Master Informatique Fondamentale - M1 Compilation

Organisation du programme objet

Paul Feautrier

ENS de Lyon
Paul.Feautrier@ens-lyon.fr
perso.ens-lyon.fr/paul.feautrier

1er mars 2007

Introduction

Avant de commencer à générer le programme objet, il faut définir un modèle d'exécution, qui doit permettre l'implémentation de tous les aspects du langage source :

- Exemple : si le langage permet des tableaux extensibles (Matlab) ou des tables hashcodée (Ocaml, Tk/Tcl, Maple), le modèle doit comporter les outils nécessaires (gestion dynamique de la mémoire).
- Le modèle doit permettre une implémentation raisonablement efficace.
- Exemples de contraintes :
  - Récursivité
  - Gestion dynamique de la mémoire
  - Compilation séparée, bibliothèques, relations avec le système d'exploitation, entrée / sorties
  - Tableaux

Contraintes, I : questions de temps

On doit distinguer le temps de la compilation et le temps de l'exécution.

- Le compilateur n'a pas accès directement à ce qui se passe à l'exécution.
- Inversement, les objets du compilateur ont disparu (en général) au moment de l'exécution.

  Par exemple, problème des tables de symboles au moment du deboguage, option -g du compilateur C.

- Pour agir à l'exécution, le compilateur doit insérer du code dans le programme objet.

Récursivité, règles de visibilité

- Dans presque tous les langages les programmes peuvent être structurés en procédures (fonctions, sous-programmes) qui peuvent être récursive.
- Il s'en suit que la gestion de la mémoire ne peut être statique – le compilateur ne peut pas calculer toutes les adresses.
- Tous les langages ont des règles de visibilité qui définissent des espaces de nommage indépendants (exceptions : langages interprétifs).
- Les règles de visibilité diffèrent d'un langage à l'autre (exemples : C et Pascal).

Gestion dynamique de la mémoire

Certains langages permettent la création dynamique de structures de données.

- Le programme objet doit dialoguer avec le système d'exploitation
- Si l'on veut une gestion automatique des données, le compilateur doit instrumenter les structures :
  - prévoir un compteur de références
  - marquer les pointeurs ou les identifier d'une autre façon

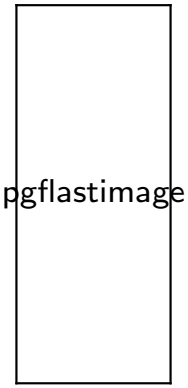
Compilation séparée

- Possibilité de découper un programme en modules pouvant être compilés indépendamment.
- En particulier, utilisation d'une bibliothèque
- Deux approches :
  - à la Fortran (exemple C) : il n'y a aucune communication entre modules. Pour maintenir un minimum de cohérence, on demande au programmeur de *recopier* certaines informations (les déclarations).
  - l'ordre des compilations est arbitraire
  - à la Ada (exemple Ocaml) : le compilateur enregistre une partie de la table des symboles après chaque compilation.
  - l'ordre des compilations est contraint

Entrées /sorties

- Certains langages n'ont pas d'instructions d'entrée sortie (C, Ocaml) ; elles sont déléguées à une bibliothèque.
- Inconvénient : les fonctions ne peuvent pas être génériques. Exemple du `close_out` en Ocaml.
- D'autres langages ont des instructions d'entrée sortie, qui doivent être compilées.

Organisation de la mémoire, I



- Ne pas confondre espace d'adressage et espace mémoire.
- Le compilateur suppose toujours que le programe qu'il construit est seul en mémoire.
- En général, c'est le dispositif de mémoire virtuelle qui permet cette hypothèse.
- La disposition ci-dessus est la plus courante, mais les conventions peuvent changer suivant le processeur (certains processeurs réservent la moitié de l'espace d'adressage pour le système)

Organisation de la mémoire, II

Si le processeur n'a pas de mémoire virtuelle (systèmes embarqués, microcontroleur).

- **Code relatif** : on spécialise un registre pour contenir l'adresse de base du code, et ce registre est additionné à toutes les adresses. Même méthode pour les données.
- Pour déplacer le code, il suffit de changer la valeur du registre de base.
- **Code translatable** : On écrit le code comme si l'adresse de base était 0.
- Au chargement, on additionne l'adresse de base à toutes les adresses.
- Il faut prévoir une zone hors texte pour savoir si une instruction contient une adresse ou non.

Pile d'exécution, principe

- Chaque fois qu'une procédure est activée, elle a besoin de place en mémoire pour ses variables locales et ses arguments.
- Si une procédure est récursive, il peut y avoir plusieurs activations en cours à un instant donné.
- L'espace nécessaire doit donc être alloué dynamiquement au moment de l'appel, et libéré au moment du retour.
- Mais on peut faire mieux que d'utiliser un outil comme `malloc` parce que les entrées et sorties de procédures se font dans l'ordre Last In First Out.
- On peut donc utiliser une *pile*.

Pile d'exécution, structure

pgflastimage

- La pile est divisée en *cadres* ou enregistrements d'activation.
- Chaque cadre comporte :
  - Les arguments de la procédure
  - L'adresse de retour
  - Un lien vers le cadre de la procédure appelante
  - Les variables locales, et en particulier un emplacement pour le résultat de la procédure.
- Un register spécialisé, la base, pointe vers le lien descendant

Comment ça marche

- Adressage des variables locales : `*(base - 4)`
- Adressage des arguments : `*(base + 8)`
- Séquence d'appel et de retour :

```
  tos += sizeof(vars);  //tos :: Top of Stack
  *(++tos) = arg1;
  *(++tos) = arg2;
..
  call();
  tos -= sizeof(vars) + sizeof(args);
```

- Le compilateur déduit des déclarations la taille des variables et des arguments ; il calcule les déplacements par accumulation et les enregistre dans la table des symboles

Résultat de la fonction

- Si le processeur le permet, le résultat est retourné dans un registre.
- Si le résultat est trop gros, il peut être retourne sur la pile.
- Cas extrème : le résultat est rangé sur le tas, et un pointeur est retourné dans un registre.

Appel par valeur, appel par référence

- En C tous les paramètres sont transmis par copie : la procédure appelée dispose d'une copie du paramètre qui fonctionne comme une variable locale. Les modifications ne sont pas répercutées.
- Dans certains langages (Pascal, C++), un paramètre peut être transmis par réfarence. Equivalent à la transmission d'un pointeur en C.
- Mais c'est le compilateur qui doit déréférencer le pointeur, alors qu'en C c'est à la charge du programmeur.

Gestion des registres

La procédure appelante ne peut pas savoir ce que la procédure appelée va faire des registres du processeurs. Deux solutions :

- La procédure appelante sauvegarde les registres qu'elle juge utile avant l'appel
- La procédure appelée sauvegarde les registres qu'elle va utiliser avant de commencer
- la sauvegarde se fait sur la pile d'exécution
- certains processeurs ont des instructions de sauvegarde / restauration rapide des registres
- La décision dépend de plusieurs facteur : politique d'allocation de registres, nombre de registres, etc.

Compilation séparée

Pour construire de gros programmes, il est indispensable de pouvoir les découper en *modules* compilables séparément.

- Un module est à la fois une unité de compilation (le plus souvent, une procédure) et une unité de stockage (un fichier).
- Quand il traite un module, le compilateur ne possède pas toutes les informations nécessaires
  - au contrôle des types
  - au calcul des adresses

Contrôle des types à la C

En C, le compilateur demande que les types des objets éloignés soient dupliqués dans le module utilisateur.

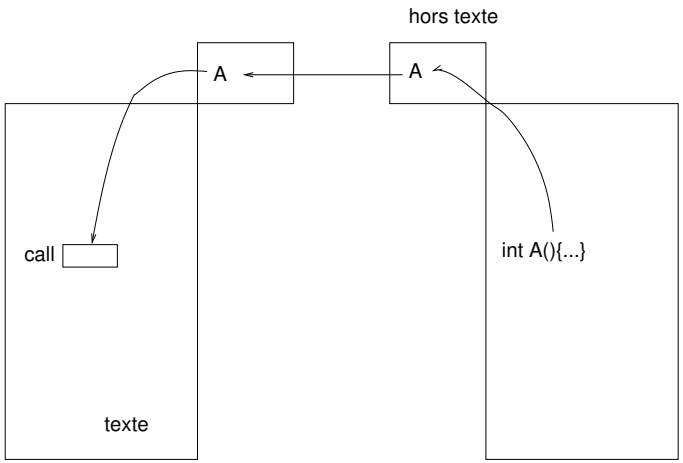
```
 float power(float x, int n){
    corps de la fonction
}
------------------------------
 float power(float, int);

 main(int argc, char *argv[]){
   ...
   u = power(v, k);
   ...
}
```

- La correction de la compilation repose sur la bonne foi du programmeur
- Si les types sont omis, le compilateur choisit des types par défaut, parfois surprenants
- Le préprocesseur permet d'importer sans erreur des fichiers de déclarations.

Contrôle de type à la Ada

- Le langage permet de savoir (et même de décider) quels sont les objets visibles à l'extérieur du module.
- Le compilateur crée un fichier contenant la partie de la table des symboles relative aux objets exportés.
- Cette table sert à la compilation des autres modules.
- Un module A dépend d'un module B si A utilise un objet de B.
- Le graphe de dépendance doit être acyclique (pas de récursivité entre modules).
- La compilation doit se faire selon un tri topologique du GD

Edition de liens



- Le compilateur ne peut pas terminer la compilation séparée : le programme comporte des trous.
- Le fichier *objet* comporte des trous référencés dans un hors-texte.
- C'est l'éditeur de lien qui rapproche les hors-texte et remplit les trous.
- Le format du fichier objet est normalisé.
- Bien souvent, on laisse au compilateur C ou à l'assembleur le soin de générer l'objet.

Accessibilité

Suivant les langages, les règles d'accès aux objets étrangers varient.

- En C et Java, les procédures ne s'emboitent pas, et les objets sont soit locaux (accés relatif à l'enregistrement d'activtion courant), soit globaux (adressage absolu). Voir cependant les extensions sauvages du compilateur gcc.
- En Pascal, ADA, les procedures s'emboitent, et il est possible d'accéder aux objets des procédures englobantes, si toute fois ils ne sont pas masqués par un objet local.
- En particulier, il n'y a qu'un seul enregistrement accessible par procédure récursive, le plus récent.
- On peut accéder aux objets non locaux en déroulant la chaîne des pointeurs d'activation, mais c'est inefficace.

Méthode du *display*

- Chaque procédure a une profondeur lexicale (sa profondeur dans l'arbre de syntaxe abstraite.
- A chaque instant, il n'y a qu'une seul procédure visible à chaque profondeur lexicale.
- On peut donc préparer un tableau, le *display*, qui contient pour chaque profondeur lexicale un pointeur vers l'enregistrement d'activation de la procédure visible.
- Ce tableau doit être entretenu à chaque entrée dans une procédure et à chaque retour s'il y a eu possibilité de récursivité.

Relations avec le système d'exploitation

Si le matériel le permet, elles sont réduites au minimum – en fait, camouflées dans des fonctions de bibliothèque.

- Extension de la pile : transparente, sur détection d'une violation de la protection mémoire.
- Appels au système (en particulier entrées/sorties, terminaison et gestion du tas), implémentées sous forme d'une bibliothèque
- Lancement d'un programme : le système fait un saut vers une adresse fixe, où l'éditeur de lien a placé un programme qui simule un appel de la fonction `main` (en C).

Gestion dynamique de la mémoire

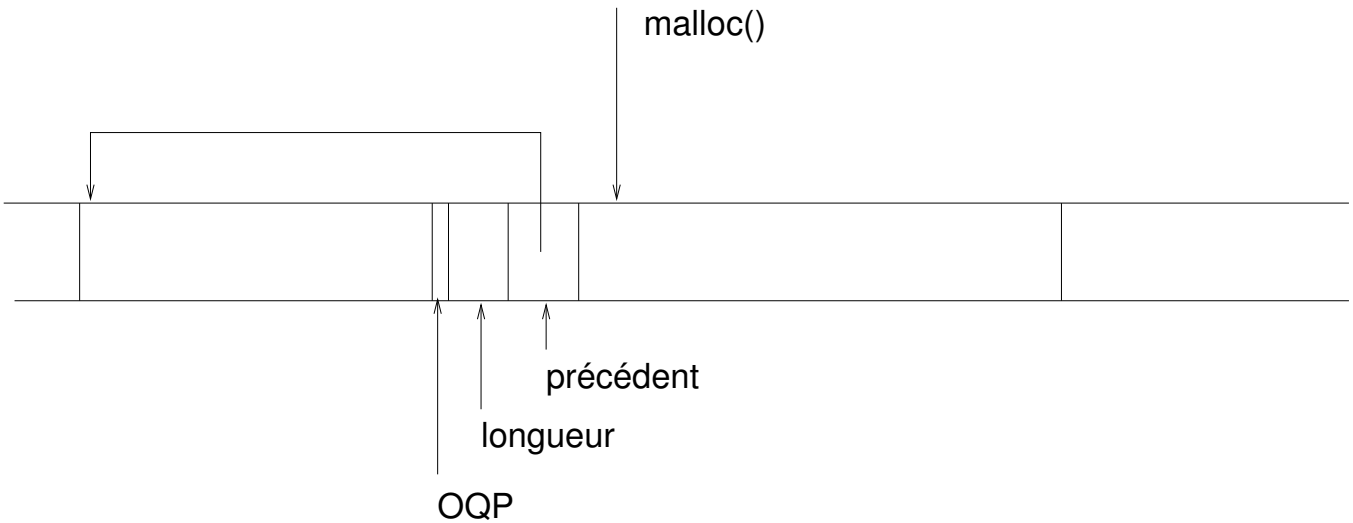
Indispensable dans le langages modernes. Deux classes de méthodes :

- La mémoire est gérée par le programmeur (C, C++).
- Avantage : pas de contrainte sur les structures de données et la manipulation des pointeurs.
- Inconvénient : c'est difficile à écrire, et il est presque impossible de ne pas faire d'erreurs (*fuites de mémoire*).
- La mémoire est gérée automatiquement par le système d'exécution (Java, Lisp, Ocaml)
- Avantage : facile et sûr
- Inconvenient : contraintes sur l'utilisation des pointeurs
- Inconvénient : le système de gestion peut dégrader les performances

`malloc`, `free`, principes

- Le tas est divisé en blocs qui peuvent être libres ou occupés. Chaque bloc est précédé d'un mot contenant sa longueur.
- Les blocs libres sont chaînés entre eux.
- Pour allouer un bloc, on parcourt la liste libre jusqu'à trouver un bloc de taille suffisante, on alloue et on reconnecte la liste, si besoin en récupérant la place inutilisée. S'il n'a pas de bloc suffisant, on demande au système d'exploitation d'agrandir le tas.

free



- Pour libérer, on reconnecte le bloc à la liste libre. Pour éviter la fragmentation, on tente de fusionner le bloc libéré avec ses voisins. Il faut pour cela que la liste libre soit triée, et aussi un pointeur arrière.
- Il s'ensuit que les blocs doivent avoir une taille minimale, de l'ordre de 16 octets.

Méthode des compteurs de références

Chaque bloc contient un champ qui compte les pointeurs qui le référencent.

- Si le compteur passe à 0, on peut libérer le bloc.
- Inconvénient : il faut instrumenter chaque manipulation de pointeur, et l'arithmétique sur pointeur devient impossible.
- La méthode ne marche pas pour les structures circulaires.

Ramasse-miette, I

Trois conditions :

- On doit pouvoir retrouver tous les locs alloués (par exemple parce qu'ils sont chaînés)
- Dans un bloc, on doit pouvoir repérer les pointeurs (par exemple, parce qu'ils ont été regroupés en tête)
- Dans la pile, on doit pouvoir retrouver les pointeurs (par exemple, parce qu'ils ont été regroupés en tête de l'enregistrement d'activation).
- Il faut que le langage interdise le calcul d'adresse et que le compilateur coopère.

Ramasse-miette, II

Algorithme :

- Marquer tous les blocs "inaccessibles" et "non visité"
- Visiter récursivement à partir des pointeurs de la pile d'activation et marquer "accessible"
- Récupérer tous les blocs qui sont restés inaccessibles. En général, on en profite pour compacter l'espace libre.

Inconvénient : le programme reste bloqué pendant l'exécution du ramasse-miette (problème en temps réél).

Gestion des tableaux, I

La gestion des tableaux dépend fortement du langage.

- **Tableaux de taille fixe** (Fortran, C ANSI)
  - Le tableau est alloué soit dans l'espace des données statiques, soit dans l'enregistrement d'activation.
  - Le compilateur connaît son adresse ou son déplacement par rapport à la base.
- **Tableaux de taille variable** (Pascal, addition récente à gcc)
  - Concerne seulement les variables locales
  - La taille du tableau doit pouvoir être calculée à l'entrée de la procédure.
  - Le compilateur doit générer du code pour calculer la taille du tableau et son déplacement dans l'enregistrement d'activation

Gestion des tableaux, II

Certains compilateurs associent à chaque tableau un *descripteur* de taille fixe (dépendant du nombre de dimension).

- Le descripteur peut être alloué sur la pile
- Le tableau proprement dit est alloué sur le tas
- Le descripteur contient un pointeur sur le tableau et la liste des dimensions.
- **Avantages** Le descripteur regroupe tout ce qu'il faut savoir sur le tableau. Il est facile de programmer des opérations sur tableaux (langages à la Fortran 95). La transmission en paramêtre est plus simple.
- **Inconvenient** Multiplication des accès mémoire.

Gestion des tableaux, III

Dans tous les cas, le compilateur doit engendrer le code de calcul d'adresse :

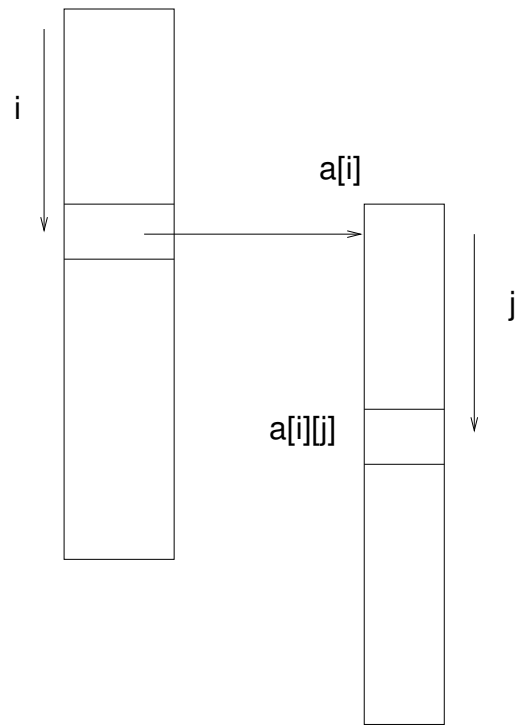
```
float a[n1][n2][n3];
```

```
a[i][j][k]
```

$$\&\texttt{a[i][j][k]} = \&a + (n_1 i + j)n_2 + k$$

De plus, dans certains langages (Pascal), le compilateur doit engendrer un code de vérification ($0 \leq i < n_1$, etc.)

Gestion des tableaux, IV



- En Java, il n'y a que des tableaux à une dimension.
- Un tableau à plusieurs dimensions contient des pointeurs sur des tableaux de dimension inférieure.
- Tous les tableaux sont alloués sur le tas

Gestion des tableaux, V

Le calcul d'adresse se fait par indexations successives :

```
float a[n1][n2][n3];
```

```
a[i][j][k]
```

$$\begin{aligned} \&\texttt{a[i]} &= \&a + i, \\ \&\texttt{a[i][j]} &= \texttt{a[i]} + j, \\ \&\texttt{a[i][j][k]} &= \texttt{a[i][j]} + k \end{aligned}$$

Le compilateur doit également tester les indices.
Inconvénient lourdeur : multiples accés mémoire

Chaînes de caractères

Dans certains langages (C) les chaînes de caractères sont des tableaux ; dans d'autres (Ocaml, Pascal), elles constituent un type de base.

- Particularité : taille variable.
- Solution : surdimensionnement.
- Problème : la représentation de la longueur effective.
- Pascal : le premier caractère code la longueur, longueur limitée à 255 caractères.
- C : marqueur de fin.

En général, le compilateur délègue les opérations sur chaînes de caractères à une bibliothèque.

Représentation des booléens

- La convention usuelle est 0 = faux, 1 = vrai.
- Un booléen peut être logé dans un octet.
- Certains langages traitent les *vecteurs de bits* (C, les langages de description du matériel).
- Tous les processeurs ont des opérations de manipulation des bits et vecteurs de bits.
- On peut compiler les expressions booléennes comme les expressions arithmétiques (voir plus lion).
- Mais très souvent les booléens ne sont que des intermédiaires qui ne sont pas affectés à une variable.

Comparaisons

- Les processeurs possèdent en général un registre de *flags* qui enregistre des informations sur le résultat de la dernière opération arithmétique.
- On peut ensuite exécuter un branchement conditionné par ces flags et en déduire un booléen.

Tests, I

En présence d'un test portant sur une expression booléenne complexe :

- Evaluer les comparaisons
- Calculer l'expression booléenne
- Tester le booléen résultant

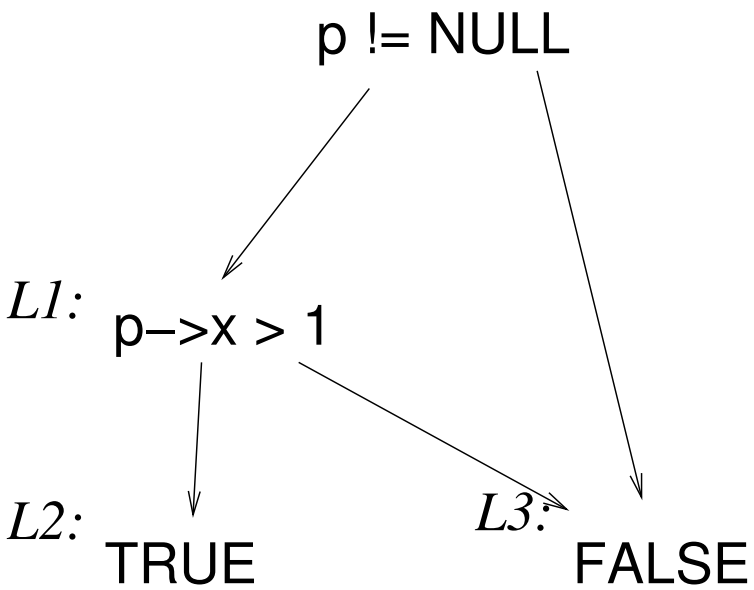
Tests, II

Autre méthode :

- On construit un graphe de décision à partir de l'expression booléenne
- On code le graphe de décision sous forme d'une cascade de branchements conditionnels
- La qualité du résultat dépend fortement de l'odre du codage.
- La méthode implémente naturellement *l'évaluation paresseuse*

Exemple

```
if(p != NULL && p->x > 1) S;
```




```
    if(p == NULL) goto L3;
    if(p->x <= 1) goto L3;
L2 : S;
L3 :;
```

Construction d'un graphe de décision (OBDD)

Principe :

- Sommets : les variables booléennes, **true** et **false**.
- Deux arcs par sommet, vrai et faux.
- Construction récursive à l'aide de la formule de Shannon :

$$f(x, y, \ldots) = x.f(1, y, \ldots) \vee \overline{x}.f(0, y, \ldots)$$

- Identifier les sou-graphes isomorphes.
- Supprimer les sommets dont les deux sous-arbres sont isomorphes.

Switch, case, etc

Deux façons d'implémenter :

- Comme une cascade de tests
- A l'aide d'une table de branchement

Principe de choix :

- La cascade de tests est plus économe en mémoire s'il y a peu de cas (par exemple, 5 caractères dans un alphabet de 256 lettres).
- La table de branchement est plus rapide s'il y a beaucoup de cas.

Boucles

- Boucles `while` : coder comme une instruction conditionnelle.
- Boucle `DO` (Fortran) ou `for` (Pascal) : deux solutions :
  - Convertir en boucle `while`
  - Optimiser la gestion du compte-tour : placement en registre, décomptage
- Boucle `for` (C) : décider si c'est une boucle `DO` ou une boucle `while`. Nécessite une analyse sémantique, voir plus loin.

Compilation des langages à objets

Le problème essentiel : l'héritage = inclusion de types.

- Exemple : la classe `Cat` hérite de `Mammal` hérite de `Animal`.

```
Animal creature;
Cat minou = new Cat(...);
//coercion du particulier au général: OK
creature = minou;
Cat angora;
//erreur: du général au particulier sans coercion
angora = creature;
//OK mais vérification à l'exécution
angora = (Cat) creature;
```

- La coercion du général au particulier doit être signalée, et la JVM la vérifie à l'exécution. L'objet `minou` doit donc comporter un type manifeste (pointeur vers la classe).

Sélection de méthodes

Distinguer soigneusement entre surcharge et redéfinition.

- Surcharge : deux méthodes ayant le même nom mais qui diffèrent par le nombre ou le type de leurs arguments
- Redéfinition : deux méthodes ayant même nom et des arguments de même type, mais attachées à des classes différentes.
- La sélection des méthodes surchargées se fait à la compilation.
- La sélection des méthodes redéfinies se fait à l'exécution : chaque classe a une table de méthodes.

Techniques d'expansion

Le problème : le compilateur doit en permanence ajouter du code au programme source. Exemple : le code de calcul de l'adresse dans un tableau.

- Ce code peut être du texte (exemple : de l'assembleur) ou des structures de données de la RI.
- Il est dangeureux de "souder" ce code dans le programme du compilateur.
- Il vaut mieux fabriquer une base de *modèles* (par exemple stockés dans un fichier) que le compilateur personalise au fur et à mesure des besoins.

Un exemple ancien

Le stockage des instructions assembleur dans le compilateur `pcc` (Portable C Compiler) :

- Chaque instruction est une chaîne de caractères utilisable par `printf` :

```
#define ADD 1
char *template[MAXCODE];

template[ADD] = "    add R%d, R%d, R%d\n";

fprintf(assemblyFile, template[ADD], rx, ry, rz);
```

- On peut perfectionner ce schéma à volonté.