# Les Compilateurs

## *Compilers*

**Paul Feautrier**

*Université de Versailles Saint-Quentin*
*45 Avenue des Etats Unis*
*78035 VERSAILLES CEDEX FRANCE*
*Paul.Feautrier@prism.uvsq.fr*

RÉSUMÉ. *On présente l'état de l'art et l'évolution récentes des techniques de compilation.*

ABSTRACT. *This paper presents the state of the art and recent developments in compilation techniques.*

MOTS-CLÉS : *compilateurs, parallélisation automatique, optimisation.*

KEY WORDS: *compilers, automatic parallelization, optimization.*

**1ère Soumission à TSI, version 1.1 Volume 1 – n° 1/2000, pages 1 à x**

### 1. Introduction

La paresse est la première qualité de l'informaticien. L'informatique ne s'est développée que parce que certains en ont eu assez de faire des multiplications en série ou de trier des fiches en carton. Les ordinateurs ont permis de remplacer ces activités répétitives par l'écriture d'un programme. Mais l'activité de programmation est elle-même en partie répétitive, et le principe du moindre effort veut qu'elle soit automatisée dans la mesure du possible.

L'utilisation d'un compilateur permet de programmer dans un langage de haut niveau, plus concis et moins sujet à erreur que le langage machine. Un programme écrit dans un tel langage doit être traduit pour être exécuté. Depuis la réalisation du premier compilateur par John Backus, la qualité de cette traduction est un souci permanent : il s'agit de ne pas trop perdre de performance par rapport à la programmation en langage machine. Ce sont les méthodes d'optimisation qui permettent de limiter cette perte de performance, et ce sont sur elles que se concentrent la recherche actuelle du domaine. Après une rapide présentation de la

structure d'un compilateur, cet article traitera donc surtout des méthodes d'optimisation, aussi bien pour architectures séquentielles que parallèles.

## 2. **Structure d'un compilateur**

### *2.1 Analyse syntaxique*

Le rôle de l'analyse syntaxique est de vérifier que le programme source est conforme à la grammaire du langage utilisé et de l'enregistrer sous une forme facile à manipuler dite « représentation intermédiaire » (RI). Comme la syntaxe d'un langage est presque toujours définie par une grammaire hors-contexte, on peut pour cette phase utiliser un compilateur de compilateur dont le plus connu est Yacc. Cette phase est aujourd'hui bien maîtrisée. On se reportera au traité [AHO 86].

### *2.2 La représentation intermédiaire*

Le choix de la RI est très important pour la suite des évènements, parce que c'est sur elle que travailleront les phases suivantes du compilateur. Actuellement, on utilise souvent l'arbre d'analyse syntaxique (souvent simplifié) comme représentation intermédiaire. Cette solution a l'avantage de ne perdre aucune des informations présentes dans le programme.

L'arbre d'analyse syntaxique peut être insuffisant pour certaines applications. Il peut arriver par exemple que l'on souhaite manipuler des expressions arithmétiques. On est alors amené à utiliser des représentations empruntées aux outils de calcul formel, voire même à se connecter à un tel outil.

### *2.3 Optimisation*

Cette phase consiste en des transformations de la RI, qui doivent améliorer les performances du programme sous la contrainte que les résultats n'en soient pas modifiés. Certaines optimisations sont indépendantes de la cible : il est (presque) toujours bénéfique de ne pas refaire plusieurs fois le même calcul. D'autres, au contraires, dépendent de l'état courant de l'architecture informatique : ce n'est que depuis que les ordinateurs ont des caches que l'on se préoccupe de la gestion de la localité, et ces recherches deviendront inutiles si l'on trouve d'autres moyens pour compenser la lenteur des accès à la mémoire. Les techniques d'optimisation seront présentées en détail en Section 3.

### *2.4 Génération de code*

Très généralement, il s'agit d'appliquer une suite de transformations à la représentation intermédiaire jusqu'à la rendre équivalente à un programme en langage machine ou en langage d'assemblage, qu'il n'y a plus qu'à imprimer. La phase de transformation doit tenir compte des particularités de la machine cible. On peut en faire une implémentation *ad hoc*, et il faudra la reprendre pour chaque nouveau processeur. On a cherché à construire des générateurs de code universels

paramétrés par une description de la machine cible. Cette description prend en général la forme d'un système de ré-écriture de la représentation intermédiaire, parfois préalablement linéarisée (générateurs de Graham-Glanville).

## 3. Techniques d'optimisation

### *3.1 Analyse sémantique*

Pour respecter la règle d'invariance des résultats, le compilateur doit avoir une compréhension, même sommaire, du fonctionnement du programme, et doit pouvoir juger de l'impact d'une modification. La recherche des informations nécessaires à cette compréhension constitue la phase d'analyse sémantique du programme. La difficulté, c'est qu'une analyse parfaite est impossible. Elle permettrait, en effet, de répondre à certaines questions (comme la terminaison ou l'équivalence) qui sont indécidables dès que le langage permet de simuler une machine de Turing. Il y a deux façons de résoudre ce problème. On peut tout d'abord se contenter d'analyses approximatives. On doit choisir soigneusement le sens de l'approximation : une analyse approximative ne doit jamais conduire à la génération d'un programme erroné. L'autre possibilité est de restreindre la classe des programmes analysés. Si ceux-ci n'ont plus la puissance de la machine de Turing (par exemple, si la terminaison est garantie), rien ne s'oppose à une analyse exacte et complète. On verra un exemple de cette situation en section 4.

Les analyses locales ont pour but de prouver que la valeur d'une variable (ou d'un groupe de variables) a une certaine propriété en un point du programme. Par exemple, si dans la portion de code :

```
      real x[100] ;
      …
S :   x[i]=.. ;
```

on peut prouver la propriété $0 \leq i < 100$ en S, on pourra se dispenser du test de validité de l'indice.

Mais dans de nombreux cas, on a besoin du concept de dépendance. Deux instructions sont en dépendance si on ne peut les intervertir sans courir le risque de changer les résultats du programme. Il y a des dépendances de données (par exemple quand une instruction utilise le résultat de celle qui la précède), et des dépendances de contrôle (on ne peut intervertir le test d'une conditionnelle et ses branches). Les dépendances de données les plus importantes sont les dépendances de flot, qui relient entre elles les créations de nouvelles valeurs (les instructions d'affectation), et les utilisations de ces valeurs. Par exemple, dans le code :

```
      S : x = 0 ;
      …
      T : y = x*x ;
```

on cherchera à établir que la seule source (*reaching definition*) de `x` dans l'instruction *T* est l'instruction *S*. Il faut pour cela montrer que la partie du code non représentée ne contient pas d'affectation à `x` et qu'il est impossible de parvenir en *T*

sans passer par *S*. Si on y parvient, on aura démontré que x = 0 en *T* et que l'on peut remplacer cette instruction par `y = 0;`.

Les méthodes d'analyse des programmes sont maintenant bien comprises et ressortent toutes soit de la méthode de l'interprétation abstraite pour laquelle nous renvoyons à l'article de P. Cousot dans ce numéro de TSI, soit de méthodes géométriques que nous présenterons en Section 4.1

### *3.2 Optimisations classiques*

Leur but est d'éliminer les calculs inutiles. Par exemple, la propagation des constantes consiste à effectuer au moment de la compilation tous les calculs dont les arguments sont connus, de façon à ne pas avoir à les refaire pendant l'exécution.

L'optimisation la plus importante en ce genre est celle qui consister à éliminer les invariants de boucle (*hoisting*). Une expression est invariante dans une boucle si chaque itération calcule la même valeur. Il est avantageux de déplacer l'évaluation de cette valeur unique avant la boucle. On reconnaît une expression invariante à ce que les sources de ses variables sont toutes extérieures au corps de la boucle.

### *3.3 Optimisations liées à l'architecture de la cible.*

Les gains de performance des processeurs modernes sont dus pour partie aux progrès de la technologie, et pour une autre partie aux progrès de l'architecture. De ce coté, on peut identifier deux sources d'accélération : l'exploitation du parallélisme caché dans les programmes séquentiels, et l'exploitation de la localité dans les accès à la mémoire. Dans les deux cas, les architectes ont prévu de cacher ces mécanismes, et l'utilisateur peut programmer comme s'il disposait d'un ordinateur séquentiel muni d'une mémoire à temps d'accès uniforme.

Cependant, on s'est aperçu depuis longtemps que ces dispositifs fonctionnent mieux si on les prend en compte au moment de la programmation. Par exemple, le détecteur de parallélisme caché fonctionne mieux si on regroupe ensemble des opérations indépendantes. Ce type d'optimisation est très complexe et non portable. Il est donc en général caché dans le compilateur.

Comme il n'est pas possible de présenter ici toutes les optimisations liées à l'architecture, nous nous bornerons à présenter une méthode, le pipeline logiciel. Le problème de l'amélioration de la localité sera traité en section 4.

Les processeurs modernes disposent d'unités de calcul indépendantes, souvent spécialisé. Soit un processeur équipé d'un additionneur et d'un multiplieur. Pour simplifier on suppose que chaque opération prend un cycle, et on ignore les accès à la mémoire. Soit l'instruction :

```
y = a*x + b ;
```

Il n'y a pas de parallélisme, car l'additionneur doit attendre la fin de la multiplication pour opérer. Soit maintenant la boucle :

```
for(i=0 ; i<n ; i++)
```

```
    y[i] = a*x[i]+b ;
```

On peut effectuer en parallèle, à l'itération *i*, le produit `a*x(i)` et l'addition de l'itération *i-1*. On peut représenter cette optimisation comme une transformation de la boucle, qui devient :

```
r[0] = a*x[0] ;
for(i=1 ; i<n-1 ; i++){
    r[i%2] = a*x[i] ;
    y[i-1] = r[(i-1)%2] + b ;
    }
y[n-1]=r[n-1]+b ;
```

Naturellement, `r` n'est pas un tableau, mais une paire de registres utilisés de façon cyclique. La boucle ainsi transformée devient un pipeline logiciel. Les deux instructions qui précèdent et suivent la nouvelle boucle en sont le prélude et le postlude. Si $n$ est grand, la performance de l'ordinateur est doublée par rapport à une compilation naïve. Existe-t-il des méthodes générales pour la construction de tels pipelines logiciels ?

Il est assez facile d'imaginer des méthodes d'ordonnancement pour du code sans boucle. On peut par exemple utiliser une table de réservation : une matrice dont les lignes représentent les unités de calcul et dont les colonnes représentent le temps. On marque une case si l'unité correspondant à la ligne est occupée pendant le cycle correspondant à la colonne. On cherche à placer chaque opération le plus tôt possible (le plus à gauche) en respectant le flot des données et les réservations déjà faites. Cette méthode ne donne pas nécessairement le code optimal, mais elle est simple et on peut borner la perte de performance qu'elle entraîne.

La méthode de l'ordonnancement modulo (*modulo scheduling*) fonctionne de façon analogue, a ceci près que l'on commence par choisir la période *T* du pipeline logiciel. Si l'on utilise une ressource à l'instant *t*, elle sera également utilisée aux instants *t+T*, *t+2T*, etc. et on marque en conséquence la table de réservation. Il se peut alors qu'il soit impossible de trouver un ordonnancement valide. Dans ce cas, on recommence pour une valeur supérieure de *T*, et ceci jusqu'à succès.

Il est intéressant de voir apparaître sur cet exemple un cas d'interférence entre optimisations. Les processeurs modernes ont tous des registres, qu'il est important de bien utiliser pour éviter des accès coûteux à la mémoire. La boucle originale utilise un registre pour stocker le produit `a*x[i]`, alors que le pipeline logiciel en utilise deux. Il est fréquent qu'une version optimisée d'une boucle ait besoin de plus de registres que la version initiale, ce qui peut imposer des écritures de résultats intermédiaires en mémoire, lesquelles peuvent faire perdre les gains de performance dus au pipeline logiciel. Il s'agit là d'une situation très courante en optimisation, et ce genre d'interaction est impossible à prendre en compte pour le matériel, et très difficile à traiter pour le programmeur. Des études récentes cherchent à coupler le pipeline logiciel et l'allocation des registres à l'aide de la programmation linéaire en nombres entiers.

## 4. La parallélisation automatique et le modèle polyédrique

La parallélisation automatique est un cas particulier de d'optimisation où la cible est une machine parallèle ou vectorielle. Dans le cas des programmes de calcul numérique intensif, qui sont basés sur des algorithmes d'algèbre linéaire hautement répétitifs (boucles `DO`), agissant sur des tableaux et utilisant des mécanismes d'indexation assez simples (on parle de programmes réguliers), il a été possible de développer une théorie complète : le modèle polyédrique.

### *4.1 L'approche géométrique.*

Dans l'espace à $n$ dimensions, un polyèdre est un ensemble de points satisfaisant à un système d'inégalités linéaires. Par exemple, dans l'espace à deux dimensions, l'ensemble $D=\{x,y|0\leq x<n,\ 0\leq y<x\}$ est un polyèdre (en fait, un triangle rectangle isocèle). Soit maintenant la boucle :

```
      for(i=0 ; i<n ; i++)
            for(j=0 ; j<i ; j++)
      S :         y[i]=y[i]+a[i][j]*x[j] ;
```

L'instruction $S$ va être exécutée $i$ fois pour chaque valeur de $i$, soit $n(n+1)/2$ fois au total. Il faut nommer chacune des exécutions ou opérations : il suffit pour cela d'associer à chacune d'elles les valeurs courantes de $i$ et de $j$. Nous parlerons de l'opération $<i,j>$. De l'analyse des bornes des boucles on déduit $0\leq i<n$ et $0\leq j<i$, ce qui signifie que le vecteur $<i,j>$ appartient au polyèdre $D$ ci-dessus. Mais les compte-tours d'une boucle sont des entiers. Le domaine d'itération de $S$ est donc l'ensemble des points entiers contenus dans $D$, objet mathématique que l'on appelle un Z-polyèdre, et dont la dimension n'est autre que le nombre de boucles entourant $S$. Cette façon de représenter un programme a de multiples avantages. La théorie des polyèdres et des Z-polyèdres a été bien développée pour les besoins de la Recherche Opérationnelle. De nombreuses opérations (par exemple l'intersection de deux polyèdres) sont triviales, et deux algorithmes bien connus, le Simplex et l'algorithme de Fourier, permettent de décider si un polyèdre est vide ou non. La succession temporelle des opérations est donnée par l'ordre lexicographique des vecteurs d'itération. Vu les contraintes que nous imposons aux fonctions d'indice, les opérations d'indexation conduisent à calculer des images de polyèdres par des fonctions affines, et les dépendances peuvent être décrites exactement par des Z-polyèdres, c'est-à-dire comme des systèmes de contraintes linéaires en nombres entiers.

Il y a ensuite plusieurs façons d'exhiber le parallélisme caché du programme. La plus parlante consiste à dater chaque opération : deux opérations exécutées à la même date sont exécutées en parallèle. L'ordonnancement doit satisfaire une contrainte de causalité : une opération ne peut commencer que si tous ses opérandes sont disponibles, ou encore si toutes ses sources sont terminées. On obtient de cette façon un système de contraintes qui lui aussi se résout par la programmation linéaire.

Pour paralléliser l'exemple ci-dessus, il faut d'abord rechercher les sources de l'instruction *S*. Les tableaux `A` et `x` ne sont pas modifiés ; on peut supposer qu'ils existent au moment où la boucle commence : ils n'imposent pas de contrainte à l'ordonnancement. Par contre, la source de `y[i]` à l'itération $<i,j>$ est clairement l'itération $<i,j-1>$ . L'ordonnancement $t$ doit donc vérifier :

$t(i,j) > t(i,j-1)$

et il est facile de voir que $t(i,j) = j$ est une solution.

La génération d'un programme parallèle à partir d'un ordonnancement est un problème d'énumération des opérations dans l'ordre des temps croissants. Il a reçu beaucoup d'attention et de nombreuses solutions ont été proposées, mais son étude est assez technique et nous entraînerai trop loin ici. Dans le cas de l'exemple, il faut trouver quelles sont les opérations à effectuer à l'instant $t$. Vu l'ordonnancement choisi, elles satisfont $0 \leq i < n$, $j=t$, $0 \leq j < i$, ce qui se simplifie en $t<i<n$. Ceci conduit au programme :

```
for(t=0 ; t<n ; t++)
      forall(i=t+1 ; t<n ; t++)
            y[i]=y[i]+a[i][t]*x[t] ;
```

où nous avons utilisé le mot clef `forall` pour signaler une boucle parallèle.

Ces quelques indications constituent les linéaments d'une méthode de compilation pour ordinateurs parallèles. Pour en faire une méthode pratique, d'autres problèmes plus techniques doivent être résolus. Par exemple, la méthode dégage en général trop de parallélisme pour les architectures courantes. Le problème de l'ordonnancement sous contrainte de ressources (aussi bien le nombre de processeurs que la taille de la mémoire) n'a pas pour le moment de solution satisfaisante, et on doit se contenter de méthodes de *pavage* plus ou moins *ad hoc*.

### *4.2 L'amélioration de la localité*

On dit qu'un programme a de la localité lorsque, en un point donné de son déroulement, il a tendance à n'utiliser qu'une faible fraction de son espace mémoire. Cette fraction constitue l'espace de travail (working set) du programme ; naturellement, elle évolue au fur et à mesure de l'avancement des traitements. Un programme a d'autant plus de localité que ses espaces de travail sont plus petits.

Il est important qu'un programme ait beaucoup de localité, qu'il soit exécuté sur une architecture parallèle ou séquentielle. Par exemple, lorsque l'on a construit un code parallèle, et donc alloué un ensemble d'opérations à chaque processeur, il faut rapprocher les données du processeur qui va les utiliser. Dans le modèle polyédrique, on y parvient de la façon suivante.

Il s'agit d'attribuer un numéro de processeur à chaque cellule de tableau et à chaque opération. Pour cela, on postule que ce numéro est une fonction affine des indices du tableau (ou du vecteur d'itération de l'opération). Pour qu'il n'y ait pas de communications, il faut que chaque opération soit exécutée par le processeur dont la mémoire héberge ses opérandes et son résultat. En écrivant cette *condition de placement* pour chaque instruction, on obtient un système d'équations linéaires et

homogènes dont les inconnues sont les coefficients des fonctions de placement. Ce système a toujours la solution triviale, qui correspond au *collapsus* du calcul sur le processeur 0. Il n'en a en général pas d'autres. Pour trouver un placement intéressant, il faut ignorer certaines conditions de placement, ce qu'il faut compenser en prévoyant des communications résiduelles. De nombreuses heuristiques ont été développées pour bien choisir ces communications : par exemples celles de plus petit volume ou celles que l'architecture cible réalise efficacement.

La localité est tout aussi importante pour les machines séquentielles à cause de l'utilisation des caches pour mieux tolérer la latence de la mémoire. Mais la complexité du fonctionnement des caches ainsi que la non linéarité du problème (calcul du « volume d'un polyèdre »), font que seules des solutions partielles sont actuellement connues.

## 5. Conclusion : nouvelles machines, nouveaux langages

En résumé, le rôle d'un compilateur est d'assurer le passage des langages de haut niveau vers l'architecture des processeurs modernes. Du coté des langages, la tendance est à la concision maximale, le compilateur étant chargé de fournir les détails manquants par un processus qui se rapproche de plus en plus de la programmation automatique. De l'autre coté, les architectures deviennent de plus en plus complexes, et des performances satisfaisantes ne peuvent être obtenues que par un processus d'optimisation presque hors de porté du programmeur, tant les facteurs à prendre en compte sont nombreux. La compilation devient une activité si complexe qu'il est difficile d'en avoir la maîtrise complète. On se rabat sur des enchaînements de compilateurs ou de phases de compilateur, qui sont de plus en plus difficiles à agencer, et qui ne fournissent qu'une solution sous-optimale au problème général de l'écriture d'un programme efficace. Il manque ici une théorie de « grande unification », qui devrait prendre en compte à la fois l'efficacité du code et les contraintes de ressources (nombre de registres, d'unités fonctionnelles, taille des caches et de la mémoire).

Un autre problème, qui a été beaucoup discuté mais qui n'est pas entièrement résolu est celui du compilateur paramétrable, capable de s'adapter le plus vite possible à l'évolution des processeurs. On voit qu'il y a encore beaucoup à faire en compilation !

## 6. Références.

Plutôt que de donner des références sous la forme usuelle, ce qui aurait pris trop de place, j'ai préféré indiquer quelques points d'entrée qui permettront au lecteur de mener sa propre recherche bibliographique.

Le traité classique est :

[AHO 86] A.J. Aho, R. Sethi, J.D. Ullman :*Compilers, Principle, Techniques and Tools,* Addison-Wesley, 1986.

On peut également signaler :

[MUCH 97] S. S. Muchnick : *Advanced Compiler Design and Implementation*, Morgan Kaufmann, 1997.

[DARTE 96] G-R. Perrin et A. Darte (Eds.), *The Data Parallel Programming Model*, Springer, LNCS 1132, 1996.

Pour tout ce qui concerne les polyèdres et la programmation linéaire, on consultera

[SCHR 86] A. Schrijver, *Theory of Linear and Integer Programming*, Wiley, 1986.

On trouvera de nombreux articles sur la compilation dans la revue « Transaction on Programming Languages and Systems » de l'ACM. Les articles relatifs à la parallélisation automatiques se trouvent dans des revues plus spécialisées, comme « Int. J. of Parallel Programming » (Plenum), « J. of Parallel and Distributed Computing » (Academic Press) et « Parallel Computing » (Elsevier).

Les principales conférences du domaine sont « Principles of Programming Languages » (ACM), « Programming Languages Design and Implementation » (ACM) et, pour la partie parallèle, « Principles and Practice of Parallel Programming » (ACM) et « Parallel Architectures and Compiler Techniques » (IEEE).