter2 unify_now args_a args_b;
      List.iter2 unify_now rets_a rets_b

  | Struct rfa, Struct rfb ->
      unify_structs rfa rfb

  | _ -> type_clash sta stb
```

FIGURE 7.10: Fonction d'unification

```
let unify_structs rfa rfb =
  let fa = !rfa in
  let fb = !rfb in

  let new_a = ref [] in
  let new_b = ref [] in

  let unify_fields = function
    | _, InBoth (ta, tb) -> unify_now ta tb
    | k, OnlyL f -> new_b := (k,f) :: !new_b
    | k, OnlyR f -> new_a := (k,f) :: !new_a
  in

  List.iter unify_fields (compare_lists fa fb);
  let by_offset (x, _) (y, _) =
    compare x y
  in
  rfa := List.sort by_offset (!new_a @ !rfa);
  rfb := List.sort by_offset (!new_b @ !rfb)
```

FIGURE 7.11: Structures

```
      ) : Int
    );
}
```

L'opérateur `coerce[a,b]` est un artefact de Newspeak, destiné à détecter les débordements d'entiers lors d'une analyse de valeurs par interprétation abstraite. Dans le cas de notre analyse, les valeurs ne sont pas pertinentes et cet opérateur peut être vu que comme un "plus" unaire typé (INT) → INT.

A contrario, lorsqu'il n'est pas possible d'inférer des types compatibles, l'analyseur s'arrête avec une erreur.

```
void f(int *p) {
    /*!npk userptr p */
    *p = 3;
}
```

Le commentaire `/*!npk userptr p */` est interpréter par l'analyseur et le fait unifier le type de p avec *t* @, c'est-à-dire qu'il force son qualificateur à être USER.

```
04-addrof.c:4#4 - Cannot unify qualifiers:
  Kernel
  User
```

On signale l'emplacement où la dernière unification a échoué.

```
let rec shorten_q = function
  | QVar ({contents = Instanciated (QVar _ as t)} as vt) ->
      let t2 = shorten_q t in
      vt := Instanciated t;
      t2
  | QVar {contents = Instanciated t} -> t
  | t -> t

let rec unify_quals a b =
  let sa = shorten_q a in
  let sb = shorten_q b in
  match (sa, sb) with
  | User, User -> ()
  | Kernel, Kernel -> ()
  | ((QVar ({contents = Unknown na} as ra)),
     (QVar  {contents = Unknown nb})) ->
       begin
         if na <> nb then
           ra := Instanciated sb
       end
  | ((QVar ({contents = Unknown {id = n}} as r)), q)
    ->
      begin
        if occurs_q n q then
          occurs_q_check_failed sa sb
        else
          r := Instanciated q
      end
  | (_, (QVar ({contents = Unknown _}))) -> unify_quals sb sa
  | _ -> Utils.error "Cannot unify qualifiers:\n  %s\n  %s\n"
                       (string_of_qual sa)
                       (string_of_qual sb)
```

FIGURE 7.12: Qualificateurs

```
let unify a b =
  if !Options.lazy_unification then
    Queue.add (Unify (a, b)) unify_queue
  else
    unify_now a b
```

FIGURE 7.13: Unification directe ou retardée

CHAPITRE 8

ÉTUDE DE CAS : UN PILOTE DE CARTE GRAPHIQUE

Dans ce chapitre, un exemple de mise en œuvre du système de type décrit dans le chapitre 6 et implanté dans le chapitre 7.

8.1 Linux

Le noyau Linux, abordé dans le chapitre 2, est un noyau de système d'exploitation développé depuis le début des années 90 et "figure de proue" du mouvement open-source. Au départ écrit par Linus Torvalds sur son ordinateur personnel, il a au fil des années était porté sur de nombreuses architectures et s'est enrichi de nombreux pilotes de périphériques. En 20XX, son code source comporte 8.8 millions de lignes de code (en grande majorité du C) dont 55% de pilotes.

Même si le noyau est monolithique (la majeure partie des traitements s'effectue au sein d'un même fichier objet), les sous-systèmes sont indépendants. C'est ce qui permet d'écrire des pilotes de périphériques et des modules.

8.2 GNU C

Linux est écrit dans le langage C, mais pas sa version normalisée. Il utilise le dialecte GNU C qui est celui que supporte GCC.

Pour traduire ce dialecte, il a été nécessaire d'adapter `c2newspeak`. La principale particularité est la notation `__attribute__((...))` qui peut décorer les déclarations de fonctions, de variables ou de types. De nouvelles fonctionnalités sont aussi présentes.

Par exemple, il est possible de manipuler des étiquettes de première classe : si `lbl:` est présent avant une instruction, on peut capturer l'adresse de celle-ci avec `void *p = &&lbl` et y sauter indirectement avec `goto *p`.

Une autre fonctionnalité est le concept d'instruction-expression : `({bloc})` est une expression, dont la valeur est celle de la dernière expression évaluée lors de `i`.

Les attributs, quant à eux, rentrent dans trois catégories :

- les annotations de compilation ; par exemple, `used` désactive l'avertissement "cette variable n'est pas utilisée ".
- les optimisations ; par exemple, les objets marqués `hot` sont groupés de telle manière qu'ils se retrouvent en cache ensemble.
- les annotations de bas niveau ; par exemple, `aligned(n)` spécifie qu'un objet doit être aligné sur au moins `n` bits.

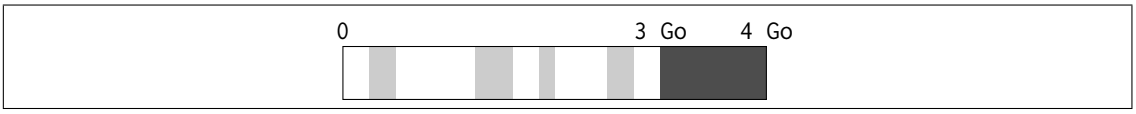


FIGURE 8.1: L'espace d'adressage d'un processus. En gris clair, les zones accessibles à tous les niveaux de privilèges : code du programme, bibliothèques, tas, pile. En gris foncé, la mémoire du noyau, réservée au mode privilégié.

Dans notre cas, toutes ces annotations peuvent être ignorées.

8.3 Configuration

Pour que le code noyau soit compilable, il est nécessaire de définir certaines macros. En particulier, le système de configuration de Linux utilise des macros nommées `CONFIG_*` pour inclure ou non certaines fonctionnalités. Il a donc fallu faire un choix ; nous avons choisi la configuration par défaut. Pour analyser des morceaux plus importants du noyau, il faudrait définir un fichier de configuration plus important.

8.4 Appels systèmes sous Linux

Dans cette section, nous allons voir comment ces mécanismes sont implantés dans le noyau Linux. Une description plus détaillée pourra être trouvée dans [BC05], ou pour le cas de la mémoire virtuelle, [Gor04].

Deux rings sont utilisés : en *ring* 0, le code noyau et en *ring* 3, le code utilisateur.

Une notion de tâche similaire à celle décrite dans la section 2.3 existe : elles s'exécutent l'une après l'autre, le changement s'effectuant sur interruptions.

Pour faire appel aux services du noyau, le code utilisateur doit faire appel à des appels systèmes, qui sont des fonctions exécutées par le noyau. Chaque tâche doit donc avoir deux piles : une pile "utilisateur" qui sert pour l'application elle-même, et une pile "noyau" qui sert aux appels système.

Grâce à la mémoire virtuelle, chaque processus possède sa propre vue de la mémoire dans son espace d'adressage (figure 8.1), et donc chacun un ensemble de tables de pages et une valeur de CR3associée. Au moment de changer le processus en cours, l'ordonnanceur charge donc le CR3du nouveau processus.

Les adresses basses (inférieures à `PAGE_OFFSET` = 3 Gio = `0xc0000000`) sont réservées à l'utilisateur. On y trouvera par exemple :

- le code du programme
- les données du programmes (variables globales)
- la pile utilisateur
- le tas (mémoire allouée par `malloc` et fonctions similaires)
- les bibliothèques partagées

Au dessus de `PAGE_OFFSET`, se trouve la mémoire réservée au noyau. Cette zone contient le code du noyau, les piles noyau des processus, etc.

Les programmes utilisateur s'exécutant en *ring* 3, ils ne peuvent pas contenir d'instructions privilégiées, et donc ne peuvent pas accéder directement au matériel (c'était le but !). Pour qu'ils puissent interagir avec le système (afficher une sortie, écrire sur le disque...), le mécanisme des appels système est nécessaire. Il s'agit d'une interface de haut niveau entre

les *rings* 3 et 0. Du point de vue du programmeur, il s'agit d'un ensemble de fonctions C "magiques" qui font appel au système d'exploitation pour effectuer des opérations.

Voyons ce qui se passe derrière la magie apparente. Une explication plus détaillée est disponible dans la documentation fournie par Intel [Int10].

Dans la bibliothèque C

Il y a bien une fonction `getpid` présente dans la bibliothèque C du système. C'est la fonction qui est directement appelée par le programme. Cette fonction commence par placer le numéro de l'appel système (noté `__NR_getpid`, valant 20 ici) dans EAX, puis les arguments éventuels dans les registres (EBX, ECX, EDX, ESI puis EDI). Une interruption logicielle est ensuite déclenchée (`int 0x80`).

Dans la routine de traitement d'interruption

Étant donné la configuration du processeur [1], elle sera traitée en *ring* 0, à un point d'entrée prédéfini (`arch/x86/kernel/entry_32.S`, `ENTRY(system_call)`).

```
        # system call handler stub
ENTRY(system_call)
        RING0_INT_FRAME                     # can't unwind into user space anyway
        pushl %eax                        # save orig_eax
        CFI_ADJUST_CFA_OFFSET 4
        SAVE_ALL
        GET_THREAD_INFO(%ebp)
                                        # system call tracing in operation / emulation
        testl $_TIF_WORK_SYSCALL_ENTRY,TI_flags(%ebp)
        jnz syscall_trace_entry
        cmpl $(nr_syscalls), %eax
        jae syscall_badsys
syscall_call:
        call *sys_call_table(,%eax,4)
        movl %eax,PT_EAX(%esp)              # store the return value
        # ...
        INTERRUPT_RETURN
```

L'exécution reprend donc en *ring* 0, avec dans ESP le pointeur de pile noyau du processus. Les valeurs des registres ont été préservées, la macro `SAVE_ALL` les place sur la pile. Ensuite, à l'étiquette `syscall_call`, le numéro d'appel système (toujours dans EAX) sert d'index dans le tableau de fonctions `sys_call_table`.

Dans l'implantation de l'appel système

Puisque les arguments sont en place sur la pile, comme dans le cas d'un appel de fonction "classique", la convention d'appel *cdecl* est respectée. La fonction implantant l'appel système, nommée `sys_getpid`, peut donc être écrite en C.

On trouve cette fonction dans `kernel/timer.c` :

1. Il est impropre de dire que le processeur est configuré — tout dépend uniquement de l'état de certains registres, ici la *Global Descriptor Table* et les *Interrupt Descriptor Tables*.

```
SYSCALL_DEFINE0(getpid)
{
        return task_tgid_vnr(current);
}
```

La macro `SYSCALL_DEFINE0` nomme la fonction `sys_getpid`, et définit entre autres des points d'entrée pour les fonctionnalités de débogage du noyau. À la fin de la fonction, la valeur de retour est placée dans EAX, conformément à la convention *cdecl*.

Retour vers le ring 3

Au retour de la fonction, la valeur de retour est placée à la place de EAX là où les registres ont été sauvegardés sur la pile noyau (`PT_EFLAGS(%esp)`). L'instruction `iret` (derrière la macro `INTERRUPT_RETURN`) permet de restaurer les registres et de repasser en mode utilisateur, juste après l'interruption. La fonction de la bibliothèque C peut alors retourner au programme appelant.

8.5 Bug

On décrit le cas d'un pilote video qui contenait un bug de pointeur utilisateur. Il est répertorié sur `http://freedesktop.org` en tant que bug #29340.

Pour changer de mode graphique, les pilotes de GPU peuvent supporter le *Kernel Mode Setting* (KMS).

Pour configurer un périphérique, l'utilisateur communique avec le pilote noyau avec le mécanisme d'*ioctls*[2]. Ils sont similaires à des appels système, mais spécifique à un périphérique particulier. Les fonctions implantant un *ioctl* sont donc vulnérables à la même classe d'attaques que les appels système, et donc doivent être écrits avec une attention particulière.

Le code suivant est présent dans le pilote KMS pour les GPU AMD Radeon :

```
/* from drivers/gpu/drm/radeon/radeon_kms.c */
int radeon_info_ioctl(struct drm_device *dev,
                      void *data,
                      struct drm_file *filp) {
struct radeon_device *rdev =
            dev->dev_private;
struct drm_radeon_info *info;
struct radeon_mode_info *minfo =
            &rdev->mode_info;
uint32_t *value_ptr;
uint32_t value;
struct drm_crtc *crtc;
int i, found;

info = data;
value_ptr = (uint32_t *)
            ((unsigned long)info->value);
value = *value_ptr;
```

2. Ce nom vient de la fonction `ioctl()` pour *Input/Output Control*.

```
--- a/drivers/gpu/drm/radeon/radeon_kms.c
+++ b/drivers/gpu/drm/radeon/radeon_kms.c
@@ -112,7 +112,9 @@

        info = data;
        value_ptr = (uint32_t *)((unsigned long)info->value);
-       value = *value_ptr;
+       if (DRM_COPY_FROM_USER(&value, value_ptr, sizeof(value)))
+               return -EFAULT;
+
        switch (info->request) {
        case RADEON_INFO_DEVICE_ID:
                value = dev->pci_device;
```

FIGURE 8.2: Patch résolvant le problème de pointeur utilisateur.

```
        [...]
}
```

On peut voit que l'argument `data` est converti en un `struct drm_radeon_info *`. Un pointeur `value_ptr` est extrait de son champ `value`, et finalement ce pointeur est déréférencé.

Cependant, l'argument `data` est un pointeur vers une structure (allouée en espace noyau) du type suivant, dont les champs proviennent d'un appel utilisateur de `ioctl()`.

```
/* from include/drm/radeon_drm.h */
struct drm_radeon_info {
  uint32_t    request;
  uint32_t    pad;
  uint64_t    value;
};
```

Pour mettre ce problème en évidence, nous avons annoté la fonction `radeon_info_ioctl` de telle manière que son second paramètre soit un pointeur noyau vers une structure contenant un champ USER, `value`. Ceci est possible puisqu'avant la traduction, on efface les types présents dans le programme C. Ainsi, un pointeur ne peut pas être distingué d'un entier transtypé en un pointeur. Avec cette configuration, nous obtenons une erreur de type à la ligne 16.

L'intégralité de ce code peut être trouvée en annexe A.

La bonne manière de faire a été publiée avec le numéro de *commit* `d8ab3557` (figure 8.2) (`DRM_COPY_FROM_USER` étant une simple macro pout `copy_from_user`). Dans ce cas, on n'obtient pas d'erreur de typage.

8.6 Détails

Pour utiliser notre système de types, plusieurs étapes sont nécessaires en plus de traduire le noyau linux en SAFESPEAK.

Afin de réaliser l'analyse, il faut annoter les sources pour créer un environnement initial (via la variable `exttbl` décrite en section 7.3).

Ensuite, il faut réécrire les fonctions de manipulation de pointeurs fournies par le noyau : `get_user`, `put_user`, `copy_from_user`, `copy_to_user`, etc. Leur implantation revient à réaliser un test puis à faire la copie. À leur place on utilise les primitives copy_from_user$(\cdot,\cdot)$ et copy_to_user$(\cdot,\cdot)$.

Enfin, on peut lancer l'inférence de type.

CONCLUSION DE LA PARTIE III

Ccl III.

Quatrième partie

Conclusion

CHAPITRE 9

CONCLUSION

On fait ici un bilan des travaux présentés, en commençant par un bilan des contributions réalisées. On fait ensuite un tour des aspects posant problème, ou traités de manière incomplète, en évoquant les travaux possibles pour enrichir l'expressivité de ce système.

9.1 Contributions

Un langage impératif bien typable Le système de types de C est trop rudimentaire pour permettre de d'obtenir des garanties sur l'exécution des programmes bien typés. En interdisant certaines constructions dangereuses et en en annotant certaines, nous avons isolé un langage impératif bien typable, pour lequel on peut définir un système de types sûr.

Un système de types abstraits En partant de ce système de types, on a décrit une extension permettant de créer des pointeurs pour lesquels l'opération de déréférencement est restreinte à certaines fonctions. Dans le contexte d'un noyau de système d'exploitation, cette restriction permet de vérifier statiquement qu'à aucun moment le noyau ne déréférence un pointeur dont la valeur est contrôlée par l'espace utilisateur, évitant ainsi un problème de sécurité.

9.2 Limitations et travaux futurs

Arithmétique En C, on dispose de plusieurs types entiers, pouvant avoir plusieurs tailles et être signés ou non signés, ainsi que des types flottants qui diffèrent par leur taille. Au contraire, en SAFESPEAK on ne conserve qu'un seul type d'entier. La raison pour cela est que nous ne nous intéressons pas du tout aux problématiques de sémantique entière : les débordements, dénormalisations, etc, sont supposés ne pas arriver.

Comme en C il n'est pas possible de définir de nouveaux types de base (des entiers de 24 bits par exemple), il est facile d'étendre le système de types de SAFESPEAK pour ajouter tous ces nouveaux types. La traduction depuis Newspeak insère déjà des opérateurs de transtypage pour lesquels il est facile de donner une sémantique (pouvant lever une erreur en cas de débordement, comme en Ada) et un typage. Les littéraux numériques peuvent poser problème, puisqu'ils deviennent alors polymorphes. Une solution peut être de leur donner le plus grand type entier et d'insérer un opérateur de transtypage à chaque littéral. Haskell utilise une solution similaire : les littéraux entier ont le type de précision arbitraire `Integer` et

est converti dans le bon type en appelant la fonction `fromInteger` du type synthétisé à partir de l'environnement.

Transtypage Puisque l'approche retenue est basée sur le typage statique, il est impossible de capturer de nombreuses constructions qui sont possibles, ou même idiomatiques en C : les unions, les conversions de types (explicites ou implicites) et le *type punning* (défini ci-desosus). Les deux premières sont équivalentes. Bien qu'on puisse remplacer chaque conversion explicite d'un type t_1 vers un type t_2 par l'appel à une fonction cast_{t_1,t_2}, on ajoute alors un "trou" dans le système de types. Cette fonction devrait en effet être typée $(t_1) \to t_2$, autrement dit le type "maudit" $\alpha \to \beta$ de `Obj.magic` en OCaml ou `unsafeCoerce` en Haskell.

Le *type punning* consiste à modifier directement la suite de bits de certaines données pour la manipuler d'une manière efficace. Par exemple, il est commun de définir un ensemble de macros pour accéder à la mantisse et à l'exposant de flottants IEEE754. Ceci peut être fait avec des unions ou des masques de bits.

Dans de tels cas, le typage statique est bien sûr impossible. Pour traiter ces cas, il faudrait encapsuler la manipulation dans une fonction et y ajouter une information de type explicite, comme `float_exponent` : (FLOAT) → INT.

En fait, les casts entre types entiers ne posent pas de problèmes. Ceux entre types pointeurs ne posent pas de problème non plus puisqu'au moment de la traduction en SAFESPEAK on efface les étiquettes de types sur les pointeurs.

Mais les casts entre entiers et pointeurs posent problème. Si on autorise les casts, on peut néanmoins trouver le solution suivante.

$$\frac{\Gamma \vdash e : t*}{\Gamma \vdash (\text{INT})\ e : \text{INT}}\ (\text{PTRINT-BAD}) \qquad \frac{\Gamma \vdash e : \text{INT}}{\Gamma \vdash (\text{PTR})\ e : t*}\ (\text{INTPTR-BAD})$$

Cela pose problème car il est alors possible de créer une fonction qui convertit n'importe quel type pointeur en n'importe quel autre type pointeur :

$$\vdash \text{fun}(p)\{\text{RETURN}((\text{PTR})\ (\text{INT})\ p)\} : (t_a*) \to t_b*$$

L'opération problématique est en fait l'opérateur (INT) de conversion d'un pointeur vers un entier, car l'entier résultant ne pose pas de problème de sûreté ni de sécurité. Pour lui donner une sémantique, on peut supposer que l'environnement d'exécution est paramétré par une fonction de placement en mémoire $\text{addr}^\natural : \Phi \to \text{INT}$ qui à un chemin associe un entier. On étend alors la sémantique par la règle suivante.

$$\frac{}{\langle (\text{INT})\ \widehat{\&}\varphi, m\rangle \to \langle \text{addr}^\natural(\varphi), m\rangle}\ (\text{CASTINT})$$

Quant au transtypage dans le sens inverse, il est plus compliqué à traiter. À l'exécution il est en effet nécessaire d'avoir à disposition une fonction $\text{interp}_{(t*)} : \text{INT} \to t*$ qui permette d'interpréter un entier comme pointeur qualifié q vers un objet de type t de manière qu'on ajoute la règle suivante.

$$\frac{}{\langle (\text{PTR})\ \widehat{n}, m\rangle \to \langle \text{interp}_{(t*)}(\widehat{n}), m\rangle}\ (\text{CASTPTR})$$

L'exécution de l'analogue de $\text{addr}^\natural$ demanderait de créer un pointeur depuis un entier.

La solution la plus sûre est donc d'interdire totalement ces conversions.

Environnement d'exécution La sémantique opérationnelle utilise un environnement d'exécution pour certains cas. Contrairement à C, les débordements de tampon et les déréférencements de pointeurs sont vérifiés dynamiquement. Mais ce n'est pas une caractéristique cruciale de cette approche : en effet, si on suppose que les programmes que l'on analyse sont corrects de ce point de vue, on peut désactiver ces vérifications et le reste des propriétés est toujours valable.

Un autre endroit, plus problématique, où des tests dynamiques sont faits est lorsqu'on recherche en mémoire des pointeurs référençant un cadre d'appel qui n'est plus valide (à l'aide de l'opérateur Cleanup$(\cdot)$). Supprimer ce test rend l'analyse incorrecte, car il est alors possible de faire référence à une variable avec un type différent.

De même, si on peut avoir une garantie statique que les adresses des variables locales ne seront plus accessibles au retour d'une fonction, alors on peut supprimer le nettoyage en posant Cleanup$(m) = m$. Cette garantie peut être obtenue avec une analyse statique préalable [DDMPn10].

Allocation dynamique La plupart des programmes, et le noyau Linux en particulier, utilisent la notion d'allocation dynamique de mémoire. C'est une manière de créer dynamiquement une zone de mémoire qui restera accessible après l'exécution de la fonction courante. Cette mémoire pourra être libérée à l'aide d'une fonction dédiée. Dans l'espace utilisateur, les programmes peuvent utiliser les fonctions `malloc()`, `calloc()`, et `realloc()` pour allouer des zones de mémoire et `free()` pour les libérer. Dans le noyau Linux, ces fonctions existent sous la forme de `kmalloc()`, `kfree()`, etc. Une explication détaillée de ces mécanismes peut être trouvée dans [Gor04].

Ces fonctions manipulent les données en tant que zones mémoires opaques, à en renvoyant un pointeur vers une zone mémoire d'un nombre d'octets donnés. Cela présuppose un modèle mémoire plus bas niveau. Pour se rapprocher de la sémantique de SAFESPEAK, une manière de faire est de définir un opérateur de plus haut niveau prenant une expression et retournant l'adresse d'une cellule mémoire contenant cette valeur. Le typage est alors direct :

$$\frac{\Gamma \vdash e : t}{\Gamma \vdash \text{ALLOC}(e) : t*}\ (\text{MALLOC}) \qquad \frac{\Gamma \vdash e : t*}{\Gamma \vdash \text{FREE}(e)}\ (\text{FREE})$$

En ce qui concerne l'exécution, on peut ajouter une troisième composante aux états mémoire : $m = (s, g, h)$ où h associe à une étiquette numérique une valeur (comme les globales mais indexées par des entiers plutôt que par leur nom). La libération de mémoire est problématique parce qu'il faut faire confiance au programmeur pour ne pas accéder aux zones mémoires libérées. Encore une fois, on peut utiliser une analyse préalable comme [DDMPn10]. Il est sûr (mais pas possible en pratique) d'ignorer les commandes de libérations de mémoire.

Structures récursives Une autre limitation est que seules les structures simples sont supportées. Les structures récursives, contenant un pointeur vers un objet du même type structure, ne sont pas typables. La raison est syntaxique : l'accès à un champ est noté $lv.l_S$ afin d'avoir à disposition le type complet de la structure.

Par exemple, une liste chaînée d'entiers peut être décrite de la manière suivante en C :

```
struct int_list {
    int value;
    struct int_list *next;
};
```

La compilation est possible car la définition complète de `struct int_list` n'est pas nécessaire pour connaître la taille d'un pointeur vers un objet de ce type.

Pour compiler `l.next` en SAFESPEAK, il faudrait écrire le type en extension :

$$l.next_{value:\text{INT};next:\{value:\text{INT};next:...\}\ *}$$

On peut fermer cette récursion infinie en ajoutant directement un opérateur de point fixe qui rend accessible le type d'un pointeur vers le type structure en train d'être défini.

$$S ::= \ldots \quad | \quad \text{fix}(p \rightarrow S)$$

On écrit alors :

$$l.next_{\text{fix}(p \rightarrow \{value:\text{INT};next:p\})}$$

Les structures mutuellement récursives peuvent être également exprimées de cette manière.

```
struct a {
    int value;
    struct b *next;
};

struct b {
    float value;
    struct a *next;
};
```

Ces structures sont respectivement traduites en :

$$A = \text{fix}(p \rightarrow \{value : \text{INT}; next : \{value : \text{FLOAT}; next : p\}\ *\})$$
$$B = \text{fix}(p \rightarrow \{value : \text{FLOAT}; next : \{value : \text{INT}; next : p\}\ *\})$$

Ce schéma suffit à compiler toutes les types de structures possibles en C. En effet, les seuls cas posant problème arrivent lorsqu'on fait référence à la structure en cours de définition ; et il n'est alors possible que d'y accéder par pointeur.

Assembleur Comme la plupart des outils d'analyse de code C, il est impossible de traiter l'assembleur en ligne qui peut se trouver entre deux instructions. Dans le cas de Linux, le code est fait pour être portable, et les parties dépendantes d'une certaine architecture (et donc le code assembleur) sont isolées explicitement. On peut alors si nécessaire ajouter une annotation de type sur une fonction dont l'implantation est faite en assembleur, mais au sein de cette implantation on ne peut bien sûr rien dire.

Analyse du noyau Linux Ici nous avons présenté l'analyse expérimentale d'une fonction problématique d'une interface programmation particulière. Mais le principe de l'analyse est applicable à toutes les fonctions de traitement d'`ioctl`s des pilotes KMS, et même à toutes les fonctions faisant partie des différentes interfaces recevant un pointeur de l'espace utilisateur.

Autres types abstraits Notre approche est basée sur le fait de rendre abstrait un type de C et d'y interdire certaines opérations : ici, on marque un certain type de pointeur comme "utilisateur" et on interdit l'opérateur `*` dessus.

Le langage C n'ayant pas de types abstraits [1] , on ne peut pas séparer la réprésentation d'un type (par exemple : entier signé de 32 bits) des opérations qui y sont attachées.

Dans de nombreuses interfaces, on emploie des types entiers qui servent d'étiquettes. Par exemple, les descripteurs de fichiers renvoyés par la fonction `open()` et passés au fonctions `read()` et `write()` ont le type `int`. Le langage autorise donc par exemple de multiplier entre eux deux descripteurs de fichiers, ce qui ne correspond pas à une opération concrète.

On peut aussi distinguer plusieurs types abstraits entre eux. Par exemple, si un encodeur vidéo manipule des numéros de *frame* et des identifiants de codec tous les deux entiers, on peut interdire d'utiliser avec un identifiant de codec une fonction prenant en paramètre un numéro de *frame*.

1. La seule abstraction possible est lorsqu'on manipule une structure par pointeur. Il n'est alors pas nécessaire de connaître sa définition totale. L'idiome est alors de placer uniquement une déclaration en avance (`struct s;`) dans l'en-tête (`.h`) et de renseigner la définition complète dans l'implantation (`.c`).

Annexes

MODULE RADEON KMS

```
/* from drivers/gpu/drm/radeon/radeon_kms.c */
int radeon_info_ioctl(struct drm_device *dev, void *data, struct drm_file *filp)
{
        struct radeon_device *rdev = dev->dev_private;
        struct drm_radeon_info *info;
        struct radeon_mode_info *minfo = &rdev->mode_info;
        uint32_t *value_ptr;
        uint32_t value;
        struct drm_crtc *crtc;
        int i, found;

        info = data;
        value_ptr = (uint32_t *)((unsigned long)info->value);
        value = *value_ptr;
        switch (info->request) {
        case RADEON_INFO_DEVICE_ID:
                value = dev->pci_device;
                break;
        case RADEON_INFO_NUM_GB_PIPES:
                value = rdev->num_gb_pipes;
                break;
        case RADEON_INFO_NUM_Z_PIPES:
                value = rdev->num_z_pipes;
                break;
        case RADEON_INFO_ACCEL_WORKING:
                /* xf86-video-ati 6.13.0 relies on this being false for evergreen */
                if ((rdev->family >= CHIP_CEDAR) && (rdev->family <= CHIP_HEMLOCK))
                        value = false;
                else
                        value = rdev->accel_working;
                break;
        case RADEON_INFO_CRTC_FROM_ID:
                for (i = 0, found = 0; i < rdev->num_crtc; i++) {
                        crtc = (struct drm_crtc *)minfo->crtcs[i];
                        if (crtc && crtc->base.id == value) {
```

```
                                struct radeon_crtc *radeon_crtc = to_radeon_crtc(crtc);
                                value = radeon_crtc->crtc_id;
                                found = 1;
                                break;
                        }
                }
                if (!found) {
                        DRM_DEBUG_KMS("unknown crtc id %d\n", value);
                        return -EINVAL;
                }
                break;
        case RADEON_INFO_ACCEL_WORKING2:
                value = rdev->accel_working;
                break;
        case RADEON_INFO_TILING_CONFIG:
                if (rdev->family >= CHIP_CEDAR)
                        value = rdev->config.evergreen.tile_config;
                else if (rdev->family >= CHIP_RV770)
                        value = rdev->config.rv770.tile_config;
                else if (rdev->family >= CHIP_R600)
                        value = rdev->config.r600.tile_config;
                else {
                        DRM_DEBUG_KMS("tiling config is r6xx+ only!\n");
                        return -EINVAL;
                }
        case RADEON_INFO_WANT_HYPERZ:
                mutex_lock(&dev->struct_mutex);
                if (rdev->hyperz_filp)
                        value = 0;
                else {
                        rdev->hyperz_filp = filp;
                        value = 1;
                }
                mutex_unlock(&dev->struct_mutex);
                break;
        default:
                DRM_DEBUG_KMS("Invalid request %d\n", info->request);
                return -EINVAL;
        }
        if (DRM_COPY_TO_USER(value_ptr, &value, sizeof(uint32_t))) {
                DRM_ERROR("copy_to_user\n");
                return -EFAULT;
        }
        return 0;
}

/* from include/drm/radeon_drm.h */
struct drm_radeon_info {
        uint32_t                request;
```

```
        uint32_t                pad;
        uint64_t                value;
};

/* from drivers/gpu/drm/radeon/radeon_kms.c */
struct drm_ioctl_desc radeon_ioctls_kms[] = {
        /* KMS */
        DRM_IOCTL_DEF(DRM_RADEON_INFO, radeon_info_ioctl, DRM_AUTH|DRM_UNLOCKED)
};

/* from drivers/gpu/drm/radeon/radeon_drv.c */

static struct drm_driver kms_driver = {
        .driver_features =
            DRIVER_USE_AGP | DRIVER_USE_MTRR | DRIVER_PCI_DMA | DRIVER_SG |
            DRIVER_HAVE_IRQ | DRIVER_HAVE_DMA | DRIVER_IRQ_SHARED | DRIVER_GEM,
        .dev_priv_size = 0,
        .ioctls = radeon_ioctls_kms,
        .name = "radeon",
        .desc = "ATI Radeon",
        .date = "20080528",
        .major = 2,
        .minor = 6,
        .patchlevel = 0,
};

/* from drivers/gpu/drm/drm_drv.c */
int drm_init(struct drm_driver *driver)
{
        DRM_DEBUG("\n");
        INIT_LIST_HEAD(&driver->device_list);

        if (driver->driver_features & DRIVER_USE_PLATFORM_DEVICE)
                return drm_platform_init(driver);
        else
                return drm_pci_init(driver);
}
```

ANNEXE B

RÈGLES D'ÉVALUATION

$$\frac{\langle e, m\rangle \to \langle e', m'\rangle}{\langle C(\!|e|\!), m\rangle \to \langle C(\!|e'|\!), m'\rangle}\ (\textsc{Ctx}) \qquad \frac{\langle l\,v, m\rangle \to \langle l\,v', m'\rangle}{\langle C_L(\!|l\,v|\!)_L, m\rangle \to \langle C_L(\!|l\,v'|\!)_L, m'\rangle}\ (\textsc{Ctx-Lv})$$

$$\frac{\langle i, m\rangle \to \langle i', m'\rangle}{\langle C_I(\!|i|\!)_I, m\rangle \to \langle C_I(\!|i'|\!)_I, m'\rangle}\ (\textsc{Ctx-Instr})$$

Contextes

$$\begin{aligned} C ::= {} & C_L \\ \mid {} & C \boxplus e \\ \mid {} & v \boxplus C \\ \mid {} & \boxminus C \\ \mid {} & C \leftarrow e \\ \mid {} & \varphi \leftarrow C \\ \mid {} & \{l_1 : v_1; \ldots; l_i : C; \ldots; l_n : e_n\} \\ \mid {} & [v_1; \ldots; C; \ldots; e_n] \\ \mid {} & C(e_1, \ldots, e_n) \\ \mid {} & f(v_1, \ldots, C, \ldots, e_n) \end{aligned}$$

Contextes (left-values)

$$\begin{aligned} C_L ::= {} & \bullet \\ \mid {} & *\,C_L \\ \mid {} & C_L.l_S \\ \mid {} & C_L[e] \\ \mid {} & \varphi[C] \end{aligned}$$

Contextes (instructions)

$$\begin{aligned} C_I ::= {} & C_I; i \\ \mid {} & \textsc{If}(C)\{i_1\}\textsc{Else}\{i_2\} \\ \mid {} & \textsc{Return}(C) \\ \mid {} & \textsc{Decl}\ x = C\ \textsc{in}\{i\} \\ \mid {} & C \end{aligned}$$

FIGURE B.1: Règles d'évaluation – contextes

$\boxed{\Xi \to \Omega}$

$$\frac{}{\langle \Omega, m\rangle \to \Omega}\ (\textsc{Exp-Err}) \qquad \frac{\langle e, m\rangle \to \Omega}{\langle C(\!|e|\!), m\rangle \to \Omega}\ (\textsc{Eval-Err})$$

FIGURE B.2: Règles d'évaluation – erreurs

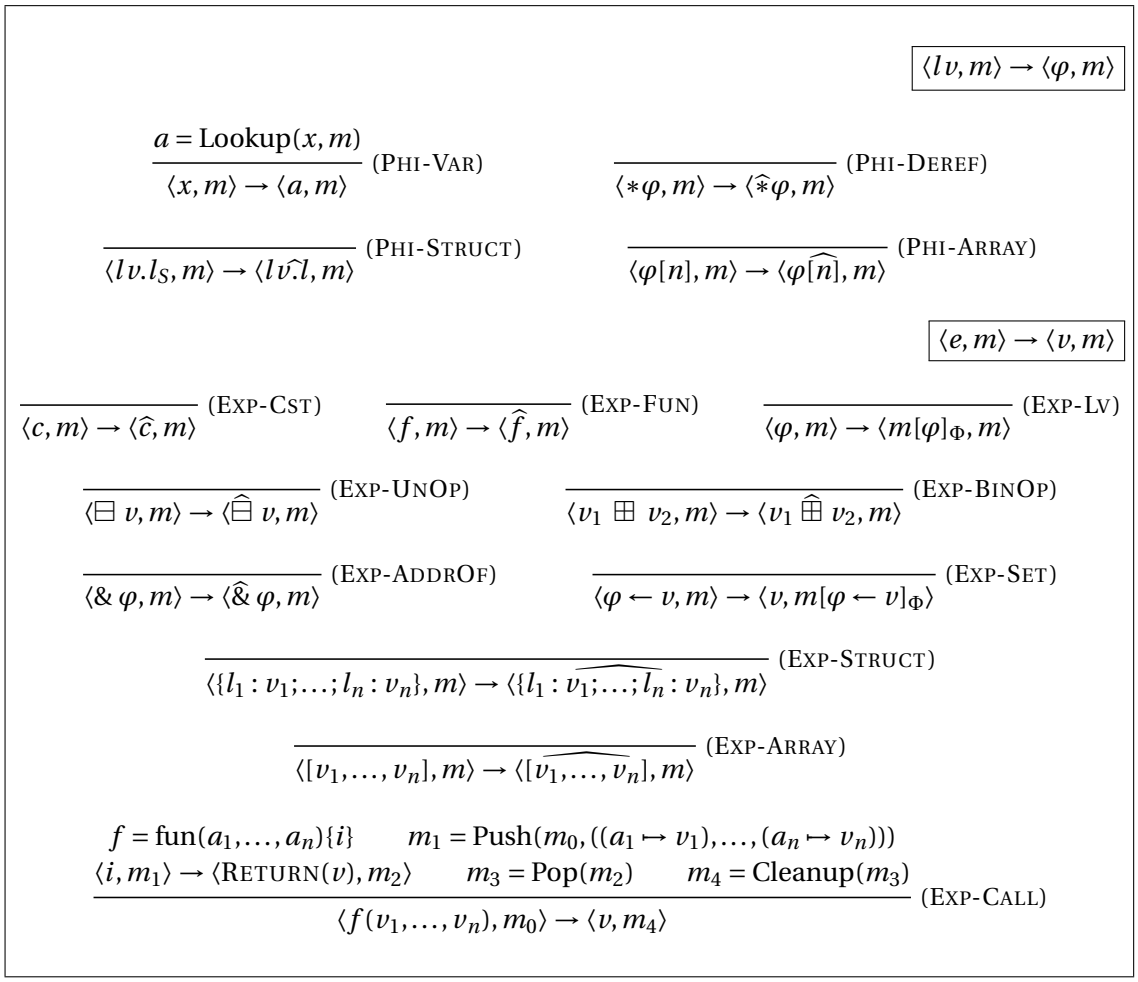


FIGURE B.3: Règles d'évaluation – left-values et expressions

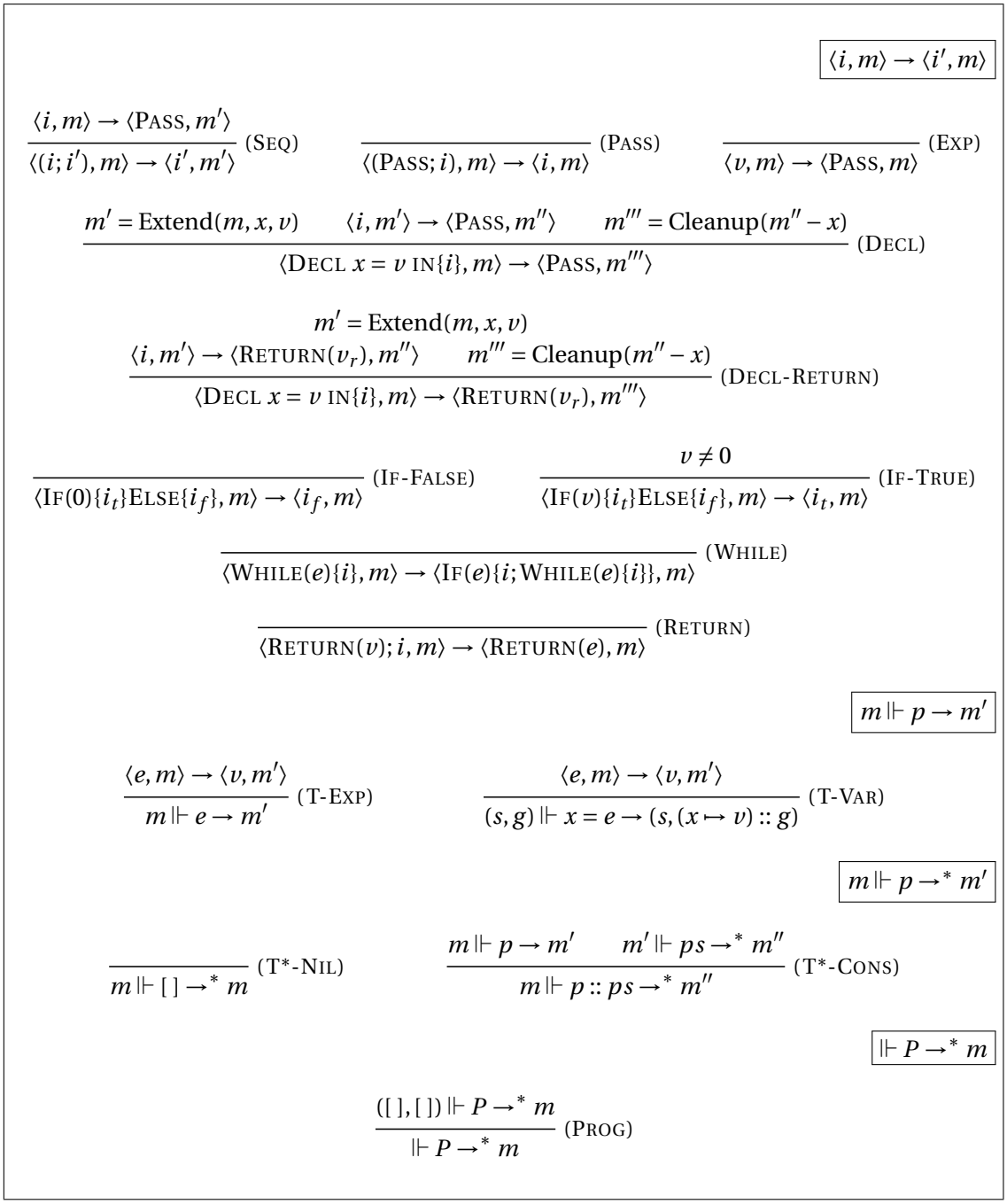


FIGURE B.4: Règles d'évaluation – instructions et phrases

$$\frac{\langle \varphi_d \leftarrow \varphi_s, m\rangle \rightarrow \langle \textsc{Pass}, m'\rangle}{\langle \text{copy\_from\_user}(\varphi_d, \widehat{\Diamond}\ \varphi_s), m\rangle \rightarrow \langle 0, m'\rangle}\ (\textsc{User-Get-OK})$$

$$\frac{\nexists\, \varphi_s, \varphi = \widehat{\Diamond}\ \varphi_s}{\langle \text{copy\_from\_user}(\varphi_d, \varphi), m\rangle \rightarrow \langle -1, m\rangle}\ (\textsc{User-Get-Err})$$

$$\frac{\langle \varphi_d \leftarrow \varphi_s, m\rangle \rightarrow \langle \textsc{Pass}, m'\rangle}{\langle \text{copy\_to\_user}(\widehat{\Diamond}\ \varphi_d, \varphi_s), m\rangle \rightarrow \langle 0, m'\rangle}\ (\textsc{User-Put-OK})$$

$$\frac{\nexists\, \varphi_d, \varphi = \widehat{\Diamond}\ \varphi_d}{\langle \text{copy\_to\_user}(\varphi, \varphi_s), m\rangle \rightarrow \langle -1, m\rangle}\ (\textsc{User-Put-Err})$$

FIGURE B.5: Règles d'évaluation – extensions noyau

ANNEXE C

RÈGLES DE TYPAGE

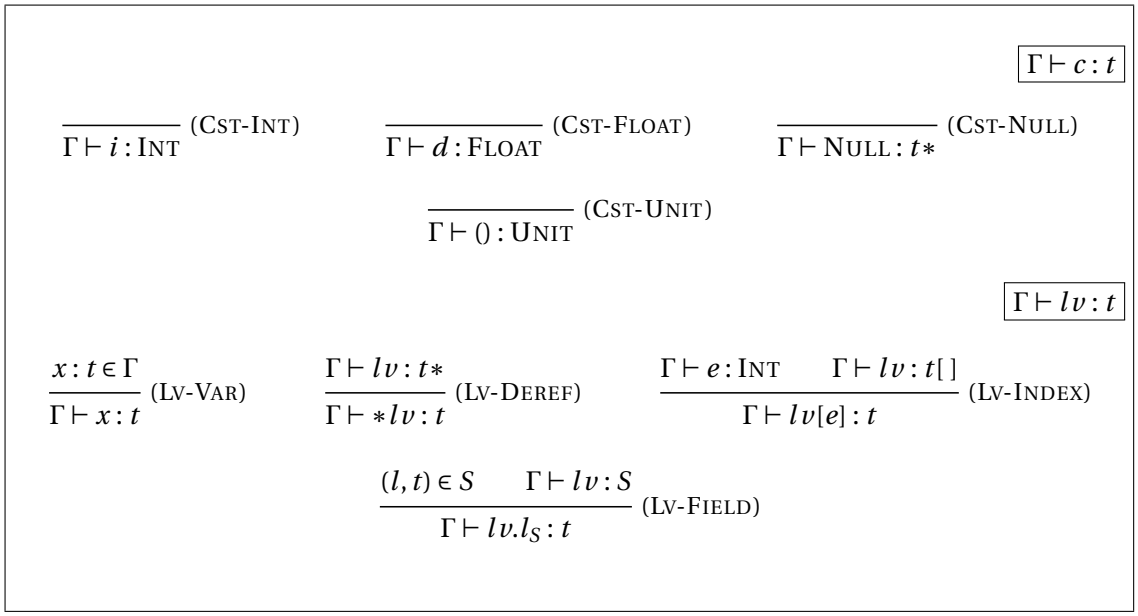


FIGURE C.1: Règles de typage – constantes et variables

$\boxed{\Gamma \vdash \boxminus e : t}$

$$\frac{\Gamma \vdash e : \text{INT}}{\Gamma \vdash +e : \text{INT}} \text{ (UNOP-PLUS-INT)} \qquad \frac{\Gamma \vdash e : \text{FLOAT}}{\Gamma \vdash +.e : \text{FLOAT}} \text{ (UNOP-PLUS-FLOAT)}$$

$$\frac{\Gamma \vdash e : \text{INT}}{\Gamma \vdash -e : \text{INT}} \text{ (UNOP-MINUS-INT)} \qquad \frac{\Gamma \vdash e : \text{FLOAT}}{\Gamma \vdash -.e : \text{FLOAT}} \text{ (UNOP-MINUS-FLOAT)}$$

$$\frac{\boxminus \in \{\sim, !\} \qquad \Gamma \vdash e : \text{INT}}{\Gamma \vdash \boxminus e : \text{INT}} \text{ (UNOP-NOT)}$$

$\boxed{\Gamma \vdash e_1 \boxplus e_2 : t}$

$$\frac{\boxplus \in \{+, -, \times, /, \&, |, \hat{}, \&\&, ||, \ll, \gg, \leq, \geq, <, >\} \qquad \Gamma \vdash e_1 : \text{INT} \qquad \Gamma \vdash e_2 : \text{INT}}{\Gamma \vdash e_1 \boxplus e_2 : \text{INT}} \text{ (OP-INT)}$$

$$\frac{\boxplus \in \{+., -., \times., /., \leq., \geq., <., >.\} \qquad \Gamma \vdash e_1 : \text{FLOAT} \qquad \Gamma \vdash e_2 : \text{FLOAT}}{\Gamma \vdash e_1 \boxplus e_2 : \text{FLOAT}} \text{ (OP-FLOAT)}$$

$$\frac{\boxplus \in \{=, \neq\} \qquad \Gamma \vdash e_1 : t \qquad \Gamma \vdash e_2 : t \qquad \text{EQ}(t)}{\Gamma \vdash e_1 \boxplus e_2 : \text{INT}} \text{ (OP-EQ)}$$

$$\frac{\boxplus \in \{+_p, -_p\} \qquad \Gamma \vdash e_1 : t* \qquad \Gamma \vdash e_2 : \text{INT}}{\Gamma \vdash e_1 \boxplus e_2 : t*} \text{ (PTR-ARITH)}$$

$\boxed{\text{EQ}(t)}$

$$\frac{t \in \{\text{INT}, \text{FLOAT}\}}{\text{EQ}(t)} \text{ (EQ-NUM)} \qquad \frac{}{\text{EQ}(t*)} \text{ (EQ-PTR)} \qquad \frac{\text{EQ}(t)}{\text{EQ}(t[\,])} \text{ (EQ-ARRAY)}$$

$$\frac{\forall i \in [1; n].\text{EQ}(t_i)}{\text{EQ}(\{l_1 : t_1; \ldots l_n : t_n\})} \text{ (EQ-STRUCT)}$$

FIGURE C.2: Règles de typage – opérateurs

$\boxed{\Gamma \vdash e : t}$

$$\frac{\Gamma \vdash lv : t}{\Gamma \vdash \&lv : t*}\ (\textsc{Addr}) \qquad \frac{\forall i \in [1;n], \Gamma \vdash e_i : t_i}{\Gamma \vdash \{l_1 : e_1; \ldots; l_n : e_n\} : \{l_1 : t_1; \ldots; l_n : t_n\}}\ (\textsc{Struct})$$

$$\frac{\Gamma \vdash e : (t_1, \ldots, t_n) \to t \qquad \forall i \in [1;n], \Gamma \vdash e_i : t_i}{\Gamma \vdash e(e_1, \ldots, e_n) : t}\ (\textsc{Call}) \qquad \frac{\Gamma \vdash lv : t \qquad \Gamma \vdash e : t}{\Gamma \vdash lv \leftarrow e : t}\ (\textsc{Set})$$

$$\frac{\forall i \in [1;n], \Gamma \vdash e_i : t}{\Gamma \vdash [e_1; \ldots; e_n] : t[\,]}\ (\textsc{Array}) \qquad \frac{\Gamma' = (\Gamma - \underline{R}), \vec{a} : \vec{t}, \underline{R} : t_r \qquad \Gamma' \vdash i}{\Gamma \vdash \text{fun}(\vec{a})\{i\} : \vec{t} \to t_r}\ (\textsc{Fun})$$

$\boxed{\Gamma \vdash i}$

$$\frac{}{\Gamma \vdash \textsc{Pass}}\ (\textsc{Pass}) \qquad \frac{\Gamma \vdash i_1 \qquad \Gamma \vdash i_2}{\Gamma \vdash i_1; i_2}\ (\textsc{Seq}) \qquad \frac{\Gamma \vdash e : t}{\Gamma \vdash e}\ (\textsc{Exp})$$

$$\frac{\Gamma \vdash e : t \qquad \Gamma, x : t \vdash i}{\Gamma \vdash \textsc{Decl}\ x = e\ \textsc{in}\{i\}}\ (\textsc{Decl}) \qquad \frac{\Gamma \vdash e : \textsc{Int} \qquad \Gamma \vdash i_1 \qquad \Gamma \vdash i_2}{\Gamma \vdash \textsc{If}(e)\{i_1\}\textsc{Else}\{i_2\}}\ (\textsc{If})$$

$$\frac{\Gamma \vdash e : \textsc{Int} \qquad \Gamma \vdash i}{\Gamma \vdash \textsc{While}(e)\{i\}}\ (\textsc{While}) \qquad \frac{\Gamma \vdash \underline{R} \leftarrow e}{\Gamma \vdash \textsc{Return}(e)}\ (\textsc{Return})$$

$\boxed{\Gamma \vdash p \to \Gamma'}$

$$\frac{\Gamma \vdash e : t}{\Gamma \vdash e \to \Gamma}\ (\textsc{T-Exp}) \qquad \frac{\Gamma \vdash e : t \qquad \Gamma' = (x, t), \Gamma}{\Gamma \vdash x = e \to \Gamma'}\ (\textsc{T-Var})$$

FIGURE C.3: Règles de typage

$$\frac{}{\Gamma \vdash \text{copy\_from\_user} : (t*, t@) \to \textsc{Int}}\ (\textsc{GetU}) \qquad \frac{}{\Gamma \vdash \text{copy\_to\_user} : (t@, t*) \to \textsc{Int}}\ (\textsc{PutU})$$

$$\frac{\Gamma \vdash lv : t}{\Gamma \vdash \&lv : t*}\ (\textsc{AddrOf-Kernel}) \qquad \frac{\Gamma \vdash e : t*}{\Gamma \vdash *e : t}\ (\textsc{Lv-Deref-Kernel})$$

$$\frac{\boxplus \in \{+_p, -_p\} \qquad \Gamma \vdash e_1 : t@ \qquad \Gamma \vdash e_2 : \textsc{Int}}{\Gamma \vdash e_1 \boxplus e_2 : t@}\ (\textsc{Ptr-Arith-User})$$

FIGURE C.4: Règles de typage – extensions noyau

ANNEXE D

PREUVES

D.1 Composition de lentilles

Démonstration. Pour prouver que $\mathscr{L}_1 \ggg \mathscr{L}_2 \in \text{LENS}_{A,C}$, il suffit de prouver les trois propriétés caractéristiques.

GetPut

$$\begin{aligned}
\text{put}_{\mathscr{L}}(\text{get}_{\mathscr{L}}(r), r) &= \text{put}_{\mathscr{L}}(\text{get}_{\mathscr{L}_2}(\text{get}_{\mathscr{L}_1}(r)), r) && \{ \text{ définition de get}_{\mathscr{L}} \ \} \\
&= \text{put}_{\mathscr{L}_1}(\text{put}_{\mathscr{L}_2}(\text{get}_{\mathscr{L}_2}(\text{get}_{\mathscr{L}_1}(r)), \text{get}_{\mathscr{L}_1}(r)), r) && \{ \text{ définition de put}_{\mathscr{L}} \ \} \\
&= \text{put}_{\mathscr{L}_1}(\text{get}_{\mathscr{L}_1}(r), r) && \{ \text{ GETPUT sur } \mathscr{L}_2 \ \} \\
&= r && \{ \text{ GETPUT sur } \mathscr{L}_1 \ \}
\end{aligned}$$

PutGet

$$\begin{aligned}
\text{get}_{\mathscr{L}}(\text{put}_{\mathscr{L}}(a, r)) &= \text{get}_{\mathscr{L}_2}(\text{get}_{\mathscr{L}_1}(\text{put}_{\mathscr{L}}(a, r))) && \{ \text{ définition de get}_{\mathscr{L}} \ \} \\
&= \text{get}_{\mathscr{L}_2}(\text{get}_{\mathscr{L}_1}(\text{put}_{\mathscr{L}_1}(\text{put}_{\mathscr{L}_2}(a, \text{get}_{\mathscr{L}_1}(r)), r))) && \{ \text{ définition de put}_{\mathscr{L}} \ \} \\
&= \text{get}_{\mathscr{L}_2}(\text{put}_{\mathscr{L}_2}(a, \text{get}_{\mathscr{L}_1}(r))) && \{ \text{ PUTGET sur } \mathscr{L}_1 \ \} \\
&= a && \{ \text{ PUTGET sur } \mathscr{L}_2 \ \}
\end{aligned}$$

PutPut

$$
\begin{aligned}
& \mathrm{put}_{\mathscr{L}}(a', \mathrm{put}_{\mathscr{L}}(a, r)) \\
= {} & \mathrm{put}_{\mathscr{L}}(a', \mathrm{put}_{\mathscr{L}_1}(\mathrm{put}_{\mathscr{L}_2}(a, \mathrm{get}_{\mathscr{L}_1}(r)), r)) \\
& \{ \text{définition de } \mathrm{put}_{\mathscr{L}} \} \\
= {} & \mathrm{put}_{\mathscr{L}_1}(\mathrm{put}_{\mathscr{L}_2}(a', \mathrm{get}_{\mathscr{L}_1}(\mathrm{put}_{\mathscr{L}_1}(\mathrm{put}_{\mathscr{L}_2}(a, \mathrm{get}_{\mathscr{L}_1}(r)), r))), \mathrm{put}_{\mathscr{L}_1}(\mathrm{put}_{\mathscr{L}_2}(a, \mathrm{get}_{\mathscr{L}_1}(r)), r)) \\
& \{ \text{définition de } \mathrm{put}_{\mathscr{L}} \} \\
= {} & \mathrm{put}_{\mathscr{L}_1}(\mathrm{put}_{\mathscr{L}_2}(a', \mathrm{put}_{\mathscr{L}_2}(a, \mathrm{get}_{\mathscr{L}_1}(r))), \mathrm{put}_{\mathscr{L}_1}(\mathrm{put}_{\mathscr{L}_2}(a, \mathrm{get}_{\mathscr{L}_1}(r)), r)) \\
& \{ \textsc{GetPut} \text{ sur } \mathscr{L}_1 \} \\
= {} & \mathrm{put}_{\mathscr{L}_1}(\mathrm{put}_{\mathscr{L}_2}(a', \mathrm{get}_{\mathscr{L}_1}(r)), \mathrm{put}_{\mathscr{L}_1}(\mathrm{put}_{\mathscr{L}_2}(a, \mathrm{get}_{\mathscr{L}_1}(r)), r)) \\
& \{ \textsc{PutPut} \text{ sur } \mathscr{L}_2 \} \\
= {} & \mathrm{put}_{\mathscr{L}_1}(\mathrm{put}_{\mathscr{L}_2}(a', \mathrm{get}_{\mathscr{L}_1}(r)), r) \\
& \{ \textsc{PutPut} \text{ sur } \mathscr{L}_1 \} \\
= {} & \mathrm{put}_{\mathscr{L}}(a', r) \\
& \{ \text{définition de } \ggg \}
\end{aligned}
$$

□

D.2 Progrès

On prouve de manière récursive mutuelle les théorèmes 5.1 et 5.2

Démonstration. On procède par induction sur la dérivation de $\Gamma \vdash e : t$, et plus précisément sur la dernière règle utilisée.

CST-INT : e est alors de la forme n, qui est une valeur.

CST-FLOAT : e est alors de la forme d, qui est une valeur.

CST-NULL : e est alors égale à NULL, qui est une valeur.

CST-UNIT : e est alors égale à $()$, qui est une valeur.

LV-VAR : Puisque $(x, t) \in \Gamma$ et $\Gamma \vDash m$, il existe $(x \mapsto v) \in m$. La règle d'évaluation PHI-VAR s'applique donc.

LV-DEREF : Appliquons l'hypothèse de récurrence à $l\,v$ (vue en tant qu'expression).

- $l\,v = v$. Puisque $\Gamma \vdash v : t*$, on déduit du lemme 5.2 que $v = \varphi$ ou $v = \mathrm{NULL}$.
  Dans le premier cas, la règle PHI-DEREF s'applique : $\langle e, m\rangle \to \langle \hat{*}\varphi, m\rangle$. Dans le second, puisque $\langle *\mathrm{NULL}, m\rangle \to \Omega_{ptr}$, on a $\langle e, m\rangle \to \Omega_{ptr}$.
- $\langle l\,v, m\rangle \to \langle e', m'\rangle$. De CTX avec $C = *\bullet$, on obtient $\langle e, m\rangle \to \langle *e', m'\rangle$.
- $\langle l\,v, m\rangle \to \Omega$. En appliquant EVAL-ERR avec $C = *\bullet$, on obtient $\langle e, m\rangle \to \Omega$.

LV-INDEX : De même, on applique l'hypothèse de récurrence à $l\,v$.

- $l\,v = v$.
  Comme $\Gamma \vdash v : t[]$, on déduit du lemme 5.2 que $v = [v_1; \ldots; v_p]$. Appliquons l'hypothèse de récurrence à e.
  - $e = v'$. Puisque $\Gamma \vdash e :$ INT, on réapplique le lemme 5.2 et $v' = n$. D'après PHI-ARRAY, $\langle l\,v[e], m\rangle \to \langle [v_1; \ldots; v_p]\widehat{[n]}, m\rangle$. Deux cas sont à distinguer : si $n \in [0; p-1]$, la partie droite vaut v_{n+1} et donc $\langle l\,v[e], m\rangle \to \langle v_{n+1}, m\rangle$. Sinon elle vaut Ω_{array} et $\langle l\,v[e], m\rangle \to \Omega_{array}$ par EXP-ERR.
  - $\langle e, m\rangle \to \langle e', m'\rangle$. En appliquant CTX avec $C = v[\bullet]$, on en déduit
  - $\langle l\,v[e], m\rangle \to \langle l\,v[e'], m'\rangle$.
  - $\langle e, m\rangle \to \Omega$. Avec EVAL-ERR sous ce même contexte, $\langle l\,v[e], m\rangle \to \Omega$
- $\langle l\,v, m\rangle \to \langle e', m'\rangle$. On applique alors CTX avec $C = \bullet[e]$, et $\langle l\,v[e], m\rangle \to \langle e'[e], m'\rangle$.
- $\langle l\,v, m\rangle \to \Omega$. Toujours avec $C = \bullet[e]$, de EVAL-ERR il vient $\langle l\,v[e], m\rangle \to \Omega$.

LV-FIELD : On applique l'hypothèse de récurrence du théorème 5.2 à $l\,v$.

- $l\,v = \varphi$ Alors PHI-STRUCT s'applique. Puisque $(l, t) \in S$, l'accès au champ l ne provoque pas d'erreur Ω_{field}. Donc $\langle e, m\rangle \to \langle \varphi[l], m\rangle$.
- $\langle l\,v, m\rangle \to \langle l\,v', m'\rangle$ En appliquant CTX avec $C = \bullet.l_S$, il vient $\langle l\,v, m\rangle \to \langle l\,v', m'\rangle$.
- $\langle l\,v, m\rangle \to \Omega$ En appliquant EVAL-ERR avec $C = \bullet.l_S$, on a $\langle l\,v, m\rangle \to \Omega$.

OP-INT : Cela implique que $e = e_1 \boxplus e_2$. Par le lemme 5.1, on en déduit que $\Gamma \vdash e_1 :$ INT et $\Gamma \vdash e_2 :$ INT.
Appliquons l'hypothèse de récurrence sur e_1. Trois cas peuvent se produire.

- $e_1 = v_1$. On a alors $\langle e_1, m\rangle \to \langle v_1, m'\rangle$ avec $m' = m$.
  On applique l'hypothèse de récurrence à e_2.
  - $e_2 = v_2$: alors $\langle e_2, m'\rangle \to \langle v_2, m''\rangle$ avec $m'' = m$. On peut alors appliquer EXP-BINOP, sauf dans le cas d'une division par zéro ($\boxplus \in \{/; \%; /.\}$ et $v_2 = 0$) où alors $v_1 \,\widehat{\boxplus}\, v_2 = \Omega_{div}$. Dans ce cas, on a alors par EXP-ERR $\langle e, m\rangle \to \Omega_{div}$.
  - $\exists (e_2', m''), \langle e_2, m'\rangle \to \langle e_2', m''\rangle$.
    En appliquant CTX avec $C = v_1 \boxplus \bullet$, on en déduit $\langle v_1 \boxplus e_2, m'\rangle \to \langle v_1 \boxplus e_2', m''\rangle$ soit $\langle e, m\rangle \to \langle v_1 \boxplus e_2', m''\rangle$.
  - $\langle e_2, m'\rangle \to \Omega$. De EVAL-ERR avec $C = v_1 \boxplus \bullet$ vient alors $\langle e, m\rangle \to \Omega$.
- $\exists (e_1', m'), \langle e_1, m\rangle \to \langle e_1', m'\rangle$. En appliquant CTX avec $C = \bullet \boxplus e_2$, on obtient $\langle e_1 \boxplus e_2, m\rangle \to \langle e_1' \boxplus e_2, m'\rangle$, ou $\langle e, m\rangle \to \langle e_1' \boxplus e_2, m'\rangle$.
- $\langle e_1, m\rangle \to \Omega$. D'après EVAL-ERR avec $C = \bullet \boxplus e_2$, on a $\langle e, m\rangle \to \Omega$.

OP-FLOAT : Ce cas est similaire à OP-INT.

OP-EQ : Ce cas est similaire à OP-INT.

OP-COMPARABLE : Ce cas est similaire à OP-INT.

UNOP-PLUS-INT : Alors $e = + e_1$. En appliquant l'hypothèse d'induction sur e_1 :

- soit $e_1 = v_1$. Alors en appliquant EXP-UNOP, $\langle + v_1, m\rangle \to \langle \widehat{+} v_1, m\rangle$, c'est-à-dire en posant $v = \widehat{+} v_1$, $\langle e, m\rangle \to \langle v, m\rangle$.
- soit $\exists e_1', m', \langle e_1, m\rangle \to \langle e_1', m'\rangle$. Alors en appliquant CTX avec $C = + \bullet$, on obtient $\langle e, m\rangle \to \langle e_1', m'\rangle$.
- soit $\langle e_1, m\rangle \to \Omega$. De EVAL-ERR avec $C = + \bullet$ il vient$\langle e, m\rangle \to \Omega$.

UNOP-PLUS-FLOAT : Ce cas est similaire à UNOP-PLUS-INT.

UNOP-MINUS-INT : Ce cas est similaire à UNOP-PLUS-INT.

UNOP-MINUS-FLOAT : Ce cas est similaire à UNOP-PLUS-INT.

UNOP-NOT : Ce cas est similaire à UNOP-PLUS-INT.

PTR-ARITH : Le schéma est similaire au cas OP-INT. Le seul cas intéressant arrive lorsque e_1 et e_2 sont des valeurs. D'après le lemme 5.2 :

- $e_1 = \text{NULL}$ ou $e_1 = \varphi$
- $e_2 = n$

D'après EXP-BINOP, $\langle e, m\rangle \to \langle e_1 \mathbin{\widehat{\boxplus}} n, m\rangle$.

On se réfère ensuite à la définition de $\widehat{\boxplus}$ (page 56) : si e_1 est de la forme $\varphi[m]$, alors $e_1 \mathbin{\widehat{\boxplus}} n = \varphi[m+n]$. Donc $\langle e, m\rangle \to \langle \varphi[m+n], m\rangle$.

Dans les autres cas ($e_1 = \text{NULL}$ ou $e_1 = \varphi$ avec φ pas de la forme $\varphi'[m]$), on a $e_1 \mathbin{\widehat{\boxplus}} n = \Omega_{ptr}$. Donc d'après EXP-ERR, $\langle e, m\rangle \to \Omega_{ptr}$.

ADDR :

$$\left(\frac{\Gamma \vdash l\,v : t}{\Gamma \vdash \&l\,v : t*}\ (\text{ADDR}) \right)$$

On applique l'hypothèse de récurrence du théorème 5.2 à $l\,v$.

Les cas d'évaluation et d'erreur sont traités en appliquant respectivement CTX et EVAL-ERR avec $C = \&\bullet$. On détaille le cas où $l\,v = \varphi$.

$$\left(\frac{}{\langle \&\,\varphi, m\rangle \to \langle \widehat{\&}\,\varphi, m\rangle}\ (\text{EXP-ADDROF}) \right)$$

SET :

$$\left(\frac{\Gamma \vdash l\,v : t \qquad \Gamma \vdash e : t}{\Gamma \vdash l\,v \leftarrow e : t}\ (\text{SET}) \right)$$

ARRAY : On va appliquer l'hypothèse de récurrence à e_1, puis si $e_1 = v_1$, on l'applique à e_2, etc. Alors on se retrouve dans un des cas suivants :

- $\exists p \in [1;n], e'_p, m : e_1 = v_1, \ldots, e_{p-1} = v_{p-1}, \langle e_p, m\rangle \to \langle e'_p, m'\rangle$. Alors on peut appliquer CTX avec $C = [v_1, \ldots, v_{p-1}, \bullet, e_{p+1}, \ldots, e_n]$.
- $\exists p \in [1;n], \Omega : e_1 = v_1, \ldots, e_{p-1} = v_{p-1}, \langle e_p, m\rangle \to \Omega$. Dans ce cas EVAL-ERR est applicable avec ce même C.
- $e_1 = v_1, \ldots, e_n = v_n$. Alors on peut appliquer EXP-ARRAY en construisant un tableau.

STRUCT : Le schéma de preuve est similaire au cas ARRAY. En cas de pas d'évaluation ou d'erreur, on utilise le contexte $C = \{l_1 : v_1, \ldots, l_{p-1} : v_{p-1}, \bullet, l_{p+1} : e_{p+1}, \ldots, l_n : e_n\}$; et dans le cas où toutes les expressions sont évaluées, on applique EXP-STRUCT.

CALL : On commence par appliquer l'hypothèse de récurrence à e. Dans le cas d'un pas d'évaluation ou d'erreur, on applique respectivement CTX ou EVAL-ERR avec $C = \bullet(e_1, \ldots, e_n)$. Reste le cas où e est une valeur : d'après le lemme 5.2, e est de la forme $f = \text{fun}(\vec{a})\{i\}$.

Ensuite, appliquons le même schéma que pour ARRAY. En cas de pas d'évaluation ou d'erreur, on utilise CTX ou EVAL-ERR avec $C = f(v_1, \ldots, v_{p-1}, \bullet, e_{p+1}, \ldots, e_n)$. Le seul cas restant est celui où l'expression considérée a pour forme $f(v_1, \ldots, v_n)$ avec $f = \text{fun}(\vec{a})\{i\}$: EXP-CALL s'applique alors.

FUN : Ce cas est direct : la règle EXP-FUN s'applique.

□

D.3 Préservation

(théorème 5.3)

Démonstration. On procède par induction sur la forme de l'expression e.

Constante c On détaille par exemple le cas d'une constante entière.

D'une part, $\Gamma \vdash e : t$ donc d'après le lemme 5.2, $t = \text{INT}$.

D'autre part, la seule règle d'évaluation possible est EXP-CST qui évalue en une constante entière.

Accès mémoire $l\,v$

Opération unaire $\boxminus\, e$

Opération binaire $e \boxplus e$

Pointeur $\&l\,v$

Affectation $l\,v \leftarrow e$

Structure $\{l_1 : e_1; ; l_n : e_n\}$

Tableau $[e_1; \ldots; e_n]$

Fonction f

Appel de fonction $e(e_1, \ldots, e_n)$

□

D.4 Progrès pour les types qualifiés

(théorème 6.1)

D.5 Préservation pour les types qualifiés

(théorème 6.2)

TABLE DES FIGURES

Liste des définitions

Liste des théorèmes et propriétés

RÉFÉRENCES WEB

[1] The Objective Caml system, documentation and user's manual – release 3.12
http://caml.inria.fr/pub/docs/manual-ocaml/

[2] Haskell Programming Language – Official Website
http://www.haskell.org/

[3] Python Programming Language – Official Website
http://www.python.org/

[4] Perl Programming Language – Official Website
http://www.perl.org/

[5] Sparse - a Semantic Parser for C
https://sparse.wiki.kernel.org/index.php/Main_Page

[6] CIL - C Intermediate Language
http://kerneis.github.com/cil/

[7] The C - - language
http://www.cminusminus.org/

[8] Penjili project
http://www.penjili.org/

[9] The Rust Programming Language
http://www.rust-lang.org/

BIBLIOGRAPHIE

[AB07] Andrew W. Appel and Sandrine Blazy. Separation logic for small-step Cminor (extended version). Research report 6138, INRIA, 2007. 29 pages. 7

[ABD+07] Alex Aiken, Suhabe Bugrara, Isil Dillig, Thomas Dillig, Brian Hackett, and Peter Hawkins. An overview of the saturn project. In *Proceedings of the 7th ACM SIGPLAN-SIGSOFT workshop on Program analysis for software tools and engineering*, PASTE '07, pages 43–48, New York, NY, USA, 2007. ACM. 25

[AH07] Xavier Allamigeon and Charles Hymans. Analyse statique par interprétation abstraite. In Eric Filiol, editor, *5ème Symposium sur la Sécurité des Technologies de l'Information et des Communications (SSTIC'07)*, Rennes, France, June 2007. 22

[BA08] S. Bugrara and A. Aiken. Verifying the Safety of User Pointer Dereferences. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 325–338, 2008. 25

[BBC+10] Al Bessey, Ken Block, Ben Chelf, Andy Chou, Bryan Fulton, Seth Hallem, Charles Henri-Gros, Asya Kamsky, Scott McPeak, and Dawson Engler. A few billion lines of code later : using static analysis to find bugs in the real world. *Commun. ACM*, 53(2) :66–75, February 2010. 24

[BC05] Daniel P. Bovet and Marco Cesati. *Understanding the Linux Kernel, Third Edition*. O'Reilly Media, third edition edition, November 2005. 104

[BDH+09] Julien Brunel, Damien Doligez, René Rydhof Hansen, Julia L. Lawall, and Gilles Muller. A foundation for flow-based program matching using temporal logic and model checking. In *The 36th Annual ACM SIGPLAN - SIGACT Symposium on Principles of Programming Languages*, pages 114–126, Savannah, GA, USA, January 2009. 22

[BDL06] Sandrine Blazy, Zaynah Dargaye, and Xavier Leroy. Formal verification of a C compiler front-end. In *FM 2006 : Int. Symp. on Formal Methods*, volume 4085 of *Lecture Notes in Computer Science*, pages 460–475. Springer, 2006. 7

[BDN09] Ana Bove, Peter Dybjer, and Ulf Norell. A brief overview of agda — a functional language with dependent types. In *Proceedings of the 22nd International Conference on Theorem Proving in Higher Order Logics*, TPHOLs '09, pages 73–78, Berlin, Heidelberg, 2009. Springer-Verlag. 27

[BLS05] Mike Barnett, K. Rustan M. Leino, and Wolfram Schulte. The spec# programming system : an overview. In *Proceedings of the 2004 international conference on Construction and Analysis of Safe, Secure, and Interoperable Smart Devices*, CASSIS'04, pages 49–69, Berlin, Heidelberg, 2005. Springer-Verlag. 27

[CC77] Patrick Cousot and Radhia Cousot. Abstract interpretation : a unified lattice model for static analysis of programs by construction or approximation of fixpoints. In *POPL '77 : Proceedings of the 4th ACM SIGACT-SIGPLAN symposium on Principles of Programming Languages*, pages 238–252, New York, NY, USA, 1977. ACM. 22, 23

[CC92] P. Cousot and R. Cousot. Abstract interpretation and application to logic programs. *Journal of Logic Programming*, 13(2–3) :103–179, 1992. (The editor of Journal of Logic Programming has mistakenly published the unreadable galley proof. For a correct version of this paper, see `http://www.di.ens.fr/~cousot`.). 22

[CCF+05] Patrick Cousot, Radhia Cousot, Jérôme Feret, Laurent Mauborgne, Antoine Miné, David Monniaux, and Xavier Rival. The ASTREÉ analyzer. In Shmuel Sagiv, editor, *ESOP*, volume 3444 of *Lecture Notes in Computer Science*, pages 21–30. Springer, 2005. 24

[CCF+09] Patrick Cousot, Radhia Cousot, Jérôme Feret, Laurent Mauborgne, Antoine Miné, and Xavier Rival. Why does Astrée scale up ? *Formal Methods in System Design*, 35(3) :229–264, 2009. 24

[DDMPn10] Javier De Dios, Manuel Montenegro, and Ricardo Peña. Certified absence of dangling pointers in a language with explicit deallocation. In *Proceedings of the 8th international conference on Integrated formal methods*, IFM'10, pages 305–319, Berlin, Heidelberg, 2010. Springer-Verlag. 117

[DM82] Luis Damas and Robin Milner. Principal type-schemes for functional programs. In *Proceedings of the 9th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, POPL '82, pages 207–212, New York, NY, USA, 1982. ACM. 91

[DRS00] Nurit Dor, Michael Rodeh, and Mooly Sagiv. CSSV : Towards a realistic tool for statically detecting all buffer overflows in C, 2000. 27

[EH94] Ana Erosa and Laurie J. Hendren. Taming control flow : A structured approach to eliminating goto statements. In *In Proceedings of 1994 IEEE International Conference on Computer Languages*, pages 229–240. IEEE Computer Society Press, 1994. 90

[FFA99] Jeffrey S. Foster, Manuel Fähndrich, and Alexander Aiken. A theory of type qualifiers. In *Programming language design and implementation*, PLDI '99, pages 192–203, 1999. 25

[FGM+07] J. Nathan Foster, Michael B. Greenwald, Jonathan T. Moore, Benjamin C. Pierce, and Alan Schmitt. Combinators for bidirectional tree transformations : A linguistic approach to the view-update problem. *ACM Trans. Program. Lang. Syst.*, 29(3), May 2007. 39

[FJKA06] Jeffrey S. Foster, Robert Johnson, John Kodumal, and Alex Aiken. Flow-insensitive type qualifiers. *ACM Trans. Program. Lang. Syst.*, 28 :1035–1087, November 2006. 24

[Flo67] Robert W. Floyd. Assigning Meanings to Programs. In J. T. Schwartz, editor, *Proceedings of a Symposium on Applied Mathematics*, volume 19 of *Mathematical Aspects of Computer Science*, pages 19–31, Providence, 1967. American Mathematical Society. 26

[FTA02] Jeffrey S. Foster, Tachio Terauchi, and Alex Aiken. Flow-sensitive type qualifiers. In *Proceedings of the ACM SIGPLAN 2002 Conference on Programming language design and implementation*, PLDI '02, pages 1–12, New York, NY, USA, 2002. ACM. 25

[GGTZ07] Stephane Gaubert, Eric Goubault, Ankur Taly, and Sarah Zennou. Static analysis by policy iteration on relational domains. In Rocco Nicola, editor, *Programming Languages and Systems*, volume 4421 of *Lecture Notes in Computer Science*, pages 237–252. Springer Berlin Heidelberg, 2007. 23

[GMJ+02] Dan Grossman, Greg Morrisett, Trevor Jim, Michael Hicks, Yanling Wang, and James Cheney. Region-based memory management in Cyclone. *SIGPLAN Not.*, 37(5) :282–293, May 2002. 26

[Gon05] Georges Gonthier. A computer-checked proof of the Four Colour Theorem. 2005. 27

[Gor04] Mel Gorman. *Understanding the Linux Virtual Memory Manager*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2004. 104, 117

[Gra92] Philippe Granger. Improving the results of static analyses programs by local decreasing iteration. In *Proceedings of the 12th Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 68–79, London, UK, UK, 1992. Springer-Verlag. 23

[Har88] Norm Hardy. The confused deputy (or why capabilities might have been invented). *ACM Operating Systems Review*, 22(4) :36–38, October 1988. 4, 19

[HJLT05] Thomas Hallgren, Mark P. Jones, Rebekah Leslie, and Andrew Tolmach. A principled approach to operating system construction in haskell. In *Proceedings of the tenth ACM SIGPLAN international conference on Functional programming*, ICFP '05, pages 116–128, New York, NY, USA, 2005. ACM. 6

[HL08] Charles Hymans and Olivier Levillain. Newspeak, Doubleplussimple Minilang for Goodthinkful Static Analysis of C. Technical Note 2008-IW-SE-00010-1, EADS IW/SE, 2008. 8, 89, 90

[HLA+05] G. Hunt, J. Larus, M. Abadi, M. Aiken, P. Barham, M. Fahndrich, C. Hawblitzel, O. Hodson, S. Levi, N. Murphy, B. Steensgaard, D. Tarditi, T. Wobber, and B. Zill. An Overview of the Singularity Project. Technical report, Microsoft Research, 2005. 6

[Hoa69] C. A. R. Hoare. An axiomatic basis for computer programming. *Commun. ACM*, 12(10) :576–580, October 1969. 26

[Int10] Intel, Santa Clara, CA, USA. *Intel® 64 and IA-32 Architectures Software Developer's Manual*, June 2010. 13, 105

[ISO99] ISO. The ANSI C standard (C99). Technical Report WG14 N1124, ISO/IEC, 1999. 6

[JMG+02] Trevor Jim, J. Greg Morrisett, Dan Grossman, Michael W. Hicks, James Cheney, and Yanling Wang. Cyclone : A safe dialect of c. In *Proceedings of the General Track of the annual conference on USENIX Annual Technical Conference*, ATEC '02, pages 275–288, Berkeley, CA, USA, 2002. USENIX Association. 25

[JW04] Robert Johnson and David Wagner. Finding user/kernel pointer bugs with type inference. In *USENIX Security Symposium*, pages 119–134, 2004. 25, 83

[KcS07] Oleg Kiselyov and Chung chieh Shan. Lightweight static capabilities. *Electr. Notes Theor. Comput. Sci.*, 174(7) :79–104, 2007. 24

[Ker81] Brian W. Kernighan. Why Pascal is not my favorite programming language. Technical report, AT&T Bell Laboratories, April 1981. 5

[KR88] Brian W. Kernighan and Dennis M. Ritchie. *The C Programming Language Second Edition*. Prentice-Hall, Inc., 1988. 6

[LA04] Chris Lattner and Vikram Adve. LLVM : A Compilation Framework for Lifelong Program Analysis & Transformation. In *Proceedings of the 2004 International Symposium on Code Generation and Optimization (CGO'04)*, Palo Alto, California, Mar 2004. 7

[LBR99] Gary T. Leavens, Albert L. Baker, and Clyde Ruby. Jml : A notation for detailed design, 1999. 27

[LZ06] Peng Li and Steve Zdancewic. Encoding information flow in Haskell. In *Proceedings of the 19th IEEE Workshop on Computer Security Foundations (CSFW '06)*, Washington, DC, USA, 2006. IEEE Computer Society. 24

[Mau04] Laurent Mauborgne. ASTRÉE : Verification of absence of run-time error. In René Jacquart, editor, *Building the information Society (18th IFIP World Computer Congress)*, pages 384–392. The International Federation for Information Processing, Kluwer Academic Publishers, Aug 2004. 24

[Mer03] J. Merrill. GENERIC and GIMPLE : a new tree representation for entire functions. In *GCC developers summit 2003*, pages 171–180, 2003. 7

[Mil78] Robin Milner. A theory of type polymorphism in programming. *Journal of Computer and System Sciences*, 17(3) :348–375, December 1978. 5

[MMR+13] Anil Madhavapeddy, Richard Mortier, Charalampos Rotsos, David Scott, Balraj Singh, Thomas Gazagnaire, Steven Smith, Steven Hand, and Jon Crowcroft. Unikernels : library operating systems for the cloud. In *Proceedings of the eighteenth international conference on Architectural support for programming languages and operating systems*, ASPLOS '13, pages 461–472, New York, NY, USA, 2013. ACM. 6

[NCH+05] George C. Necula, Jeremy Condit, Matthew Harren, Scott McPeak, and Westley Weimer. Ccured : type-safe retrofitting of legacy software. *ACM Trans. Program. Lang. Syst.*, 27(3) :477–526, May 2005. 25

[New00] Tim Newsham. Format string attacks. *Phrack*, 2000. 25

[NMRW02] George C. Necula, Scott McPeak, Shree Prakash Rahul, and Westley Weimer. Cil : Intermediate language and tools for analysis and transformation of c programs. In *Proceedings of the 11th International Conference on Compiler Construction*, CC '02, pages 213–228, London, UK, UK, 2002. Springer-Verlag. 7

[NPW02] Tobias Nipkow, Lawrence C. Paulson, and Markus Wenzel. *Isabelle/HOL — A Proof Assistant for Higher-Order Logic*, volume 2283 of *LNCS*. Springer, 2002. 27

[oEE08] Institute of Electrical and Electronics Engineers. IEEE Standard for Floating-Point Arithmetic. Technical report, Microprocessor Standards Committee of the IEEE Computer Society, 3 Park Avenue, New York, NY 10016-5997, USA, August 2008. 45

[One96] Aleph One. Smashing the stack for fun and profit. *Phrack*, 1996. 13

[Pie02] Benjamin C. Pierce. *Types and Programming Languages*. MIT Press, 2002. 5

[PJNO97] Simon L. Peyton Jones, Thomas Nordin, and Dino Oliva. C- : A portable assembly language. In Chris Clack, Kevin Hammond, and Antony J. T. Davie, editors, *Implementation of Functional Languages*, volume 1467 of *Lecture Notes in Computer Science*, pages 1–19. Springer, 1997. 7

[PTS+11] Nicolas Palix, Gaël Thomas, Suman Saha, Christophe Calvès, Julia Lawall, and Gilles Muller. Faults in Linux : Ten years later. In *Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2011)*, Newport Beach, CA, USA, March 2011. 22

[Ric53] H. G. Rice. Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society*, 74(2) :pp. 358–366, 1953. 21

[Spe05] Brad Spengler. grsecurity 2.1.0 and kernel vulnerabilities. *Linux Weekly News*, 2005. 22

[SRH95] Mooly Sagiv, Thomas Reps, and Susan Horwitz. Precise interprocedural dataflow analysis with applications to constant propagation, 1995. 23

[Sta11] Basile Starynkevitch. Melt - a translated domain specific language embedded in the gcc compiler. In Olivier Danvy and Chung chieh Shan, editors, *DSL*, volume 66 of *EPTCS*, pages 118–142, 2011. 7

[STFW01] Umesh Shankar, Kunal Talwar, Jeffrey S. Foster, and David Wagner. Detecting format string vulnerabilities with type qualifiers. In *SSYM'01 : Proceedings of the 10th conference on USENIX Security Symposium*, page 16, Berkeley, CA, USA, 2001. USENIX Association. 25

[SY86] R E Strom and S Yemini. Typestate : A programming language concept for enhancing software reliability. *IEEE Trans. Softw. Eng.*, 12(1) :157–171, January 1986. 25

[Tan07] Andrew S. Tanenbaum. *Modern Operating Systems*. Prentice Hall Press, Upper Saddle River, NJ, USA, 3rd edition, 2007. 2

[The04] The Coq Development Team. *The Coq Proof Assistant Reference Manual – Version V8.0*, April 2004. `http://coq.inria.fr`. 27

[TJ92] Jean-Pierre Talpin and Pierre Jouvelot. Polymorphic type, region and effect inference. *Journal of Functional Programming*, 2 :245–271, 1992. 26

[TT93] Mads Tofte and Jean-Pierre Talpin. A theory of stack allocation in polymorphically typed languages. Technical report, 1993. 26

[TT94] Mads Tofte and Jean-Pierre Talpin. Implementation of the typed call-by-value λ-calculus using a stack of regions. In *Proceedings of the 21st ACM SIGPLAN-SIGACT symposium on Principles of programming languages*, POPL '94, pages 188–201, New York, NY, USA, 1994. ACM. 26

[TTL] Linus Torvalds, Josh Triplett, and Christopher Li. Sparse - a semantic parser for C. `https://sparse.wiki.kernel.org`. 24

[VB04] Arnaud Venet and Guillaume Brat. Precise and efficient static array bound checking for large embedded c programs. In *Proceedings of the ACM SIGPLAN 2004 conference on Programming language design and implementation*, PLDI '04, pages 231–242, New York, NY, USA, 2004. ACM. 24

[vL11] Twan van Laarhoven. Lenses : viewing and updating data structures in Haskell. http://www.twanvl.nl/files/lenses-talk-2011-05-17.pdf, May 2011. 39

Résumé

Les noyaux de système d'exploitation manipulent des données fournies par les programmes utilisateur via certaines interfaces, comme les appels système. Si celles-ci sont manipulées sans prendre une attention particulière, une faille de sécurité connue sous le nom de *Confused Deputy Problem* peut amener à des fuites de données confidentielles ou l'elévation de privilège d'un attaquant.

Le but de cette thèse est d'utiliser des techniques de typage statique afin de détecter les manipulations dangereuses de pointeurs contrôlés par l'espace utilisateur.

De nombreux systèmes d'exploitation, dont le noyau Linux sur lequel nous nous basons, sont écrits dans le langage C. Afin de pouvoir appliquer les techniques de typage statique à ce langage, on commence par isoler un sous-langage sûr nommé SAFESPEAK, pour lequel on définit une sémantique opérationnelle et un système de types. Les propriétés classiques de sûreté du typage sont établies.

Dans un second temps on étend SAFESPEAK pour intégrer la notion de valeur provenant de l'espace utilisateur. On montre que l'extension triviale du système de types brise la sûreté du typage, et que l'on peut réétablir celle-ci en donnant un type particulier aux pointeurs dont la valeur est contrôlée par l'espace utilisateur, forçant leur déférencement à se faire au sein d'un ensemble limité de fonctions sûres. Un cas d'étude est déroulé pour montrer que cette technique permet de détecter un bug qui a frappé un pilote de carte graphique AMD dans le noyau Linux.

Ces travaux ouvrent des perspectives sur deux points. D'une part, il est possible d'appliquer cette analyse à de plus grandes parties du noyau Linux. Les diverses interfaces qui exposent des données contrôlées par l'espace utilisateur peuvent bénéficier de cette vérification statique. Ceci est d'autant plus vrai que les pilotes de périphériques dont elles sont issues sont en général moins testées que le reste du code du noyau. D'autre part, cette technique revient à introduire des types abstraits opaques, séparant leur représentation de leur interface. Ce concept étant absent de C, de nombreuses erreurs de programmation notamment liées aux manipulations d'entiers peuvent également être vérifiées à l'aide de cette méthode.

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec gravida, diam ullamcorper volutpat varius, est dui blandit nibh, et posuere nisi tortor ut lectus. Aliquam in ipsum at nisl luctus pulvinar. Nulla pulvinar accumsan interdum. Aliquam varius est et ligula commodo varius quis gravida justo. Ut sit amet eros orci, at iaculis metus. Cras placerat condimentum magna ut dictum. Mauris id sollicitudin quam. Ut pretium aliquet lacus, et elementum odio blandit quis.

Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Morbi tincidunt, justo sed venenatis imperdiet, diam dui blandit nisl, et sodales urna mauris sit amet felis. Sed congue sodales sollicitudin. Integer imperdiet tellus et ligula consequat quis commodo metus posuere. Donec id vehicula mi. Nullam id tincidunt orci. Nunc luctus ligula ut urna laoreet consequat. In hac habitasse platea dictumst. Sed imperdiet ultricies diam id cursus. Quisque metus orci, tempor et dapibus eget, ullamcorper in augue. Nullam non urna metus, id accumsan lacus. Duis erat arcu, porttitor non laoreet vitae, ultrices vitae tortor. Pellentesque interdum faucibus tellus eu semper. Nulla ac aliquet mauris.

In in metus purus. Nunc ullamcorper posuere turpis, vel laoreet tortor posuere at. Ut euismod arcu sit amet dui varius vel varius neque pellentesque. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam nec mi interdum ipsum eleifend tincidunt. Duis bibendum sem vitae risus facilisis mattis. Sed varius eleifend nulla vel commodo. Morbi risus metus, egestas non commodo id, ullamcorper a enim. Integer et augue et orci feugiat rhoncus a in augue.

Nulla elementum felis eu eros pretium quis porttitor felis faucibus. Vivamus a vulputate tellus. Suspendisse potenti. Morbi dapibus, quam ac consectetur hendrerit, lorem urna faucibus dui, in vestibulum tellus nisl sit amet ligula. Morbi ultricies venenatis nisi sit amet scelerisque. Nam pharetra bibendum lacinia. Suspendisse potenti. Suspendisse elementum tellus et mi hendrerit sagittis.

Ut dignissim fermentum cursus. Praesent feugiat venenatis rutrum. Cras eget nibh id quam ultrices pellentesque eget non elit. Maecenas justo lorem, fermentum vitae dictum eu, egestas eu urna. In hac habitasse platea dictumst. Cras nec arcu at purus aliquet accumsan. Nam interdum malesuada orci et sagittis