

Projet: Réalisation d'un simulateur de CPU

Un processeur est l'unité centrale de traitement (*Central Processing Unit*, CPU) chargée d'exécuter les instructions d'un programme. Il coordonne l'accès à la mémoire, effectue des calculs et prend en charge le contrôle du flux d'exécution des programmes. Son fonctionnement repose sur un cycle d'instruction, dans lequel il récupère une instruction en mémoire, la décode, puis l'exécute en manipulant les registres et la mémoire. Le processeur est constitué de plusieurs éléments fondamentaux :

- **Les registres** : ce sont des emplacements mémoire internes au processeur permettant de stocker temporairement des valeurs nécessaires aux calculs et à l'exécution des instructions. Parmi eux, on trouve les registres généraux (AX, BX, CX, DX), les registres d'indexation et les registres de contrôle (IP, SP, FLAGS).
- **L'unité de contrôle** : elle orchestre l'exécution des instructions en gérant la récupération des instructions, leur décodage et leur exécution.
- **L'unité arithmétique et logique (ALU)** : elle est responsable des opérations mathématiques et logiques (addition, soustraction, comparaisons, etc.).
- **L'accès à la mémoire** : le processeur communique avec la mémoire vive (RAM) pour charger les instructions et stocker temporairement des données. Il utilise différentes stratégies d'adressage pour accéder aux données de manière efficace.

Dans le cadre de ce projet, nous allons implémenter une version simplifiée de ce fonctionnement, en nous concentrant sur un ensemble réduit de registres et de modes d'adressage, tout en conservant les principes fondamentaux du traitement des instructions et de la gestion de la mémoire.

Le projet sera à rendre sur moodle quelques jours avant les soutenances (date à préciser), et les soutenances auront lieu pendant la dernière séance de TD/TP. Sur moodle, vous trouverez un document récapitulatif de ce qui est attendu dans votre rendu. Pour les étudiants qui ne peuvent être présents le jour de la soutenance, vous devez impérativement contacter votre chargé de TD afin de lui fournir un justificatif et prévoir avec elle/lui une autre date de soutenance (au plus près de la semaine de rendu). En effet, les soutenances sont obligatoires, et toute absence injustifiée sera sanctionnée par la note de zéro. Enfin, nous vous rappelons que la note de ce projet fait partie de la note du module, mais ne fait pas partie du contrôle continu. Plus précisément, elle ne peut pas être compensée avec la règle du max et sera conservée en seconde session. **Attention** : Ce projet doit obligatoirement être réalisé en binôme. Si vous ne trouvez pas de binôme, vous devez envoyer un mail à votre chargé de TD pour qu'il vous en trouve un.

Exercice 1 – Implémentation d'une table de hachage générique

Dans cet exercice, nous allons implémenter une table de hachage générique en C. L'objectif est de créer une structure permettant d'associer des clés (chaînes de caractères) à des valeurs de type générique `void*`. Cette flexibilité nous permettra de concevoir un dictionnaire associant des clés à n'importe quel type de structure. La table de hachage utilisera une taille fixe de `TABLE_SIZE = 128` et appliquera un mécanisme de sondage/probing linéaire en cas de collision. Ainsi on utilisera les structures suivantes :

```
1 typedef struct hashentry{
2     char* key;
3     void* value;
4 } HashEntry;
```

```
5 typedef struct hashmap {  
6     int size;  
7     HashEntry* table;  
8 } HashMap;
```

Remarque : quand on supprime un élément dans une table de hachage avec adressage ouvert, on peut utiliser une TOMBSTONE pour éviter d’interrompre les chaînes de probing. Une TOMBSTONE est un élément spécial, distinct de la valeur nulle, qui remplace tout élément supprimé. Contrairement à NULL, elle permet de continuer la recherche et d’assurer l’accessibilité des éléments. Bien sûr, une TOMBSTONE doit être supprimée si un nouvel élément souhaite être inséré à sa position. Ici, on utilisera par exemple un TOMBSTONE définie par :

```
1 #define TOMBSTONE ((void*)-1)
```

Q 1.1 Implémentez une fonction `unsigned long simple_hash(const char *str)` permettant de convertir une chaîne de caractères en un indice dans la table de hachage.

Q 1.2 Écrivez une fonction `HashMap *hashmap_create()` permettant d’allouer dynamiquement une table de hachage et d’initialiser ses cases à NULL.

Q 1.3 Implémentez la fonction `int hashmap_insert(HashMap *map, const char *key, void *value)` permettant d’insérer un élément dans la table de hachage.

Q 1.4 Réalisez une fonction `void *hashmap_get(HashMap *map, const char *key)` permettant de récupérer un élément à partir de sa clé.

Q 1.5 Ajoutez une fonction `int hashmap_remove(HashMap *map, const char *key)` permettant de supprimer un élément de la table de hachage tout en assurant la continuité du sondage linéaire.

Q 1.6 Implémentez une fonction `void hashmap_destroy(HashMap *map)` permettant de libérer toute la mémoire allouée à la table de hachage.

Exercice 2 – Gestion dynamique de la mémoire

Dans cet exercice, nous allons implémenter un gestionnaire de mémoire simple en C. Le gestionnaire de mémoire divise l’espace mémoire en **segments** de taille variable. Chaque segment peut être alloué à un programme ou libéré pour être réutilisé. L’allocation et la libération de segments suivent des étapes bien définies. Dans notre cas, lorsqu’un segment est alloué, il est retiré de la liste des segments libres, puis il est ajouté à une **table de hachage** (correspondant aux segments alloués) en lui associant un nom unique. Lorsqu’un segment est libéré, il est supprimé de la table de hachage et réinséré dans la liste des segments libres (en le fusionnant avec d’autres segments libres adjacents s’il en existe, afin de limiter la fragmentation de la mémoire).

L’implémentation du gestionnaire reposera ainsi sur deux structures essentielles : la structure **Segment** représentant un bloc de mémoire (qui peut être libre ou alloué) et la structure **MemoryHandler** représentant l’état global du gestionnaire de mémoire.

```
1 typedef struct segment {  
2     int start; // Position de debut (adresse) du segment dans la memoire  
3     int size; // Taille du segment en unites de memoire  
4     struct Segment *next; // Pointeur vers le segment suivant dans la liste chainee  
5 } Segment;  
6
```

```
7 typedef struct memoryHandler {  
8     void **memory; // Tableau de pointeurs vers la memoire allouee  
9     int total_size; // Taille totale de la memoire gereee  
10    Segment *free_list; // Liste chainee des segments de memoire libres  
11    HashMap *allocated; // Table de hachage (nom, segment)  
12 } MemoryHandler;
```

Lors de l'initialisation du gestionnaire, la mémoire est constituée d'un **unique segment libre** couvrant la totalité de l'espace mémoire disponible. Par exemple, pour une mémoire initiale de taille 1024, la liste des segments libres contient un unique segment : `[start=0, size=1024]`.

Q 2.1 Implémentez une fonction `MemoryHandler *memory_init(int size)` permettant d'initialiser le gestionnaire de mémoire.

On souhaite à présent écrire une fonction qui permet de savoir s'il est possible d'allouer un segment commençant à une position donnée. Cette allocation est possible s'il existe un segment libre contenant le segment à allouer, c-à-d s'il existe un segment libre `seg` tel que `seg->start <= start` et `seg->end >= start+size`, où `start` est l'adresse du segment que l'on souhaite allouer et `size` est sa taille.

Q 2.2 Implémentez une fonction `find_free_segment(MemoryHandler* handler, int start, int size, Segment** prev)` qui retourne un tel segment libre s'il en existe, l'argument `prev` permettant alors de récupérer un pointeur sur le segment libre qui le précède dans la liste chaînée `free_list`. Cette fonction retourne NULL dans le cas contraire.

On s'intéresse maintenant à l'allocation de segments. L'allocation d'un segment nommé `X` de taille `size` à l'adresse `start` suit les étapes suivantes :

1. Vérifier qu'un espace mémoire libre suffisant est disponible à l'adresse indiquée (à l'aide de la fonction précédente).
2. Si c'est le cas, créer un nouveau segment `new_seg` de taille `size` à l'adresse `start`, puis l'ajouter à la table de hachage `allocated`, avec `X` comme clé.
3. Remplacer dans `free_list` le segment libre qui contient `new_seg` par deux segments libres : un avant `start` et un après `start+size`.

Exemple : supposons une mémoire initiale de taille **1024**. Si l'utilisateur commence par allouer un segment nommé `"data"` de **100 unités** à l'adresse **200**, l'état devient :

- La table des segments alloués contient le segment `[200, 100]`, avec `"data"` comme clé.
- La liste des segments libres comporte désormais deux segments : `[0, 200]` et `[300, 724]`.

Q 2.3 Implémentez une fonction `int create_segment(MemoryHandler *handler, const char *name, int start, int size)` permettant d'allouer dynamiquement un segment de mémoire de taille `size` à l'adresse mémoire `start`.

La libération d'un segment consiste essentiellement à le retirer de la table de hachage `allocated` pour l'ajouter dans la liste `free_list`. Cet ajout doit cependant se faire de manière à optimiser la réutilisation des espaces libres, c'est-à-dire en fusionnant ce nouveau segment libre avec les éventuels segments libres qui lui sont adjacents. Sur l'exemple précédent, la libération de `[200, 100]` doit ainsi conduire à l'unique segment libre `[0, 1024]` résultant de la fusion de `[200, 100]` avec les segments libres adjacents `[0, 200]` et `[300, 724]`.

Q 2.4 Implémentez une fonction `int remove_segment(MemoryHandler *handler, const char *name)` permettant de libérer un segment de mémoire alloué et de l'ajouter à la liste des segments libres comme décrit ci-dessus.

Exercice 3 – Conception d'un parser pour un langage pseudo-assembleur

Un **parser** est un programme ou une fonction qui analyse une entrée textuelle pour en extraire une structure exploitable par un ordinateur. Il est souvent utilisé pour interpréter du code source en vue de son exécution ou de sa traduction vers un autre langage. Dans notre cas, nous allons développer un parser capable de comprendre un **pseudo-assembleur**, c'est-à-dire un langage inspiré de l'assembleur mais simplifié pour des besoins pédagogiques ou spécifiques.

Un programme écrit en pseudo-assembleur est organisé en **sections**, qui sont des parties distinctes du code ayant chacune un rôle précis. Les deux sections fondamentales sont :

- **La section .DATA**, qui contient les déclarations de variables et de constantes. Par exemple, l'instruction "X DW 10" permet de déclarer une variable de type DW nommée X et initialisée à la valeur 10.
- **La section .CODE**, qui regroupe les instructions exécutables du programme. Par exemple, l'instruction "MOV AX,7" met la valeur 7 dans AX.

Le parser devra identifier ces sections et extraire les instructions qui y figurent. Pour cela, le parser reposera sur la structure **Instruction** suivante :

```
1 typedef struct {
2     char *mnemonic;    // Instruction mnemonic (ou nom de variable pour .DATA)
3     char *operand1;    // Premier operande (ou type pour .DATA)
4     char *operand2;    // Second operande (ou initialisation pour .DATA)
5 } Instruction;
```

Plus précisément :

- **mnemonic** représente le mnémonique de l'instruction (par exemple MOV, ADD, JMP...) ou le nom de la variable pour les instructions dans **.DATA**.
- **operand1** (optionnel) représente le premier opérande (registre, variable, valeur immédiate) ou le type de la variable dans **.DATA** (par exemple DW, DB...).
- **operand2** (optionnel) second opérande ou valeur de la variable dans **.DATA**.

Le parser devra aussi prendre en compte les **étiquettes/labels**, utilisés pour les sauts conditionnels et les boucles. Plus précisément, une étiquette est un identifiant symbolique permettant de marquer des positions dans le code (l'adresse de la ligne où se trouve le label). Elle est placée en tout début de ligne, suivie du caractère ":", et précède généralement le mnémonique de l'instruction. Par exemple, dans le code suivant, on a deux étiquettes : "loop" et "end".

```
1 .DATA
2 x DW 15
3 .CODE
4 loop: MOV AX,6
5 ADD BX,10
6 end: JMP loop
```

L'étiquette "loop" marque ici la position de la première ligne de code, ce qui permet à l'instruction "JMP loop" de revenir à cette ligne pour répéter les opérations "MOV AX,6" et "ADD BX,10".

Finalement, le résultat du parsing sera stockée dans la structure **PaserResult** suivante :

```
1 typedef struct {
2     Instruction **data_instructions; // Tableau d'instructions .DATA
3     int data_count;                // Nombre d'instructions .DATA
4     Instruction **code_instructions; // Tableau d'instructions .CODE
```

```

5 |     int code_count;           // Nombre d'instructions .CODE
6 |     HashMap *labels;         // labels -> indices dans code_instructions
7 |     HashMap *memory_locations; // noms de variables -> adresse memoire
8 | } ParserResult;

```

Cette structure contient tous les éléments extraits du fichier, c'est-à-dire :

- `data_instructions` : tableau de pointeurs vers des structures `Instruction` correspondant à la section `.DATA`.
- `data_count` : nombre d'éléments dans le tableau `data_instructions`.
- `code_instructions` : tableau de pointeurs vers des structures `Instruction` pour la section `.CODE`.
- `code_count` : nombre d'éléments dans le tableau `code_instructions`.
- `labels` : table de hachage associant les noms d'étiquettes (comme "loop") à leur position dans le tableau `code_instructions`.
- `memory_locations` : table de hachage associant les noms de variables à son adresse séquentielle.

Q 3.1 Implémentez une fonction `Instruction *parse_data_instruction(const char *line, HashMap *memory_locations)` qui permet d'analyser et stocker une ligne de la section `.DATA` d'un programme en pseudo-assembleur. Cette fonction décompose la ligne en trois éléments : le nom de la variable (`mnemonic`), son type (`operand1`, par exemple `DW` ou `DB`) et sa valeur (`operand2`). Ces trois éléments sont séparés par des espaces dans la ligne. Cette fonction associe également chaque variable à son adresse séquentielle dans le dictionnaire `memory_locations`, en tenant compte du nombre d'éléments déclarés. Par exemple, considérons les déclarations suivantes :

```

1 | .DATA
2 | X DW 3
3 | Y DB 4
4 | arr DB 5,6,7,8
5 | z DB 9

```

L'analyse de ces lignes par `parse_data_instruction` devrait produire :

$$\text{data_instructions : } \begin{bmatrix} \text{"X"} & \text{"DW"} & \text{"3"} \\ \text{"y"} & \text{"DB"} & \text{"4"} \\ \text{"arr"} & \text{"DB"} & \text{"5,6,7,8"} \\ \text{"z"} & \text{"DB"} & \text{"9"} \end{bmatrix} \quad \text{memory_locations : } \left\{ \begin{array}{l} \text{"X"} \rightarrow 0 \\ \text{"y"} \rightarrow 1 \\ \text{"arr"} \rightarrow 2 \\ \text{"z"} \rightarrow 6 \end{array} \right\}$$

L'adresse associée à chaque variable dans `memory_locations` représente l'emplacement mémoire du premier élément de cette variable. En particulier, `z` est ici associée à l'adresse 6 car `X` occupe l'adresse 0 (un seul élément), `y` occupe l'adresse 1 (un seul élément) et `arr` occupe les adresses 2 à 5 (4 éléments).

Remarque : Dans cet assembleur simplifié, pour le calcul des adresses mémoire, on ne distinguera pas les types `DW` et `DB`. Chaque élément occupe une seule unité de mémoire, indépendamment de son type.

Q 3.2 Implémentez une fonction `Instruction *parse_code_instruction(const char *line, HashMap *labels, int code_count)` qui permet d'analyser et stocker une ligne de la section `.CODE`. Une ligne instruction commence éventuellement par une étiquette, suivie du mnémonique, un espace, puis les deux opérandes séparées par une virgule (le deuxième opérande étant facultatif). Si une étiquette est présente, il faut l'ajouter à la table de hachage `labels`, avec pour valeur `code_count` qui correspond au numéro de la ligne dans `.CODE` (les indices commençant à zéro). Par exemple, considérons la section `.CODE` suivante :

```

1 .CODE
2 loop: MOV AX,6
3 ADD BX,10
4 end: JMP loop

```

L'analyse de ces lignes par `parse_code_instruction` devrait produire ces structures :

$$\text{code_instructions : } \begin{bmatrix} \text{"MOV"} & \text{"AX"} & \text{"6"} \\ \text{"ADD"} & \text{"BX"} & \text{"10"} \\ \text{"JMP"} & \text{"loop"} & \text{""} \end{bmatrix} \quad \text{labels : } \left\{ \begin{array}{l} \text{"loop"} \rightarrow 0 \\ \text{"end"} \rightarrow 2 \end{array} \right\}$$

Q 3.3 Implémentez une fonction `ParserResult *parse(const char *filename)` qui analyse un fichier assembleur complet en identifiant les sections `.DATA` et `.CODE` et en traitant chaque ligne de la manière appropriée. Elle doit construire une structure complète `ParserResult` qui contient tous les éléments extraits du fichier. La fonction doit gérer l'allocation dynamique des tableaux d'instructions (`data_instructions` et `code_instructions`) et redimensionner ces tableaux selon les besoins. Par exemple, pour un fichier contenant

```

1 .DATA
2 x DW 42
3 arr DB 20,21,22,23
4 y DB 10
5 .CODE
6 start: MOV AX,x
7 loop: ADD AX,y
8 JMP loop

```

la fonction `parse` doit créer une structure `ParserResult` complète :

$$\begin{aligned} \text{data_instructions : } & \begin{bmatrix} \text{"x"} & \text{"DW"} & \text{"42"} \\ \text{"y"} & \text{"DB"} & \text{"10"} \\ \text{"arr"} & \text{"DB"} & \text{"20,21,22,23"} \end{bmatrix} & \text{code_instructions : } & \begin{bmatrix} \text{"MOV"} & \text{"AX"} & \text{"x"} \\ \text{"ADD"} & \text{"AX"} & \text{"y"} \\ \text{"JMP"} & \text{"loop"} & \text{""} \end{bmatrix} \\ \text{labels : } & \left\{ \begin{array}{l} \text{"start"} \rightarrow 0 \\ \text{"loop"} \rightarrow 1 \end{array} \right\} & \text{memory_locations : } & \left\{ \begin{array}{l} \text{"x"} \rightarrow 0 \\ \text{"arr"} \rightarrow 1 \\ \text{"y"} \rightarrow 5 \end{array} \right\} \\ \text{data_count} = 3 & & \text{code_count} = 3 & \end{aligned}$$

Q 3.4 Tester votre fonction `parse` sur l'exemple de la question précédente.

Q 3.5 Écrivez une fonction `void free_parser_result(ParserResult *result)` permettant de supprimer un élément de type `ParserResult`.

Exercice 4 – Allocation d'un segment de données

Dans cet exercice, nous allons implémenter les mécanismes permettant d'allouer un segment en fonction des déclarations de variables récupérées par le parser. Le nom de ce segment de données sera "DS" (pour "data segment"). Nous allons aussi nous intéresser au stockage de ces données, à leur récupération et à leur affichage. Toutes ses opérations seront réalisées par notre processeur simulé (CPU) composé des éléments suivants :

```

1 typedef struct {
2     MemoryHandler* memory_handler; // Gestionnaire de memoire
3     HashMap *context;             // Registres (AX, BX, CX, DX)
4 } CPU;

```

Un CPU est donc composé d'un **gestionnaire de mémoire** (`MemoryHandler`) permettant d'allouer et de gérer la mémoire du programme et d'une **table de hachage** (`HashMap`) permettant de stocker les valeurs des registres (emplacements mémoire internes au processeur). Pour simplifier, nous ne considérons ici que les registres généraux `AX`, `BX`, `CX`, `DX` et nous supposons qu'ils correspondent à des entiers.

Q 4.1 Implémentez une fonction `CPU *cpu_init(int memory_size)` permettant d'initialiser le processeur simulé en :

- Allouant dynamiquement une structure CPU.
- Initialisant son gestionnaire de mémoire de taille `memory_size`.
- Créant une table de hachage contenant les registres généraux `AX`, `BX`, `CX`, `DX` en leur associant la valeur zéro.

Q 4.2 Implémentez une fonction `void cpu_destroy(CPU *cpu)` permettant de libérer toutes les ressources allouées par le processeur simulé, notamment les registres, la table de hachage et le gestionnaire de mémoire.

On s'intéresse maintenant au stockage et à la récupération des données.

Q 4.3 Implémentez une fonction `void* store(MemoryHandler *handler, const char *segment_name, int pos, void *data)` permettant de stocker une donnée `data` à la position `pos` de `segment_name`. Cette fonction devra d'abord vérifier que `segment_name` est bien un segment alloué, et que `pos` ne dépasse pas la taille du segment. Si c'est le cas, la donnée est insérée dans le champs `memory` à la bonne position (tenant compte de la position de début du segment). Par exemple, si le segment alloué est `[200, 100]` et que l'on souhaite ajouter la donnée à la position 50 de ce segment, la donnée sera stockée dans `memory` à la position 250.

Q 4.4 Implémentez une fonction `void* load(MemoryHandler *handler, const char *segment_name, int pos)` permettant de récupérer la donnée stockée à la position `pos` de `segment_name`.

L'allocation des données par le CPU consiste en les étapes suivantes :

1. **Calcul de l'espace nécessaire au stockage des variables.** Dans l'exemple de la question 3.1, les variables `x`, `y` et `z` ont besoin de 1 emplacement chacune, alors que la variable `arr` a besoin de 4 emplacements. Au total, il faudra donc déclarer un segment de taille 7. Dans le cadre de ce projet, on considère que les données à stocker sont des entiers (pour simplifier).
2. **Déclaration d'un segment nommé DS dans le MemoryHandler.** Sur l'exemple, après déclaration du segment, le `MemoryHandler` contient un seul segment libre (`[7, 1017]` si la taille initiale est 1024) et un seul segment alloué correspondant au segment de données (`{DS : [0, 7]}`).
3. **Remplissage du segment de données avec les valeurs déclarées dans la section .DATA du code.** Sur notre exemple, le tableau `memory` évolue ainsi :

`[NULL, NULL, ..., NULL] → [p1, p2, p3, p4, p5, p6, p7, NULL, ..., NULL]`

où `p1` est un pointeur sur un entier contenant 3 (la valeur de `x`), `p2` est un pointeur sur un entier contenant 4 (la valeur de `y`), `p3` est un pointeur sur un entier contenant 5 (la valeur du premier élément de `arr`), etc.

Q 4.5 Implémentez une fonction `void allocate_variables(CPU *cpu, Instruction** data_instructions, int data_count)` permettant d'allouer dynamiquement le segment de données en fonction des déclarations récupérées par le parser.

Q 4.6 Implémentez une fonction `void print_data_segment(CPU *cpu)` permettant d'afficher le contenu du segment de données (nom "DS").

Exercice 5 – Expressions régulières et résolution d'adressage

Dans cet exercice, nous allons utiliser des expressions régulières (*regex*) en C pour identifier les différents modes d'adressage en pseudo-assembleur. Chaque mode d'adressage détermine comment une opérande doit être interprétée et où se trouve la valeur manipulée. Dans le cadre de ce projet, nous nous intéressons aux modes d'adressage suivants :

- **Adressage immédiat** : l'opérande est une valeur numérique directement contenue dans l'instruction. Par exemple, la valeur 42 est immédiate dans l'instruction "MOV AX,42".
- **Adressage par registre** : l'opérande est le nom d'un registre du processeur. Par exemple, BX est le registre source de l'instruction "MOV AX,BX".
- **Adressage direct** : l'opérande spécifie directement l'adresse en mémoire où se trouve la valeur. Par exemple, on écrit "MOV AX,[5]" pour indiquer que la valeur à charger se trouve à la position 5 du segment de données.
- **Adressage indirect par registre** : l'opérande spécifie un registre qui contient l'adresse mémoire où se trouve la valeur. Par exemple, on écrit "MOV BX,[AX]" pour indiquer que la valeur à charger est celle située à l'adresse contenue dans le registre AX.

Les expressions régulières permettent de reconnaître et d'extraire des motifs spécifiques dans du texte. Elles sont couramment utilisées pour valider des entrées utilisateur, rechercher des motifs dans un fichier ou transformer des chaînes de caractères. Dans le cadre de ce projet, notre objectif est de formuler des expressions régulières permettant d'identifier le mode d'adressage. Par exemple, pour l'adressage immédiat, au lieu de tester si chaque caractère de l'opérande est bien un chiffre entre 0 et 9, il s'agit d'écrire une expression régulière (un motif) caractérisant la syntaxe d'un nombre de manière générale.

Avant de chercher à détecter les différents modes d'adressage, voyons quelques éléments fondamentaux des expressions régulières utiles à leur construction :

- "[0-9]" : Correspond à un chiffre.
- "[A-Z]" : Correspond à une lettre majuscule.
- "[a-zA-Z]" : Correspond à n'importe quelle lettre, majuscule ou minuscule.
- "A" : Correspond exactement au caractère 'A'.
- "^" : Indique le début de la chaîne de caractères.
- "\$" : Indique la fin de la chaîne de caractères.
- "+" : Quantificateur indiquant au moins une occurrence du motif qui le précède.
- "*" : Quantificateur indiquant zéro ou plusieurs occurrences du motif qui le précède.
- "{n}" : Indique exactement n occurrences du motif qui le précède.
- "()" : Groupe des expressions pour les traiter comme une unité.
- "|" : Opérateur logique "OU" permettant d'associer plusieurs motifs alternatifs.

Si on souhaite que le motif contiennent un caractère spécial (comme [,), *, +, etc.), il faut utiliser un échappement. Par exemple, l'expression régulière `^(\\+33|0)[1-9][0-9]{8}$` permet de spécifier le

format d'un numéro de téléphone français, qui commence soit par +33 soit par 0, suivi d'un chiffre entre 1 et 9, puis de 8 autres chiffres. Autre exemple, l'expression régulière `^[A-Z][a-z]*([A-Z][a-z]*)*$` permet de spécifier le format d'un nom de famille, qui commence par une majuscule, suivie de minuscules, et qui peut être suivie par d'autres mots au même format (précédés par un espace).

Nous devons maintenant utiliser ces constructions de base pour concevoir des expressions régulières qui correspondent aux différents types d'adressage en assembleur. En langage C, vous pouvez tester des expressions régulières à l'aide de la fonction suivante :

```
1 int matches(const char *pattern, const char *string) {
2     regex_t regex;
3     int result = regcomp(&regex, pattern, REG_EXTENDED);
4     if (result) {
5         fprintf(stderr, "Regex compilation failed for pattern: %s\n", pattern);
6         return 0;
7     }
8     result = regexec(&regex, string, 0, NULL, 0);
9     regfree(&regex);
10    return result == 0;
11 }
```

Cette fonction compile le motif, vérifie sa correspondance avec une chaîne, libère les ressources, puis retourne un booléen indiquant si une correspondance a été trouvée.

Adressage immédiat

L'adressage immédiat traite des constantes littérales (comme "42") contrairement aux autres modes d'adressage qui manipulent des adresses mémoire. Pour que toutes les fonctions d'adressage retournent uniformément des pointeurs `void*`, nous devons allouer de la mémoire pour ces valeurs. Pour cela, on va utiliser un pool de constantes, qui stocke chaque valeur une seule fois dans une table de hachage : la clé est la chaîne de caractères (ex : "42") et la valeur est un pointeur vers la mémoire allouée. Quand une valeur immédiate est rencontrée, on vérifie d'abord si elle existe déjà dans le pool avant de créer une nouvelle allocation. Cela évite les duplications tout en maintenant l'uniformité de l'interface.

Pour implémenter ce pool de constantes, on modifie la structure CPU de la manière suivante :

```
1 typedef struct CPU {
2     // Champs existants...
3     HashMap *constant_pool; // Table de hachage pour stocker les valeurs immédiates
4 } CPU;
```

Q 5.1 Mettez à jour les fonctions `CPU *cpu_init(int memory_size)` et `void cpu_destroy(CPU *cpu)` pour prendre en compte le constant pool.

Q 5.2 Implémentez une fonction `void *immediate_addressing(CPU *cpu, const char *operand)` permettant de traiter l'adressage immédiat. La fonction doit vérifier si l'opérande correspond à ce mode d'adressage (en utilisant une regex), stocker cette valeur dans le constant pool si elle n'y est pas déjà, et retourner un pointeur vers cette valeur stockée.

Adressage par registre

L'adressage par registre est un mode d'adressage où l'opérande est le nom d'un registre du processeur. Dans notre cas, les registres existants sont AX, BX, CX et DX.

Q 5.3 Implémentez une fonction `void *register_addressing(CPU *cpu, const char *operand)` permettant de traiter l'adressage par registre. Cette fonction doit vérifier si l'opérande fourni correspond au format d'un nom de registre (à l'aide d'une regex), puis vérifier si ce registre existe dans le contexte du CPU. Si le registre existe, la fonction retourne un pointeur vers la valeur du registre, ce qui permettra de lire ou modifier son contenu.

Adressage direct

L'adressage direct par mémoire est un mode d'adressage où l'opérande spécifie directement l'adresse en mémoire à laquelle accéder. L'opérande est alors de la forme "[i]" où i est sa position dans le segment de données.

Q 5.4 Implémentez une fonction `void *memory_direct_addressing(CPU *cpu, const char *operand)` permettant de traiter l'adressage direct par mémoire. Cette fonction doit vérifier si l'opérande fourni correspond au format d'une adresse mémoire directe (à l'aide d'une regex), extraire l'adresse numérique, puis accéder à la valeur stockée à cette adresse dans le segment de données. Elle retourne un pointeur vers cette valeur, ce qui permettra de la lire ou de la modifier dans le cadre d'une instruction.

Adressage indirect par registre

L'adressage indirect par registre est un mode d'adressage où l'opérande spécifie un registre qui contient l'adresse mémoire à laquelle accéder. L'opérande est alors de la forme "[XX]" où XX est le nom d'un registre du processeur.

Q 5.5 Implémentez une fonction `void *register_indirect_addressing(CPU *cpu, const char *operand)` permettant de traiter l'adressage indirect par registre. Cette fonction doit vérifier si l'opérande correspond au format d'un registre entre crochets (à l'aide d'une regex), extraire le nom du registre, récupérer sa valeur qui servira d'adresse mémoire, puis accéder à la valeur stockée à cette adresse dans le segment de données. Elle retourne un pointeur vers cette valeur, ce qui permettra de la lire ou de la modifier dans le cadre d'une instruction.

Les fonctions que nous venons d'écrire permettent de résoudre différents types d'adressage, en retournant un pointeur sur la valeur indiquée par l'opérande. Ceci permet notamment de simuler des instructions en pseudo-assembleur avec des fonctions C qui prennent en argument ces pointeurs.

Q 5.6 Implémentez une fonction `void handle_MOV(CPU* cpu, void* src, void* dest)` permettant de simuler le comportement de l'instruction MOV en pseudo-assembleur. Cette fonction prend un pointeur vers l'emplacement de la valeur source et un pointeur vers l'emplacement de destination, puis copie la valeur depuis la source vers la destination.

Q 5.7 En utilisant la fonction suivante (permettant de créer une structure CPU et d'initialiser les valeurs des registres), écrivez un main permettant de tester les 4 types d'adressage en utilisant la fonction `handle_MOV`.

```
1 CPU* setup_test_environment() {
2     // Initialiser le CPU
3     CPU *cpu = cpu_init(1024);
4     if (!cpu) {
5         printf("Error: -CPU- initialization - failed\n");
6         return NULL;
7     }
8 }
```

```

9      // Initialiser les registres avec des valeurs spécifiques
10     int *ax = (int *)hashmap_get(cpu->context, "AX");
11     int *bx = (int *)hashmap_get(cpu->context, "BX");
12     int *cx = (int *)hashmap_get(cpu->context, "CX");
13     int *dx = (int *)hashmap_get(cpu->context, "DX");
14
15     *ax = 3;
16     *bx = 6;
17     *cx = 100;
18     *dx = 0;
19
20     // Créer et initialiser le segment de données
21     if (!hashmap_get(cpu->memory_handler->allocated, "DS")) {
22         create_segment(cpu->memory_handler, "DS", 0, 20);
23
24         // Initialiser le segment de données avec des valeurs de test
25         for (int i = 0; i < 10; i++) {
26             int *value = (int *)malloc(sizeof(int));
27             *value = i * 10 + 5; // Valeurs 5, 15, 25, 35...
28             store(cpu->memory_handler, "DS", i, value);
29         }
30     }
31     printf("Test-environment initialized.\n");
32     return cpu;
33 }

```

Q 5.8 Implémentez une fonction `void *resolve_addressing(CPU *cpu, const char *operand)` permettant d'identifier automatiquement le mode d'adressage d'un opérande et de résoudre sa valeur. Cette fonction doit essayer successivement chaque mode d'adressage implémenté précédemment et retourner le résultat du premier mode applicable.

Exercice 6 – Allocation d'un segment de codes et exécution

Dans cet exercice, nous nous intéressons à l'allocation d'un segment correspondant aux lignes de codes récupérées par le parser. Le nom de ce segment de codes sera "CS" (pour "code segment"). Nous nous intéresserons ensuite à l'exécution du code par notre processeur simulé.

Avant d'effectuer l'allocation d'un segment de codes, nous devons réaliser un prétraitement qui consiste à substituer les noms des variables par leur adresse dans le segment de données et les noms d'étiquettes par leur adresse dans le segment de code. Pour cela, nous utiliserons la fonction `search_and_replace(char **str, HashMap *values)` qui pour chaque clé `k` de la table de hachage `values` remplace chacune de ses occurrences par sa valeur (`values[k]`).

```

1  char *trim(char *str) {
2      while (*str == ' ' || *str == '\t' || *str == '\n' || *str == '\r') str++;
3
4      char *end = str + strlen(str) - 1;
5      while (end > str && (*end == ' ' || *end == '\t' || *end == '\n' || *end == '\r'))
6          {
7              *end = '\0';
8              end--;
9          }
10     return str;
11 }
12
13 int search_and_replace(char **str, HashMap *values) {
14     if (!str || !*str || !values) return 0;

```

```

14
15     int replaced = 0;
16     char *input = *str;
17
18     // Iterate through all keys in the hashmap
19     for (int i = 0; i < values->size; i++) {
20         if (values->table[i].key && values->table[i].key != (void *)-1) {
21             char *key = values->table[i].key;
22             int value = (int)(long)values->table[i].value;
23
24             // Find potential substring match
25             char *substr = strstr(input, key);
26             if (substr) {
27                 // Construct replacement buffer
28                 char replacement[64];
29                 snprintf(replacement, sizeof(replacement), "%d", value);
30
31                 // Calculate lengths
32                 int key_len = strlen(key);
33                 int repl_len = strlen(replacement);
34                 int remain_len = strlen(substr + key_len);
35
36                 // Create new string
37                 char *new_str = (char *)malloc(strlen(input) - key_len + repl_len +
38                     1);
39                 strncpy(new_str, input, substr - input);
40                 new_str[substr - input] = '\0';
41                 strcat(new_str, replacement);
42                 strcat(new_str, substr + key_len);
43
44                 // Free and update original string
45                 free(input);
46                 *str = new_str;
47                 input = new_str;
48
49                 replaced = 1;
50             }
51         }
52     }
53
54     // Trim the final string
55     if (replaced) {
56         char *trimmed = trim(input);
57         if (trimmed != input) {
58             memmove(input, trimmed, strlen(trimmed) + 1);
59         }
60     }
61     return replaced;
62 }

```

Par exemple, pour la chaîne de caractères `str = "[X]"` et la table de hachage suivante

$$\text{values} : \left\{ \begin{array}{l} \text{"X"} \rightarrow 0 \\ \text{"Y"} \rightarrow 1 \end{array} \right\}$$

la fonction `search_and_replace` doit retourner `str = "[0]"`. Dans notre contexte, il s'agit d'utiliser cette fonction pour transformer le tableau `code_instructions` du parser. Par exemple, on doit pouvoir réaliser la transformation suivante

$$\begin{bmatrix} \text{"MOV"} & \text{"AX"} & \text{"[x]"} \\ \text{"MOV"} & \text{"BX"} & \text{"[y]"} \\ \text{"JMP"} & \text{"loop"} & \text{""} \end{bmatrix} \rightarrow \begin{bmatrix} \text{"MOV"} & \text{"AX"} & \text{"[0]"} \\ \text{"MOV"} & \text{"BX"} & \text{"[5]"} \\ \text{"JMP"} & \text{"1"} & \text{""} \end{bmatrix}$$

à partir des tables de hachage `labels` et `memory_locations` suivantes

$$\text{labels} : \left\{ \begin{array}{l} \text{"start"} \rightarrow 0 \\ \text{"loop"} \rightarrow 1 \end{array} \right\} \quad \text{memory_locations} : \left\{ \begin{array}{l} \text{"x"} \rightarrow 0 \\ \text{"arr"} \rightarrow 1 \\ \text{"y"} \rightarrow 5 \end{array} \right\}$$

Q 6.1 Implémentez une fonction `int resolve_constants(ParserResult *result)` qui remplace les variables par leur adresse dans le segment de données et les étiquettes par leur adresse dans le code.

Pour que notre processeur simulé puisse gérer l'exécution du code, nous allons ajouter à son contexte les registres suivants :

- **IP** (pour "instruction pointer") : sert à indiquer le numéro de la prochaine instruction à exécuter.
- **ZF** (pour "zero flag") : sert à indiquer si la dernière opération a abouti à un résultat nul (valeur 1) ou non (valeur 0).
- **SF** (pour "sign flag") : sert à indiquer si la dernière opération a abouti à un résultat strictement négatif (valeur 1) ou non (valeur 0).

Par exemple, l'instruction `CMP AX BX` compare la valeur des registres `AX` et `BX` en soustrayant la valeur de `BX` à `AX`. Si la valeur de `AX` est plus petite que celle de `BX`, alors `ZF` et `SF` auront pour valeur 0 et 1 respectivement après l'exécution de cette opération.

Q 6.2 Modifiez la fonction `CPU *cpu_init(int memory_size)` afin de créer les registres `IP`, `ZF` et `SF` au sein du contexte en les initialisant à zéro.

Q 6.3 Implémentez `void allocate_code_segment(CPU *cpu, Instruction **code_instructions, int code_count)` qui alloue et initialise le segment de codes (CS). Cette fonction doit allouer un segment de taille `code_count`, y stocker les instructions (préalablement prétraitées), et initialiser le registre `IP` à 0.

Q 6.4 Implémentez une fonction `int handle_instruction(CPU *cpu, Instruction *instr, void *src, void *dest)` qui généralise la fonction `handle_MOV` en permettant d'exécuter une instruction dans le CPU en fonction de son mnémonique. Cette fonction doit au moins gérer les opérations suivantes :

- `MOV dst, src` : transfère la valeur de la source vers la destination.
- `ADD dst, src` : additionne les valeurs de la source et de la destination, puis stocke le résultat dans la destination.
- `CMP dst, src` : effectue une soustraction logique `dst - src` (sans modifier les opérandes) et met à jour des drapeaux (`ZF = 1` si et seulement si `dst = src` et `SF = 1` si et seulement si `dst < src`).
- `JMP address` : modifie le pointeur d'instruction `IP` pour changer le flux d'exécution, c'est-à-dire `IP ← address`.
- `JZ address` : effectue un branchement conditionnel si le résultat précédent était zéro, c'est-à-dire si `ZF = 1` alors `IP ← address`.
- `JNZ address` : effectue un branchement conditionnel si le résultat précédent était non-zéro, c'est-à-dire si `ZF = 0` alors `IP ← address`.

- **HALT** : termine l'exécution en positionnant le pointeur d'instruction **IP** à la fin du segment de codes.

Q 6.5 Implémentez une fonction `int execute_instruction(CPU *cpu, Instruction *instr)` qui permet de résoudre les adresses des opérandes en fonction du type d'adressage, puis délègue l'exécution proprement dite à la fonction `handle_instruction`.

Q 6.6 Implémentez une fonction `Instruction* fetch_next_instruction(CPU *cpu)` qui récupère l'instruction suivante dans le segment de codes et incrémente le pointeur d'instruction (**IP**). La fonction doit vérifier que le registre **IP** et le segment "**CS**" existent, que l'**IP** est dans les limites valides, avant de charger l'instruction depuis la mémoire et d'incrémenter le registre **IP**.

Q 6.7 Implémentez une fonction `int run_program(CPU *cpu)` qui exécute un programme préalablement chargé dans le CPU en mode pas à pas. Cette fonction doit :

- afficher l'état initial du CPU (segment de données, registres),
- exécuter les instructions séquentiellement (en permettant à l'utilisateur de contrôler le déroulement),
- afficher l'état final du CPU à la fin de l'exécution.

Pour une meilleure expérience, l'utilisateur doit pouvoir appuyer sur la touche "Entrée" pour exécuter l'instruction suivante et sur "q" pour quitter l'exécution.

Exercice 7 – Implementation d'un segment de pile (Stack Segment)

Dans cet exercice, nous nous intéressons au segment de pile, nommé "**SS**" (pour "stack segment"), qui est une zone mémoire dédiée à la gestion dynamique des données temporaires durant l'exécution d'un programme. Sa manipulation repose principalement sur deux registres essentiels :

- **SP** (pour "Stack Pointer") : pointe toujours vers le sommet de la pile, indiquant la dernière adresse utilisée. Chaque opération "**PUSH**" décrémente **SP**, tandis que chaque opération "**POP**" l'incrémente.
- **BP** (pour "Base Pointer") : fournit un point de référence stable pour accéder aux paramètres et variables locales d'une fonction, servant de base pour le référencement des données sur la pile. Dans notre cas, il s'agit d'un pointeur sur le bas de la pile, c'est-à-dire l'adresse mémoire la plus grande (car la pile s'étend vers les adresses mémoire inférieures).

Q 7.1 Redéfinissez la fonction `CPU *cpu_init(int memory_size)` pour gérer le segment de pile. Plus précisément, la fonction doit allouer dynamiquement les registres **SP** et **BP**, créer le segment de pile "**SS**" de taille 128 (positionné sous le segment de données) et initialiser les pointeurs de pile.

Q 7.2 Implémentez les fonctions `int push_value(CPU *cpu, int value)` et `int pop_value(CPU *cpu, int *dest)`. Les fonctions doivent manipuler le pointeur **SP**, stocker/charger des valeurs dans le segment de pile **SS**, et retourner -1 en cas d'erreur. En particulier, une erreur se produit quand on essaie de dépiler (**POP**) un élément alors que la pile est vide (on parle de "stack underflow") ou quand on essaie d'empiler (**PUSH**) un élément alors que la pile est pleine.

Q 7.3 Étendez la fonction `handle_instruction()` pour supporter les mnémoniques **PUSH** et **POP**. L'implémentation doit :

- Gérer **PUSH** en utilisant la valeur d'un registre source (ou celle de **AX**, si aucun registre n'a été spécifié).
- Gérer **POP** en stockant la valeur dépilée dans un registre destination (ou dans **AX**, si aucun registre n'a été spécifié).

Exercice 8 – Gestion de l'allocation dynamique (Extra Segment)

Après avoir défini le segment de données, le segment de codes et le segment de pile, nous allons introduire l'extra segment (nommé "ES") qui permet l'allocation dynamique de mémoire pendant l'exécution du programme. Pour cela, nous allons commencer par présenter un nouveau mode d'adressage : l'adressage avec préfixe de segment explicite. Cette méthode permet d'accéder à des éléments situés hors du segment de données par défaut, en utilisant une syntaxe de type "[ES:REGISTRE]". Par exemple, `MOV AX,5` suivie de `MOV [ES:AX],10` permet de mettre la valeur 10 dans la sixième case du segment ES.

Q 8.1 Écrivez la fonction `void* segment_override_addressing(CPU* cpu, const char* operand)` pour gérer l'adressage avec préfixe de segment explicite de la forme [SEGMENT:REGISTRE]. La fonction doit :

- Valider le format d'adressage avec une expression régulière. À noter que nos segments sont toujours nommés avec deux lettres (DS, CS, SS, ES), tout comme les registres (AX, BX, CX, DX).
- Extraire le segment et le registre correspondants.
- Récupérer et retourner la donnée stockée dans le segment à la position spécifiée.

Q 8.2 Étendez la fonction `void *resolve_addressing(CPU *cpu, const char *operand)` pour intégrer cette nouvelle méthode d'adressage avec segments.

On s'intéresse maintenant à l'allocation mémoire du segment ES. Plusieurs stratégies d'allocation peuvent être considérées pour déterminer quel segment libre utiliser :

- **First Fit (0)** : Alloue le premier bloc mémoire libre suffisamment grand.
- **Best Fit (1)** : Recherche le bloc mémoire dont la taille est la plus proche de la taille demandée, dans l'optique de minimiser le potentiel gaspillage.
- **Worst Fit (2)** : Sélectionne le bloc avec le plus grand espace libre restant.

Q 8.3 Implémentez la fonction `int find_free_address_strategy(MemoryHandler *handler, int size, int strategy)` permettant de trouver le segment libre à utiliser selon la stratégie (0, 1 ou 2) spécifiée. La fonction doit parcourir la liste des segments libres, appliquer la stratégie spécifiée, et retourner l'adresse de début du bloc à allouer (ou -1 en cas d'échec).

Pour allouer dynamiquement un extra segment, il faut d'abord indiquer sa taille dans le registre AX, indiquer la stratégie d'allocation à utiliser dans le registre BX, puis écrire l'instruction `ALLOC` (sans mnémonique). Pour pouvoir ensuite retrouver l'adresse de l'extra segment, il faut ajouter au contexte du CPU un nouveau registre, nommé ES, qui contiendra cette adresse.

Q 8.4 Modifier la fonction `cpu_init` pour ajouter le registre ES au contexte, en lui associant la valeur -1 initialement.

Q 8.5 Implémentez la fonction `int alloc_es_segment(CPU *cpu)` pour allouer dynamiquement un segment ES avec les contraintes suivantes :

- Utiliser la taille spécifiée dans le registre AX.
- Sélectionner la stratégie d'allocation via le registre BX (0, 1 ou 2).
- Utiliser le drapeau ZF pour signaler le succès (ZF=0) ou l'échec de l'allocation (ZF=1).
- Initialiser le segment "ES" avec des zéros.
- Enregistrer l'adresse de base de l'extra segment dans le registre nommé ES.

Pour libérer l'extra segment, il suffit d'utiliser le mnémonique **FREE**.

Q 8.6 Implémentez la fonction `int free_es_segment()` pour libérer le segment **ES** avec les contraintes suivantes :

- Libérer individuellement chaque valeur allouée dans le segment.
- Supprimer le segment de la table des segments alloués.
- Réinitialiser le registre **ES** à -1 .

Q 8.7 Modifier la fonction `handle_instruction` pour gérer les mnémoniques **ALLOC** et **FREE**.

Fin du projet.