Axe 2 : programmation en assembleur

1. Programmation

1. Programmation

Langages bas niveau

Un langage de programmation bas niveau est un langage qui dépend fortement de la structure matérielle de la machine.

Cette caractéristique offre des avantages et des inconvénients divers.

Elle permet, entre autres,

- d'avoir un bon contrôle des **ressources matérielles** ;
- d'avoir un contrôle très fin sur la **mémoire** ;
- souvent, de produire du code dont l'exécution est très **rapide**.

En revanche, elle ne permet pas

- d'utiliser des techniques de programmation abstraites ;
- de programmer rapidement et facilement.

Langage machine

Un langage machine est un langage compris directement par le processeur en vue d'une exécution.

C'est un langage binaire : ses seules lettres sont les bits 0 et 1.

Chaque modèle de processeur possède son propre langage machine. Étant donnés deux modèles de processeurs P_1 et P_2, on dit que P_1 est compatible avec P_2 si toute instruction formulée pour P_2 peut être comprise et exécutée par P_1.

Dans la plupart des langages machine, une instruction commence par un opcode, une suite de bits qui porte la nature de l'instruction. Celui-ci est suivi des suites de bits codant les opérandes de l'instruction.

– Exemple –

La suite

```
01101010  00010101
```

est une instruction dont le opcode est 01101010 et l'opérande est 00010101. Elle ordonne de placer la valeur $(21)_{\text{dix}}$ en tête de la pile.

Langages d'assemblage

Un langage d'assemblage (ou assembleur) est un langage qui se trouve à mi-distance entre le programmeur et la machine en terme de facilité d'accès.

En effet, d'un côté, la machine peut convertir presque immédiatement un programme en assembleur vers du langage machine. De l'autre, l'assembleur est un langage assez facile à manipuler pour un humain.

Les opcodes sont codés via des mnémoniques, mots-clés bien plus manipulables pour le programmeur que les suites binaires associées.

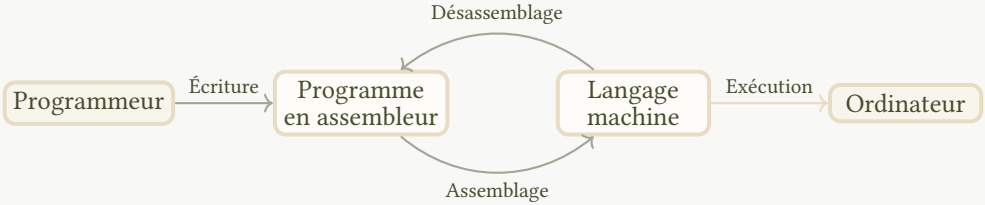
Du fait qu'un langage d'assemblage est spécifiquement dédié à un processeur donné, il existe presque autant de langages d'assemblage qu'il y a de modèles de processeurs.

Langages d'assemblage et assembleurs

L'assemblage est l'action d'un programme nommé **assembleur** qui consiste à traduire un programme en assembleur vers du langage machine.

Le désassemblage, réalisé par un **désassembleur**, est l'opération inverse. Elle permet de retrouver, à partir d'un programme en langage machine, un programme en assembleur qui lui correspond.

Voici le schéma opérationnel liant tout ceci :



L'assembleur NASM

Pour des raisons pédagogiques, nous choisissons de travailler sur une architecture **x86** en 32 bits. C'est une architecture dont le modèle de processeur est compatible avec le modèle Intel 8086.

Nous utiliserons l'**assembleur** NASM (Netwide Assembler).

Pour programmer, il faudra disposer :

1. d'un ordinateur (moderne) ;
2. d'un système Linux en 32 bits (réel ou virtuel) ou 64 bits ;
3. d'un éditeur de textes ;
4. du programme `nasm` (assembleur) ;
5. du programme `ld` (lieur) ou `gcc` ;
6. du programme `gdb` (débogueur), non indispensable à ce stade.

Généralités

Un programme assembleur est un fichier texte d'extension `.asm`.

Il est constitué de plusieurs parties dont le rôle est

1. d'invoquer des **directives** ;
2. de définir des **données initialisées** ;
3. de réserver de la mémoire pour des **données non initialisées** ;
4. de contenir une suite **instructions**.

Nous allons étudier chacune de ces parties.

Avant cela, nous avons besoin de nous familiariser avec trois ingrédients de base dans la programmation assembleur : les **valeurs**, les **registres** et la **mémoire**.

Exprimer des valeurs

Il y a plusieurs manières d'exprimer des valeurs en assembleur.

Les **entiers** sont représentés en interne selon le codage en complément à deux. Ils s'écrivent

- directement en base dix, p.ex., `0`, `10020`, `-91` ;
- en hexadécimal, avec le préfixe `0x`, p.ex., `0xA109C`, `-0x98` ;
- en binaire, avec le préfixe `0b`, p.ex., `0b001`, `0b11101`.

Les **caractères** sont représentés en interne par leur code ASCII. Ils s'écrivent

- directement, p.ex., `'a'`, `'9'` ;
- par leur code ASCII, p.ex., `10`, `120`.

Les **chaînes de caractères** sont des suites de caractères. Elles s'écrivent

- directement, p.ex., `'abbaa', 0` ;
- caractère par caractère, p.ex., `'a', 46, 36, 0`.

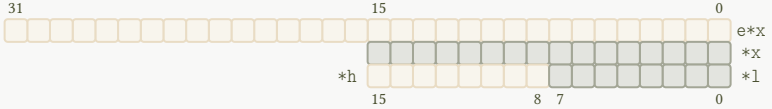
Le code ASCII du marqueur de fin de chaîne est `0` (à ne pas oublier).

Registres et sous-registres

Un registre est un emplacement de 32 bits.

On peut le considérer comme une **variable globale**.

Il y a quatre registres de travail : `eax`, `ebx`, `ecx` et `edx`. Il sont subdivisés en sous-registres selon le schéma suivant :



– Exemple –

`bh` désigne le 2^{e} octet de `ebx` et `cx` désigne les deux 1ers octets de `ecx`.

Il est possible d'écrire / lire dans chacun de ces (sous-)registres.

Attention : toute modification d'un sous-registre entraîne bien évidemment une modification du registre tout entier (et réciproquement).

Registres de diverses sortes

Il existe d'autres registres. Parmi ceux-ci, il y a

- le pointeur d'instruction `eip`, contenant l'adresse de la prochaine instruction à exécuter ;
- le pointeur de tête de pile `esp`, contenant l'adresse de la tête de la pile ;
- le pointeur de pile `ebp`, utilisé pour contenir l'adresse d'une donnée de la pile dans les fonctions ;
- le registre de drapeaux `flags`, utilisé pour contenir des informations sur le résultat d'une opération qui vient d'être réalisée.

Attention : ce ne sont pas des registres de travail, leurs rôles sont fixés. Il est possible pour certains de ces registres d'y écrire ou lire des valeurs explicitement.

L'opération `mov` — définition

L'instruction

```
mov REG, VAL
```

permet de **recopier** la valeur `VAL` dans le (sous-)registre `REG`.

– Exemples –

Voici les effets de quelques instructions :

- `mov eax, 0` eax = 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
- `mov ebx, 0xFFFFFFFF` ebx = 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
- `mov eax, 0b101` eax = 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1
- `mov al, 5` eax = * 0 0 0 0 0 1 0 1 où le symbole * dénote une valeur non modifiée.

L'opération `mov` — respect des tailles

Il est important, pour que l'instruction `mov REG, VAL` soit correcte, que la taille en octets de la valeur `VAL` soit la même que celle du (sous-)registre `REG`.

– Exemples –

Les instructions suivantes **ne sont pas correctes** :

- `mov al, 0xA5A5` Le sous-registre `al` occupe 1 octet alors que la valeur `0xA5A5` en occupe 2.
- `mov ax, ebx` Le sous-registre `ax` occupe 2 octets alors que la valeur contenue dans `ebx` en occupe 4.
- `mov eax, cx` Le sous-registre `eax` occupe 4 octets alors que la valeur contenue dans `cx` en occupe que 2.

– Exemples –

En revanche, les instructions suivantes **sont correctes** :

- `mov al, 0xA5` La valeur `0xA5` occupe bien 1 octet, tout comme `al`.
- `mov ax, 0xF` Le sous-registre `ax` occupe 2 octets alors que la valeur `0xF` n'en occupe que 1. Néanmoins, cette valeur est étendue sur 2 octets sans perte d'information en `0x000F`.

Opérations sur les registres — arithmétique

Opérations **arithmétiques** :

- `add REG, VAL` incrémente le (sous-)registre `REG` de la valeur `VAL` ;
- `sub REG, VAL` décrémente le (sous-)registre `REG` de la valeur `VAL` ;
- `mul REG` multiplie la valeur contenue dans `eax` et le (sous-)registre `REG` et place le résultat dans `edx:eax`. Cette notation désigne la suite de 64 bits dont les bits de `edx` sont ceux de poids forts et ceux de `eax` ceux de poids faibles) ;
- `div REG` place le quotient de la division de `edx:eax` par la valeur portée par le (sous-)registre `REG` dans `eax` et le reste dans `edx`.

– Exemple –

```
mov eax, 20   ; eax = 20
add eax, 51   ; eax = 71
add eax, eax  ; eax = 142

mov ebx, 3    ; ebx = 3
mul ebx       ; eax = 426

mov edx, 0    ; edx = 0
sub ebx, 1    ; ebx = 2
div ebx       ; eax = 213
```

Opérations sur les registres — logique

Opérations **logiques** :

- `not REG` place dans le (sous-)registre `REG` la valeur obtenue en réalisant le *non* bit à bit de sa valeur ;
- `and REG, VAL` place dans le (sous-)registre `REG` la valeur du *et* logique bit à bit entre les suites contenues dans `REG` et `VAL` ;
- `or REG, VAL` place dans le (sous-)registre `REG` la valeur du *ou* logique bit à bit entre les suites contenues dans `REG` et `VAL` ;
- `xor REG, VAL` place dans le (sous-)registre `REG` la valeur du *ou exclusif* logique bit à bit entre les suites contenues dans `REG` et `VAL` .

Opérations sur les registres — bit à bit

Opérations **bit à bit** :

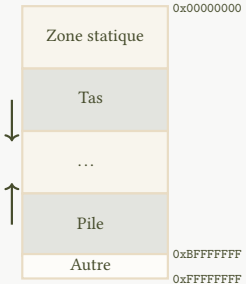
- `shl REG, NB` décale les bits du (sous-)registre `REG` à gauche de `NB` places et complète à droite par des `0` ;
- `shr REG, NB` décale les bits du (sous-)registre `REG` à droite de `NB` places et complète à gauche par des `0` ;
- `rol REG, NB` réalise une rotation des bits du (sous-)registre `REG` à gauche de `NB` places ;
- `ror REG, NB` réalise une rotation des bits du (sous-)registre `REG` à droite de `NB` places.

Mémoire

Les registres n'offrent pas assez de mémoire pour construire des programmes élaborés. C'est pour cette raison que l'on utilise la **mémoire**.

La mémoire est segmentée en plusieurs parties :

- la zone statique qui contient le code et les données statiques ;
- le tas, de taille variable au fil de l'exécution ;
- la pile, de taille variable au fil de l'exécution.



Il y a d'autres zones (non représentées ici).

En mode protégé, chaque programme en exécution possède son propre environnement de mémoire. Les adresses y sont relatives et non absolues.

De cette manière, un programme en exécution ne peut empiéter sur la mémoire d'un autre.

La lecture / écriture en mémoire suit la convention *little-endian*.

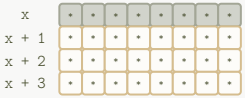
Lecture en mémoire

L'instruction

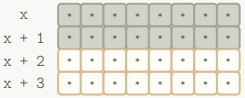
```
mov REG, [ADR]
```

place dans le (sous-)registre `REG` un, deux ou quatre octets en fonction de la taille de `REG`, **lus** à partir de l'adresse `ADR` **dans la mémoire**.

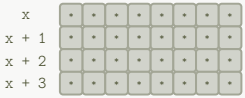
En supposant que `x` soit une adresse accessible en mémoire, les parties foncées sont celles qui sont lues et placées dans le (sous-)registre opérande de l'instruction :



`mov ah, [x]` ou `mov al, [x]`



`mov ax, [x]`



`mov eax, [x]`

Lecture en mémoire

– Exemple –

Observons l'effet des instructions suivantes sur `eax` avec la mémoire dans l'état suivant :

x	1	0	0	0	0	0	0	1
x + 1	1	1	1	1	1	1	1	1
x + 2	1	1	0	0	0	0	1	1
x + 3	1	0	1	0	1	0	1	0

```
mov eax, 0
mov al, [x]
mov ah, [x + 1]
mov ax, [x]
mov eax, [x]
mov ah, [x + 2]
mov ax, [x + 2]
```